

SMALL SAMPLE PROPERTIES OF STATIONARY
STOCHASTIC MODELS AND THE HURST
PHENOMENON IN HYDROLOGY

by

Zekâi Şen

İnş. Yük. Müh. (Istanbul), M.Sc. (London), DIC (London)

Thesis submitted for the degree of
Doctor of Philosophy in the University of London

November 1974

Dedicated to
My father Zekeriya Şen
My mother Hatice Şen
and
My wife Fatma Şen

ABSTRACT

The overall objective sought in this research work is , broadly, the improvement of techniques for simulating observed annual streamflow sequences and at the same time the substitution of rather cumbersome and costly Monte Carlo techniques used in the design of water resource systems, by exact or approximate analytical expressions which yield the same results.

In particular, the small sample properties of various parameters, namely, the variance, the standard deviation and serial correlation coefficients are analytically treated and necessary expressions derived for normal stationary processes only. Furthermore, the small sample expectations of variables intimately related to Hurst's law are analytically derived; among such variables are the range of cumulative departures, R_n , from the sample mean and the rescaled range, R_n/S . The distribution of R/S in p the case of a known population mean value of the underlying generating process has been given, and the R/S analysis for deciding whether a process is short-term or long-term persistent is analytically documented.

A new model which is referred to as the white Markov process is proposed for modelling Hurst's law. All of the necessary statistical relationships for the application of this new model to annual streamflow sequences are fully presented. Moreover, the multivariate case of the white Markov process for preserving the spatial correlations as well as the temporal correlations is developed. The white Markov process is an approximation to Fractional Gaussian noise.

The study of the small sample properties of hydrological processes is presented from a Bayesian point of view and the necessary formulation for applying a Bayesian approach in estimating the parameters of the ARIMA(1,0,1) process is given.

ACKNOWLEDGEMENTS

The writer wishes to express his indebtedness to the Scientific and Technical Research Council of TURKEY, who awarded him a NATO scholarship in order to undertake this research work. He particularly wishes to express his sincere gratitude to Dr. P.E. O'Connell for his constant guidance, helpful suggestions and encouragement which contributed a great deal to this research work. He thanks Professor J.R.D. Francis for his helpful comments.

He is grateful to Miss. E.M. Shaw for kindly reading the typescript and to his colleague Mr. P.M. Johnston for his helpful suggestions.

He thanks Miss. A. Fahri for her assistance in typing the manuscript. He finally owes much to his wife for her patience and encouragement throughout this thesis.

TABLE OF CONTENTS

	Page
Abstract.....	3
Acknowledgements.....	4
Table of Contents.....	5
List of Symbols and Abbreviations.....	8
Introduction.....	13
CHAPTER 1 SYNTHETIC HYDROLOGY.....	18
1.1 General.....	18
1.2 Models Used in Synthetic Hydrology.....	23
1.3 Model Identification.....	34
1.4 Summary.....	36
CHAPTER 2 RANGE, RESCALED RANGE AND HURST PHENOMENON.....	38
2.1 General.....	38
2.2 The Adjusted Range.....	43
2.3 The Population Range.....	45
2.4 Hurst's Law, Hurst coefficient and Hurst Phenomenon.....	48
2.5 Small Sample Expectation of the Adjusted Range.....	52
2.5.1 The Normal Independent Process.....	61
2.5.2 The Lag-One Markov Process.....	63
2.5.3 The ARIMA(1,0,1) Process.....	66
2.5.4 Fractional Gaussian Noise.....	67
2.6 Small Sample Expectation of the Population Range.....	68
2.6.1 The Normal Independent Process.....	69
2.6.2 The Lag-One Markov Process.....	70
2.6.3 The ARIMA(1,0,1) Process.....	71
2.6.4 Fractional Gaussian Process.....	73
2.7 Small Sample Expectation of the Rescaled Range of Various Hydrological Processes.....	74
2.7.1 The Rescaled Adjusted Range.....	74
2.7.2 The Rescaled Population Range.....	79
2.8 Application of Rescaled Range to Hydrological Processes..	79
2.8.1 The Normal Independent Process.....	80
2.8.2 The Lag-One Markov Process.....	82
2.8.3 The ARIMA(1,0,1) Process.....	83
2.8.4 Fractional Gaussian Noise.....	89

	Page
2.9 R_n/S Analysis.....	89
2.10 The Asymptotic Distribution of Rescaled Population Range. for NIP.....	93
2.11 Summary.....	95
CHAPTER 3 SMALL SAMPLE PROPERTIES OF PARAMETERS OF STATIONARY PROCESSES.....	96
3.1 General.....	96
3.2 The Bias Effect.....	98
3.3 Parameter Estimates.....	103
3.4 Methods of Estimation.....	106
3.4.1 The Method of Moments.....	106
3.4.2 The Maximum Likelihood Method.....	107
3.4.3 Bayesian Estimation.....	108
3.5 Small Sample Expectation of the Population Variance.....	109
3.6 Small Sample Expectation of the Population Standard Deviation.....	113
3.7 Small Sample Estimation of ρ	124
3.7.1 Small Sample Expectation of ρ	126
3.7.2 Small Sample Expectation of ρ^k for the ARIMA(1,0,1) Process.....	131
3.8 Summary.....	134
CHAPTER 4 THE WHITE MARKOV PROCESS.....	143
4.1 General.....	143
4.2 Description of the white Markov Process.....	144
4.3 Preservation of Hurst's Law.....	153
4.4 Small Sample Properties of Estimates of σ^2 , σ , ρ and h	161
4.5 The Non Gaussian Case.....	167
4.6 The Log-Normal white Markov Process.....	175
4.7 The Multivariate Case.....	177
4.7.1 Multivariate Log-Normal white Markov Process.....	188
4.8 Summary.....	189
CHAPTER 5 BAYESIAN APPROACH TO AUTOREGRESSIVE PROCESSES.....	191
5.1 General.....	191
5.2 Bayes Theorem.....	193

	Page
5.3 The Prior pdf of φ	197
5.4 The Likelihood Function of the ARIMA(1,0,1) Process.....	200
5.5 The Posterior pdf of φ	206
5.5.1 Process With Three Unknown Parameters (μ , σ , φ)	209
5.5.2 Process With Two Unknown Parameters (σ , φ)....	212
5.5.3 Process With One Unknown Parameter (φ).....	213
5.6 Bayes Estimates.....	214
5.7 Summary.....	216
CHAPTER 6 CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH.....	232
6.1 General Summary.....	232
6.2 General Conclusions.....	232
6.3 Suggestions for Future Research.....	235
APPENDICES	
A2.1 The Adjusted Range for Lag-One Markov Process.....	237
A2.2 The Adjusted Range for the ARIMA(1,0,1) Process.....	240
A2.3 Linear Independence of the Sample Mean and Sample Standard Deviation.....	242
A2.4 The Equivalence of Range and Rescaled Range.....	244
A2.5 The Asymptotic Distribution of the Rescaled Population Range.....	246
A3.1 Small Sample Expectation of the Estimate of Variance.....	250
A3.2 Small Sample Expectation of the Lag-k Serial Correlation Coefficient for the ARIMA(1,0,1) Process.....	253
A5.1 The Likelihood Function of the ARIMA(1,0,1) Process.....	262
A5.2 Three Parameter Process - Type-I	270
A5.3 Two Parameter Process.....	275
REFERENCES.....	279

LIST OF SYMBOLS AND ABBREVIATIONS

Some of the most frequently used symbols and abbreviations are given in the following list as a supplement to the main text whereas others are defined in the main text wherever they first appear.

$\underline{A}$	(mxm) matrix of coefficients in multivariate case.
ARIMA	Abbreviated form of 'Autoregressive Integrated Moving Average'.
$\underline{Q}$	Lower bound of a log-normally distributed random variate.
$\underline{Q}_i$	Lower bound of a log-normally distributed random variable at site i .
a_i	λ parameter for the Broken Line process.
B	Amount of bias in a biased estimate.
$\underline{B}$	(mxm) matrix of coefficients in multivariate case.
BL	Abbreviated form of 'Broken Line'.
$B(t)$	Ordinary Brownian motion at time t .
$B_h(t)$	Fractional Brownian motion at time t .
$B'_h(t, \delta)$	Derivative of fractional Brownian motion at time t .
$B(n, \rho)$	Magnitude of bias associated with the variance in the case of the Markov process.
$B(n, \rho, \varphi)$	Magnitude of bias associated with the variance in the case of the ARIMA(1,0,1) process.
b_j	Regression coefficient.
$\underline{C}$	(mxm) matrix of coefficients in multivariate case.
C_0	Initial storage of a reservoir, needed to avoid shortages or deficiencies.
C_i	Magnitude of storage at the end of the i -th time period.
C_m	Minimum capacity required to meet all demands.
$\underline{D}$	(mxm) matrix defined in terms of $\underline{A}$, $\underline{B}$ and $\underline{C}$.
D	Total demand withdrawn from a reservoir. In chapter 5 it denotes the historical data.
D_k	Cumulative sum of the demands up to time period k .
d_i	Volume of demand over time period from $i - 1$ to i .

dfGn	Abbreviated form of 'discrete fractional Gaussian noise'.
dB(u)	Infinitesimal increment of ordinary Brownian motion.
E(.)	Expectation operator.
E(.) _n	Denotes expected value of quantity within brackets in samples of size n.
exp	Exponential.
F _X (x)	Distribution function for the random variable x.
F	General bias correction factor for the variance.
fGn	Abbreviated form of 'fractional Gaussian noise'.
fBm	Abbreviated form of 'fractional Brownian motion'.
f(n)	Bias correction factor of the variance in the case of independent process.
f(n, ρ)	Bias correction factor of the variance in the case of the Markov process.
f(n, ρ, φ)	Bias correction factor of the variance in the case of the ARIMA(1,0,1) process.
g(n)	Bias correction factor of the standard deviation in the case of independent process.
H	Estimate of Hurst's coefficient.
$\bar{H}$	Mean of ten values of H.
h	Population value of Hurst's coefficient.
$\underline{I}$	Identity matrix.
I	Total inflow into a reservoir over n year period.
I _k	Cumulative sum of the inflows up to time period k.
J	Jacobian of a transformation.
K	Estimate of Hurst's coefficient.
L(.)	Likelihood function.
l.h.s.	Abbreviated form of 'left-hand side'.
$\underline{M}_0^X$	(mxm) matrix of lag-zero covariances of the Markov component of white Markov process.
$\underline{M}_1^X$	(mxm) matrix of lag-one covariances of the Markov component of white Markov process.
$\underline{M}_2^X$	(mxm) matrix of lag-two covariances of the Markov component of white Markov process.

M_{-0}^Z	(mxm) matrix of lag-zero covariances of the white Markov process.
M_{-1}^Z	(mxm) matrix of lag-one covariances of the white Markov process.
M_{-2}^Z	(mxm) matrix of lag-two covariances of the white Markov process.
M_{-3}^Z	(mxm) matrix of lag-three covariances of the white Markov process.
M_m	Maximum water impounded in a reservoir.
MSE	Abbreviated form of 'mean square error'.
m	Number of sites.
m_m	Minimum water impounded in a reservoir.
N	Normalizing factor.
NIP	Abbreviated form of 'Normal Independent Process'.
n	Sample duration in discrete time.
n^*	Equivalent length of independent observations.
n_m	Value of n at which break to $h = 0.5$ law occurs in pox diagram.
P_n	Largest surplus.
P_i	Peak value in the sequent peak algorithm.
pdf	Abbreviated form of 'Probability distribution function'.
$P(\omega)$	Prior pdf of random variable ω .
$P(\omega D)$	Posterior pdf of parameter ω .
Q_n	Greatest deficit.
q	A parameter for the BL process.
q_i	Monthly flow value at month i.
R_n	Adjusted range of cumulative departures from sample mean.
R_p	Population range of cumulative departures from sample mean.
R_n/S	Rescaled adjusted range.
R_p/S	Rescaled population range.
r.h.s.	Abbreviated form of 'right hand side'.
r.v.	Abbreviated form of 'random variable'.

s^2	Small sample estimate of population variance.
s	Small sample estimate of standard deviation.
s^*	Unbiased estimate of the standard deviation.
T	Sample duration in continuous time.
t	Time.
$V(.)$	Variance of some random variable.
WM	Abbreviated form of 'white Markov'
x_t	Markov component of the WM process.
z_t	Designates the WM process.
α	Parameter of the white Markov process.
β	Parameter of the white Markov process.
γ	Coefficient of skewness.
ε_t	Normal and independent random variable.
$\xi(t)$	A simple BL process.
η_t	Normal and independent random variable with zero mean and unit variance.
χ^2	Chi-square distribution.
λ	Parameter of the white Markov process.
μ	Population mean value of a random variable.
$\hat{\mu}$	Estimate of the population mean value.
ρ	Lag-one autocorrelation coefficient.
ρ_1	Lag-one autocorrelation coefficient.
$\hat{\rho}$	Estimate of population autocorrelation coefficient.
σ	Population standard deviation of a random variable.
$\hat{\sigma}$	Estimate of the population standard deviation.
σ^2	Population variance of a random variable.
$\hat{\sigma}^2$	Estimate of population variance.

τ Time lag.

ϕ_i Weights in an autoregressive process.

ϕ Parameter of the ARIMA(1,0,1) process.

INTRODUCTION

In stochastic hydrology it is necessary to use the statistical properties of a necessarily limited series of observations. Based on these properties, synthetic sequences are then generated in such a manner that they are statistically indistinguishable from the historic record i.e. they statistically resemble the historic record. In maintaining the statistical resemblance between historic and synthetic sequences the small sample properties as well as the population properties of the parameter estimates of a model play an important role. It is not always analytically possible to find exact or even approximate mathematical expressions for small sample properties of an estimate, in which case hydrologists often resort to rather cumbersome and expensive Monte Carlo techniques on a digital computer. On the other hand, the results obtained by Monte Carlo techniques will always be valid for a specified set of parameters; moreover even a slight change in one or more parameters might invalidate the result and require further Monte Carlo experiments. Because of this tedious and cumbersome approach even an approximate analytical formula is most welcome for the same purpose.

The main objective of this study is to avoid the Monte Carlo techniques as far as possible. The small sample properties of the parameters can, in general, be treated from two different points of view; one involves the employment of the classical statistical approach, especially the random sampling theory which has been extensively exploited in synthetic hydrology to date ; the other, a relative newcomer to hydrology, is the Bayesian approach which, when coupled with a decision making procedure, promises to be a powerful alternative in the design of a water resources system. The classical statistics method makes use of the information conveyed by the historic record only, whereas the Bayesian method combines this information with that obtained from other

sources and especially personal experience acquired by the hydrologist, so as to yield an improved level of information on which the estimates of parameters can be based.

In recent years much attention has been directed to the remarkable discovery of Hurst(1951,1956) who studied the long-term fluctuations within a large number of geophysical records and found that $R_n/S \sim n^h$ where R_n is the range of cumulative departures from the sample mean and S is the estimate of standard deviation. On the other hand, the small sample properties and the population properties of the range, R_n , and the rescaled range, R_n/S which are intimately related to the above referred remarkable discovery, have been analytically derived for various processes and their verification has been accomplished through using the Monte Carlo technique. In the context of reservoir storage, the importance attached to R_n is, because it represents the size of an ideal reservoir where there is no spillage or shortage over the economic life of the structure. The rescaled range, R_n/S has been proposed as a measure of the long-term persistence in empirical observations and in turn related to the parameter $0 < h < 1$, which governs the duration and intensity of periods of above and below average flow. The most remarkable fact concerning h is that its overall average value for various empirical records has been found to be 0.73 by Hurst(1951), which does not conform to the value of $h=0.5$ predicted by classical stochastic processes. This disagreement between the two values of h has been labelled the Hurst phenomenon and has led both hydrologists and statisticians alike to search for a model that could account for $h > 0.5$.

Mandelbrot (1965), Mandelbrot and Van Ness(1968) became interested in the Hurst phenomenon and presented a mathematical framework to explain Hurst's finding. This led to a new model of annual streamflow called discrete time fractional Gaussian noise (dfGn) which is based on the

fractional Brownian motion. This process made it possible to model a measure of long-term persistence specified by $0.5 < h < 1$. Mandelbrot and Wallis (1969, a, b, c, d, e) proposed approximations to dfGn for generating synthetic sequences on a digital computer. In order to overcome some of the difficulties met in these approximations Mandelbrot (1971) and Matalas and Wallis (1971b) proposed, fast and filtered dfGn approximations respectively.

However, the practical application of the above mentioned models is still rather expensive and cumbersome; therefore research for a simple stochastic model with the necessary long-term behaviour has become inevitable and eventually such a model was proposed by O'Connell (1971) under the label of the ARIMA(1,0,1) process which is shown to be an alternative to dfGn for modelling the Hurst coefficient h . The white Markov process, WM, which is proposed in this thesis is closely related to the ARIMA(1,0,1) process and has all the latter's advantages.

However, theoretical work on the properties of the range and rescaled range and the small sample properties of parameter estimates for various processes has been found lacking to date. Although these properties have been in some cases derived by Monte Carlo techniques they are in no way a substitute for analytical results which are more general and which give more insight into the problem. An analytical approach is adopted here and some new results are derived.

Chapter 1 gives a review of synthetic hydrology and the stochastic models currently available in the hydrological literature. The assumptions and simplifications necessary for applying the theory of stochastic processes to a hydrological process, have been presented. Finally, the relevant procedures for identifying a model suitable to represent the historical record have been reviewed.

The small sample expectations of the statistics R_n and R_n/S for various stochastic processes have been studied in Chapter 2. Firstly, a review of work relating to R_n or R_n/S has been presented and four different

types of statistics have been distinguished. These are, namely, the population range, R_p , the adjusted range, R_n , the rescaled population range, R_p/S and the rescaled adjusted range, R_n/S . The asymptotic expectations of the population and adjusted ranges have previously been derived for the Normal Independent Process (NIP) only, by Feller (1951). For the NIP, the small sample expectations of R_p and R_n have been given by Anis and Lloyd (1953) and Solari and Anis (1957). However, by pursuing a different kind of methodology to those authors which is discussed in chapter 2, the small sample and asymptotic expectations of R_p and R_n for any normal stationary process have been derived. Moreover, the analytical small sample and asymptotic expectations of R_p/S and R_n/S have also been derived by using the same methodology. As a result the analytical expressions found for $E(R_p/S)$ and $E(R_n/S)$ avoid any expensive computer investment through Monte Carlo techniques for studying the behaviour of R_p/S and R_n/S . In addition, the asymptotic distribution of the rescaled population range, R_p/S , has been obtained and the R_n/S analysis which helps to decide whether a process is short-term or long-term persistent has been analytically investigated. In this context some deductions of Mandelbrot and Wallis (1979e) have been analytically confirmed.

Chapter 3 relates more directly to the generation of synthetic sequences and after a review of existing work on small sample properties of some parameters, the small sample expectations for the variance, σ^2 , the standard deviation, σ and the lag-k serial correlation, ρ_k , have been explicitly described for normal stationary stochastic processes at large. The small sample expectation of the estimate of population standard deviation is derived and its application to the ARIMA(1,0,1) process is shown. The small sample properties of the estimate of the population serial correlation have been mathematically treated and necessary approximate relationships have been derived. These results can be used to provide bias correction algorithms for estimates of parameters.

Chapter 4 deals with the development and application of the white Markov (WM) process which can preserve the Hurst coefficient, h , for a length of synthetic sequence comparable with the longest geophysical records available. First of all, the statistical properties of the WM process have been studied and then the equivalence of this process to the ARIMA(1,0,1) process has been shown. On the other hand, the WM process has been shown to be a simple approximation to dfGn and to model the Hurst coefficient, h , sufficiently. The small sample properties of various parameters of the process have been analytically treated and the necessary formulae derived. Moreover, the non-Gaussian case of the WM process has been developed and the procedures for generating log-normal WM variates have been illustrated where the preservation of skewness has been considered. In addition to this, a multisite WM process with the required temporal and spatial correlation structures is developed and the analytical expressions necessary to determine the elements of unknown matrices are given. The generation of multivariate log-normal sequences is also considered.

The Bayes theorem and the estimation of the parameters of a stochastic model by Bayesian approach forms the subject matter of chapter 5. The three necessary ingredients of the Bayesian approach which are the prior probability distribution function (pdf), the maximum likelihood (ML) function and the posterior pdf, are studied in general. In particular, the application of this approach is given for the ARIMA(1,0,1) process and for the appropriate parameter values, the general analytical expressions derived are shown to reduce to those for the lag-one Markov process derived by Lenton et al (1973). Only the marginal posterior pdf of one model parameter, ϕ , is derived based on varying assumptions about the remaining parameters.

Chapter 6 comprises the main conclusions of the thesis and proposals for future research.

Chapter 1

SYNTHETIC HYDROLOGY

1.1 General

In general, the science of hydrology is concerned with the hydrological processes such as the rainfall, streamflow, infiltration, groundwater etc. all of which interact over a catchment area. The main objective of a hydrologist is to quantify the very complex behaviour of these hydrological processes and relationships existing among them.

Hydrologists have attacked the problem of investigating hydrological processes from two different points of view. One of these is to regard a complex natural process as occurring according to a definite set of physical laws, where no law of uncertainty is involved; this approach will be referred to as the deterministic approach. An alternative approach is to consider such a process as a product of some underlying random mechanism, which can only be treated by employing the laws of uncertainty and by making use of several branches of mathematics such as probability theory, the theory of stochastic processes, statistics etc. In order to apply either one of the approaches the first step is to represent the hydrological processes in a quantitative manner on which the future studies concerning the same process can be based. This quantification is achieved by sampling the process either continuously or discretely over equal intervals, along the time or space axis whichever one is the reference system; so that a sequence of observations is obtained. Although all of the hydrological processes have been existing and continuously evolving since the creation of the universe unfortunately their actual measurement has started at the earliest about one or two centuries ago; in fact most of them started only very recently. As a result there exists a sequence of observations taken at a finite set of time points, which represents an inadequate set of data insofar as the design of a water resource system is

concerned. However, the consideration of each observation as a random variable helps the hydrologist to infer the population probability structure of the process concerned, and to assess the inherent risks and uncertainties associated with the hydrological design in the light of the probability theory and statistics. The branch of hydrology dealing with the above referred studies is termed the stochastic hydrology which gives a scientific description of the structure of hydrological processes.

The general definition of a stochastic process can be given as a random phenomenon which evolves along some reference system which may be either time or space. In this thesis the terms stochastic processes and hydrological processes will be used interchangeably so that all of the known facts about stochastic processes will remain to be applicable to hydrological processes bearing in mind the assumptions which will be given in the following section. Notationally, a stochastic process is represented by $[X(t, \omega): t \in T, \omega \in \Omega]$ where t is an element of the time variable T and Ω is the sample space for the random variable $X(t, \omega)$ for each value of t , emphasizing the fact that for a fixed time instant there is a set of random variables. In general, the index set T can assume values as $T = [t: -\infty < t < +\infty]$ continuous but for a stochastic process specified at a set of discrete time points the index set is defined as $T = [t: 0, \tau_1, \tau_2, \tau_3, \dots]$. As far as the hydrological processes are concerned the index set T can take on positive values only. Particular importance is attached to the time index when the two random variables, $X(t_1, \omega)$ and $X(t_2, \omega)$ where $t_1 \neq t_2$ are interdependent, that is, the random variable $X(t_1, \omega)$ exerts an influence on the r.v. $X(t_2, \omega)$ to some extent, otherwise the ordering of the random variables is immaterial.

In the case of the lack of interdependence between the random variables the probabilistic structure i.e. the probability distribution function (pdf) of the random variable will be sufficient to represent the hydrological process entirely; this pdf is then referred to as the

probabilistic model of the hydrological process. On the contrary, the model of the hydrological process where the random variables are interrelated will be referred to as the mathematical or stochastic model. As a result the probabilistic model is a special kind of stochastic model where interdependence between the random variables does not exist. A stochastic process can be thought of as a collection of equally likely realizations which is known as an ensemble; in practice, only one of such a realization is observed over a finite time period which constitutes a time series. It is this sole time series from which a set of parameters representing the probabilistic structure of the underlying random variables and the interdependence between the random variables of the hydrological process is obtained by the method of estimation which is referred to in chapter 3 of this thesis. The incorporation of this set of parameters in a suitably chosen stochastic model for the hydrological process will then give the hydrologist the ability for generating other equally likely realizations of the ensemble. This generating scheme is the basis of the synthetic hydrology which has brought new insight into the design of water resource systems. Synthetic hydrology makes it possible to assess the performance of a hydrological system, such as a reservoir storage and flood control structures design etc., by attaching a quantitative risk to each realization routed through the system.

Synthetic hydrology is one of the few means at the disposal of the hydrologist who strives to overcome most of the inadequacies and uncertainties associated with the design of water resource systems. If a finite historic record were available only, then it would be inadequate in the sense that (a) its length might not be equal to the projected life of a water resource system and (b) the design value obtained from historic record will be uncertain in the sense that it is but one of the design values whose merit relative to the optimal design is not known. However, by the use of synthetic sequences of the same probabilistic structure as

the historic record, it is possible to generate any length of sequences compatible with the actual life of the water resource system so that, this gets rid of the above mentioned inadequacy (a), whereas the generation of an ensemble of equally likely flows eliminates the uncertainty referred to under (b). One important remark at this stage is that, the generation of an ensemble of equally likely sequences by the employment of synthetic hydrology does not create new information but provides an effective means of utilizing information conveyed by the historic observations. Therefore, the synthetic sequences are entirely based on the historic record.

The main concern of this thesis will be confined to annual streamflow sequences only, for which the underlying generating mechanism is not known exactly, but it can be approximated by the application of the theory of stochastic processes available in the statistics literature to the historic record. To this end some fundamental assumptions concerning the statistical behaviour of the hydrological process have to be made beforehand. One of the most important assumptions is that the underlying process of streamflow is a stationary stochastic process where stationary implies that the statistical properties of the process are independent of absolute time. Several types of stationarity are encountered in the theory of stochastic processes; strictly stationary stochastic processes constitute a class where all of the statistical parameters of the process are independent of absolute time, but a function of the time difference only or in another sense these processes have any order of multivariate probability distributions independent of absolute time. This property is tantamount to a statistical equilibrium of the process which will be reflected in the synthetic sequences based on the historic record. Unfortunately, it is almost impossible to observe any strictly stationary hydrological process in nature, because some of the external factors, such as artificial changes by man etc. cause the characteristics of the process to vary with time and hence a non-stationary process emerges. Non-stationary processes are difficult to

handle, because stochastic process theory has been devised almost entirely for dealing with stationary stochastic processes only and very little work relates to non-stationary processes. Another class of stationary processes has lower order moments up to a certain order k which are independent of absolute time, in which case the process is referred to as a k -th order stationary stochastic process. Hence, second order stationary stochastic process which is sometimes called a weak^{ly} stationary process, has its mean and covariance function independent of absolute time. It is this last class of stationarity which is mostly employed in synthetic hydrology. Therefore, the time series of a hydrological process is either assumed to be at least second order stationary in the case of negligible changes with time in the statistical parameters or by applying a convenient transformation such as the Fourier analysis, trend analysis etc. to the original sequence of observations to render the series second order stationary. If the underlying random variable of a second order stationary stochastic process is normally distributed, then the process automatically becomes a strictly stationary stochastic process.

Another very important assumption in synthetic hydrology is that the hydrological process concerned is ergodic. This property assures that the time averages are equal to the ensemble averages. As a result the time averages calculated from the historic record which is a sole realization available out of the ensemble characterizing hydrological process, replace the ensemble averages. Hence, the ergodicity assumption enables hydrologists to measure various probabilistic characteristics, such as the mean, covariance function, one or multi dimensional distribution functions etc. from a sequence of observations, i.e. the time series of a hydrological process. The two major assumptions made above, namely, stationarity and ergodicity of a stochastic process are fundamental to synthetic hydrology, and permit the hydrologist to generate an ensemble of equally likely synthetic realisations over the actual projected life of a water resource system and in turn, for

assessing the risks associated with the design. In the following section a brief review of stochastic processes currently used in synthetic hydrology will be presented.

1.2 Models Used in Stochastic Hydrology

Although the underlying generating mechanism of a hydrological process remains unknown an approximation to it based on extractable information from the historic sample of observations from the process can be proposed. To this end, first of all the historic record is represented in a quantitative and compact manner by a set of statistical parameters of which the most prominently used ones in hydrological studies are, the mean, μ , the standard deviation, σ , the coefficient of skewness, γ , the first order serial correlation, ρ_1 , and the Hurst coefficient h . Each one of these parameters depicts an important feature of the hydrological process; the mean is the value around which the pattern of the observations of the process fluctuates, the variance is a measure of the variability of the fluctuations i.e. the deviation of observations about the mean value; the coefficient of skewness is a measure of the deviation of the pdf of the underlying random variable from a symmetric distribution; the first order serial correlation coefficient gives a quantitative measure of the degree of linear dependence between the observations one time unit apart from each other. The magnitude of, ρ_1 , decreases as the time interval between the two successive observations increases. Consequently one would expect the daily flows to be more dependent than the monthly flows etc. Some of the parameters may not be sensitive to a particular design situation and in such a case the preservation of the parameters in the generating scheme is immaterial.

A purpose of a stochastic model is to generate synthetic sequences such that they will have the same statistical properties as a historic record. Therefore, the above mentioned prominent parameters of a hydrological

process must be preserved in any model that gives rise to the generation of synthetic sequences. As the number of the parameters to be preserved in a generating scheme increases, the structure of the model becomes more sophisticated. The earliest and the simplest stochastic model employed in hydrological studies is the Normal Independent Process (NIP) which preserves the two prominent parameters, namely, μ and σ^2 only (Barnes, 1954). The NIP completely ignores the interdependence among the observations, as a result of which the water resource system will be either overdesigned or underdesigned. For instance, in the case of reservoir design, if the interdependence is ignored then the resulting size of reservoir will be underdesigned.

The first method which took into account the dependence of observations is apparently due to Thomas and Fiering (1962) who were also pioneers in using an electronic digital computer and the sampling of random variables on the computer in the modelling of a hydrological process. The model they used was based on monthly flows which took into account the seasonality in the monthly means, standard deviations and lag-one autocorrelation coefficients and may be given as,

$$q_{i+1} = \mu_{j+1} + b_j(q_i - \mu_j) + \sigma_{j+1} \frac{(1 - \rho_j^2)^{1/2}}{\sigma_j} \epsilon_{i+1} \quad (1.1)$$

where q_{i+1} and q_i are the flow values during the (i+1)-th and i-th months respectively; μ_j and μ_{j+1} are the mean monthly flows for month j and j+1 where $j=1,2,3,4,\dots,12$, b_j is a regression coefficient resulting from regressing the flows of month (j+1) on those of month j, and finally ϵ_{i+1} is the random term independent of q_i . In the case of normal monthly flows, ϵ_{i+1} is also a normal random variable with zero mean and unit variance. The relationship between the regression coefficient, b_j and the lag-one autocorrelation coefficient may be given as

$$b_j = \rho_j \frac{\sigma_{j+1}}{\sigma_j} \quad (1.2)$$

By substituting Eq.(1.2) into Eq.(1.1) and then dividing both sides by σ_{j+1}

the following form emerges

$$\frac{q_{i+1} - \mu_{j+1}}{\sigma_{j+1}} = \rho_j \frac{q_i - \mu_j}{\sigma_j} + (1 - \rho_j^2)^{1/2} \epsilon_{i+1}$$

or in terms of standardized variables

$$z_{i+1} = \rho_j z_i + (1 - \rho_j^2)^{1/2} \epsilon_{i+1} \quad (1.3)$$

which reveals the Markovian nature of the underlying generating process given in Eq.(1.1).

In its basic form the lag-one Markov process has been defined by Matalas (1967) as

$$(x_t - \mu_x) = \rho_x (x_{t-1} - \mu_x) + (1 - \rho_x^2)^{1/2} \sigma_x \epsilon_t \quad (1.4)$$

where x_t and x_{t-1} denote the events at time instances t and $t-1$, respectively, μ_x and σ_x are the mean and standard deviation of x_t respectively, ρ_x is the lag-one serial correlation coefficient for x and finally ϵ_t is a random component which is independent of x_t and has zero mean and unit variance. The parameters μ_x , σ_x and ρ_x are unknown, but they may be estimated from a given historic record. If these estimates, namely, $\hat{\mu}_x$, $\hat{\sigma}_x$ and $\hat{\rho}_x$ are incorporated into Eq.(1.4) instead of μ_x , σ_x and ρ_x respectively then a sequence of synthetic observations statistically indistinguishable from the historic observations can be generated. As a result, the synthetic observations will resemble the historic observations in terms of $\hat{\mu}_x$, $\hat{\sigma}_x$ and $\hat{\rho}_x$.

However, if the resemblance is to be maintained in terms of coefficient of skewness, γ_x as well as μ_x , σ_x and ρ_x then the random term, ϵ_t in Eq.(1.4) has to be replaced by η_t which is defined as

$$\eta_t = \frac{2}{\gamma_\eta} \left(1 + \gamma_\eta \frac{\epsilon_t}{6} - \frac{\gamma_\eta^2}{36} \right)^3 - \frac{2}{\gamma_\eta} \quad (1.5)$$

where the skewness of η_t , denoted by γ_η , is related to the skewness, γ_x of the process X_t by

$$\gamma_x = \frac{(1 - \rho_x^2)^{3/2}}{(1 - \rho_x^3)} \gamma_\eta \quad (1.6)$$

In the case of a normally distributed term ε_t with zero mean and unit variance, then η_t will be approximately distributed as gamma with zero mean, unit variance and coefficient of skewness equal to γ_η . An estimate of γ_η , $\hat{\gamma}_\eta$, can be obtained from the estimate of γ_X , $\hat{\gamma}_X$, abstracted from a historic sample through Eq.(1.6) and then by using Eq.(1.5) η_t r.v's are generated which replace ε_t in Eq.(1.4) to yield flows which are approximately distributed as gamma.

Matalas (1967) has formulated the generation of synthetic flows which conform to a three parameter log-normal distribution and resemble historic flows in terms of μ_X , σ_X , ρ_X and γ_X . If 'a' is the lower bound of a r.v. X_t and $(X_t - a)$ is log-normally distributed then $Y_t = \ln(X_t - a)$ will be normally distributed. The relationships between the lower order moments of the r.v's X_t and Y_t have been derived by Aitchison and Brown (1957) as,

$$\mu_X = a + \exp\left(\frac{\sigma_Y^2}{2} + \mu_Y\right) \quad (1.7)$$

$$\sigma_X^2 = \exp\left(2(\sigma_Y^2 + \mu_Y)\right) - \exp(\sigma_Y^2 + 2\mu_Y) \quad (1.8)$$

and

$$\gamma_X = \frac{\exp(3\sigma_Y^2) - 3\exp(\sigma_Y^2) + 2}{[\exp(\sigma_Y^2) - 1]^{3/2}} \quad (1.9)$$

If it is assumed that the flows Y_t are generated by a lag-one Markov process then the relationship between ρ_Y and ρ_X has been given by Matalas (1967) as,

$$\rho_X = \frac{\exp(\sigma_Y^2 \rho_Y) - 1}{\exp(\sigma_Y^2) - 1} \quad (1.10)$$

which may be solved for ρ_Y provided that σ_Y and ρ_X are known. The flows generated in Y-space will have a lag-one Markovian nature, whereas their transformation into X space will constitute a non-linear process.

The lag-one serial correlation coefficient, ρ_1 , is directly related to high frequency behaviour and may not represent the low frequency behaviour

of a hydrological process. As the autocorrelation function of the lag-one Markov process is specified by ρ_1 , its main drawback is that it fails to preserve low frequency effects which are synonymous with long term persistence. Persistence in an observed sequence is referred to as the tendency for high flows to follow high flows and low flows to follow low flows. In the case of the existence of persistence in a hydrological process the dependence between far distant observations can be very small but such that the cumulative effect of such correlations is not negligible. A quantitative measurement of long-term persistence is achieved by the Hurst coefficient, h , which has its origin in the studies of Hurst (1951,1956). An empirical result given by Hurst states that

$$\frac{R_n}{S} \sim n^h \quad (1.11)$$

where R_n is the range of cumulative departures which is exclusively studied in chapter 2, and S is the standard deviation, both obtained from a historic sequence of length n . On the basis of empirical studies of 900 time series pertaining to streamflow, rainfall, tree rings, mud varves and temperature, Hurst (1951) found an average value for h of 0.73. This is contrary to $h=0.5$ which is the value yielded by the NIP and lag-one Markov process. As a result it can be said that neither the NIP nor the lag-one Markov process can account for the long-term persistence and therefore these two processes can not preserve a h value which lies in the range $0.5 < h < 1$. The deviation of the empirical h from the theoretical value of 0.5 led hydrologists to seek new models which would be capable of preserving not only the set of parameters μ , σ , γ and ρ_1 but also the Hurst coefficient, h . For this purpose Fiering (1967) tried to preserve h by using multi-lag Markovian processes and he concluded that in order to preserve h over a span of 60 years it would be necessary to employ 20 lags in his model. This result agrees with Mandelbrot and Wallis (1969a) who found the transient period before the 0.5 law holds to be three times the memory of the process.

Later, Mandelbrot (1965) and Mandelbrot and Van Ness (1968) became interested in explaining Hurst's result and they proposed a model called Fractional Gaussian Noise. By considerations from the well known Brownian process, Mandelbrot and Van Ness (1968) introduced fractional Brownian motion the increments of which give a stationary Gaussian process which is called fractional Gaussian noise. They have then defined fractional Brownian motion (fBm), $B_h(t)$ from ordinary Brownian motion $B(t)$ by the following integral transformation,

$$B_h(t) - B_h(0) = \frac{1}{\sqrt{h+0.5}} \left[\int_{-\infty}^0 \left[(t-u)^{h-0.5} - (-u)^{h-0.5} \right] dB(u) + \int_0^t (t-u)^{h-0.5} dB(u) \right] \quad 0 < h < 1 \quad (1.12)$$

where $dB(u)$ is an infinitesimal increment of ordinary Brownian motion.

One of the most appealing properties of this process is the fact that it is self-similar. This means that $[B_h(t) - B_h(0)]$ has the same distribution as $T^{-h} [B_h(tT) - B_h(0)]$. This property implies the fact that the mean range of cumulative departures of equally spaced increments of this process pursue a behaviour similar to Hurst's law given in Eq(1.11).

The process $B_h(t)$ as such has no derivative, because of the fact that the local behaviour of $B_h(t)$ is very erratic. In order to circumvent the lack of derivative of the process, $B_h(t)$ is smoothed by introducing a new process as

$$B_h(t, \delta) = \frac{1}{\delta} \int_t^{t+\delta} B_h(s) ds \quad \delta > 0$$

This type of smoothing produces an alteration in the high frequency/^{domain} such that the increments of this new process are now stationary. Hence the stationary derivative

$$B'_h(t, \delta) = \delta^{-1} \left[B_h(t+\delta) - B_h(t) \right] \quad (1.13)$$

is known as fractional Gaussian noise. A discrete time fractional Gaussian noise can be defined as

$$\Delta B_h(t) = B_h(t) - B_h(t-1) \quad (1.14)$$

where t takes only integer values.

Two approximations to fractional Gaussian noise have been proposed by Mandelbrot and Wallis (1969a) for computer simulation. The type 2 approximation is difficult to formulate and time consuming in the computer but is better than type 1 which is a cruder approximation. The mathematical form of the type 2 approximation is given as

$$X(t) = (h - 0.5) \sum_{i=t-M}^{t-1} (t-i)^{h-3/2} \epsilon_i \quad (1.15)$$

where M is a very large integer value denoting the memory, ϵ_i is the NIP with mean μ_{ϵ} and variance σ_{ϵ}^2 . A further approximation to discrete fractional Gaussian process has been developed by Wallis and Matalas (1967). This approximation is referred to as a filtered type 2 approximation which is given as

$$X(t) = (h - 0.5) \sum_{i=pt-M}^{pt-1} (pt-i)^{h-3/2} \epsilon_i \quad (1.16)$$

where $p \geq 1$ is an integer. It is evident that the r.v. $X(t)$ given in Eq.(1.16) is equivalent to sampling values at times pt from the type 2 approximation expressed by Eq.(1.15). The mean, variance and coefficient of skewness of the type 2 approximation are the same as for the filtered process and these have been derived by Matalas and Wallis (1971b) as,

$$\mu = (h - 0.5) \sum_{i=0}^{M-1} (M-i)^{h-3/2} \mu_{\epsilon} \quad (1.17)$$

$$\sigma^2 = (h - 0.5) \sum_{i=0}^{M-1} (M-i)^{2h-3} \sigma_{\epsilon}^2 \quad (1.18)$$

and

$$\gamma = \frac{\sum_{i=0}^{M-1} (M-i)^{3h-4.5}}{\left[\sum_{i=0}^{M-1} (M-i)^{2h-3} \right]^{1.5}} \quad (1.19)$$

However, the relationship between the autocorrelation functions is given as,

$$\rho_{kp} = \rho_k \quad (k=0,1,2,3,\dots)$$

where ρ_{kp} and ρ_k are the k-th order autocorrelation coefficients of the filtered type 2 approximation and original type 2 approximation respectively.

However either process is time consuming to generate on a computer.

After the appearance of dfGn in the hydrological literature, because of its requirement of large computer investment, hydrologists have begun to grapple with the problem of replacing approximations to dfGn such as the type 2 and filtered type 2 by simple and fast processes. One such process has been introduced into hydrology by O'Connell (1971) which is referred to as the ARIMA(1,0,1) process of which mathematical foundations have been laid down by Box and Jenkins (1970). The ARIMA(1,0,1) process is a mixture of a first order moving average and a first order autoregressive process and is expressed as

$$X_t = \phi X_{t-1} + \epsilon_t - \theta \epsilon_{t-1} \quad (1.20)$$

$$\rho_1 = \frac{(\varphi - \theta)(1 - \varphi\theta)}{(1 + \theta^2 - 2\varphi\theta)} \quad (1.21)$$

$$\rho_k = \varphi \rho_{k-1} \quad (1.22)$$

Although the ARIMA(1,0,1) process lies within the Brownian domain of attraction for which $h=0.5$, it has been shown by O'Connell (1974) that for suitably chosen values of φ and θ the process maintains good agreement with Hurst's law with $h>0.5$ for values of n as large as 10000; as a result both short and long-run ^{properties} may be modelled simultaneously. The ARIMA(1,0,1) process can be regarded as an approximation to dfGn. For the purpose of generating synthetic sequences the ARIMA(1,0,1) process may be formulated as

$$(X_t - \mu_x) = \varphi(X_{t-1} - \mu_x) + \sigma_x \sigma_\epsilon (\epsilon_t - \theta\epsilon_{t-1}) \quad (1.23)$$

where μ_x and σ_x are the mean and standard deviation of the process respectively. The term ϵ_t is an independent random variate with zero mean and unit variance and σ_ϵ is defined as

$$\sigma_\epsilon^2 = \frac{(1 - \varphi^2)}{(1 + \theta^2 - 2\varphi\theta)} \quad (1.24)$$

If the observed skewness of the historic sequence is to be preserved in the generating scheme then the random component ϵ_t in Eq.(1.23) is replaced by η_t given in Eq.(1.5). The relationship between the skewness γ_x of the process X_t and the skewness γ_η of the term η_t has been derived by O'Connell (1974) as,

$$\gamma_X = \left\{ \left[\frac{(1 - \varphi^3 + 3\varphi\theta^2 - 3\varphi^2\theta)}{(1 - \varphi^3)} \right] / \left[\frac{(1 + \varphi^2 - 2\varphi\theta)^{3/2}}{(1 - \varphi^2)} \right] \right\} \cdot \gamma_\eta \quad (1.25)$$

The estimate $\hat{\gamma}_\eta$ of γ_η can be obtained from Eq(1.25) provided that the estimate $\hat{\gamma}_X$ of the skewness γ_X is obtained from a given historic sequence.

A further new model known as the Broken Line (BL) process has been introduced into hydrology by Mejia et al (1972). The simple BL process was introduced by Ditlevsen (1969) in order to check, by the Monte Carlo method, some results in first passage theory. Because of the following two properties the BL process has been claimed to be a possible model for geophysical time series,

(a) the existence of a second derivative at the origin of its correlation function

(b) the possibility of modelling long spans of dependence among its components in such a manner that the simulation of the resulting process is faster than for any other process

The mathematical formulation of a simple BL process has been given by,

$$\xi(t - ka) = \sum_{i=0}^{\infty} \left[\eta_i + \frac{(\eta_{n+1} - \eta_n)(t - na)}{a} \right] I(t)_{[na, (n+1)a]} \quad (1.26)$$

where η_n are independent and identically distributed r.v's with zero mean and variance σ^2 , k is a r.v. uniformly distributed over the interval $(0,1)$, 'a' is the distance in time separating the η 's, and $I(t)_{[na, (n+1)a]}$ is an indicator function which satisfies

$$I(t)_{[na, (n+1)a]} = \begin{cases} 1 & \text{for } na \leq t \leq (n+1)a \\ 0 & \text{otherwise} \end{cases} \quad (1.27)$$

The simple BL process given in Eq(1.26) is a continuous time process and its mean, variance and correlation function are given as follows,

$$E [\xi(t)] = 0 \quad (1.28)$$

$$\text{Var} [\xi(t)] = E [\xi^2(t)] = \frac{2}{3} \sigma^2 \quad (1.29)$$

and

$$\rho(\tau) = \begin{cases} 1 - \frac{3}{4} \left(\frac{\tau}{a}\right)^2 \left[2 - \left(\frac{\tau}{a}\right)\right] & \text{for } 0 < \tau < a \\ \frac{1}{4} \left[2 - \left(\frac{\tau}{a}\right)\right]^3 & \text{for } a < \tau < 2a \\ 0 & \text{for } \tau > 2a \end{cases} \quad (1.30)$$

The addition of m independent simple BL processes gives the general BL process, as

$$X^m(t) = \sum_{i=1}^m \xi_i(t) \quad (1.31)$$

where $\xi_i(t)$ is a simple BL process with a_i and σ_i . Because of the independent superimposition of m simple BL processes, the variance and autocorrelation function of $X^m(t)$ becomes

$$V [X^m(t)] = \frac{2}{3} \sum_{i=1}^m \sigma_i^2 \quad (1.32)$$

and

$$\rho_x(\tau) = \frac{\sum_{i=1}^m \sigma_i^2 \rho_i(\tau)}{\sum_{i=1}^m \sigma_i^2} \quad (1.33)$$

$\rho_i(\tau)$ is the correlation function of $\xi_i(t)$. Mejia (1971) has applied the following restrictions to the parameters

$$a_i = a_1 q^{i-1} \quad \text{where } q > 1$$

and

$$\sigma_i = \sigma_1 \left(\frac{a_1}{a_i}\right)^{0.5} = \sigma_1 \left(\frac{1}{q^{i-1}}\right)^{0.5} \quad (1.34)$$

The memory of the process, which is equal to the time lag for which the

correlation function is ^{not} zero, is given by 2a which is controlled by q and m . Mejia et al (1972) have shown that the BL process may be derived as an approximation to dfGn. A detailed review of the process has been given by O'Connell (1974).

1.3. Model Identification.

Confronted with the problem of generating synthetic annual streamflow sequences, the main task of the hydrologist after having extracted the estimates of the necessary statistical parameters such as μ , σ , γ , ρ and h from given historic data, is to identify a model among the currently used hydrological processes cited in the previous section, in such a way that it best represents the historic record. Some of the parameters are not sensitive to the model identification procedure, such as μ , σ and γ which can be preserved by most generating models using a suitable distribution function as mentioned in the previous section.

However, there remains ρ and h only to provide guidance in the identification scheme both of which are directly related to the autocorrelation structure of the process concerned. Moreover, ρ is intimately related to the behaviour of the correlogram for small lags whereas, h pertains, to the behaviour of the correlogram for large lags. On the other hand in terms of frequencies ρ is representative of high frequencies while h is related to low frequency effects in hydrologic sequences. When long-term persistence is absent in a sequence that is to say $h=0.5$ then the autocorrelation function, acf, will be the primary means of identification. A complementary identification tool, the partial autocorrelation function explained by Box and Jenkins (1968) and Jenkins and Watts (1968) can also be used.

Various types of stochastic processes can be distinguished

according to their theoretical acf. In the case of the NIP the acf is equal to unity for lag-zero whereas, for all the other lags its value is zero. If the sample acf obtained from a historic sequence conforms to this type of theoretical acf then the underlying generating mechanism is identified as NIP. However, numerous tests for establishing statistical independence in a hydrologic sequence are available in the statistics literature; a selection is given by Matalas (1967). One of the most powerful tests is proposed by Anderson (1942) who shows the distribution of the estimate $\hat{\rho}$ for NIP to be normal with mean $-1/(n-1)$ and variance $(n-2)/(n-1)^2$, hence the confidence limits for $\hat{\rho}$ is given by

$$C.L.(\hat{\rho}) = \left[-1 \pm Z_{\alpha} (n-2)^{1/2} \right] / (n-1) \quad (1.35)$$

where Z_{α} is the standard normal deviate corresponding to a probability level α . For instance, if $\hat{\rho}$ value lies within the 95% confidence limits, then it can be concluded that the process is NIP.

Another two broad classes of stochastic processes are the pure moving average process of order q given as

$$X_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \dots - \theta_q a_{t-q} \quad (1.36)$$

where $\theta_1, \theta_2, \dots, \theta_q$ are the weights, for which the acf is terminated, being zero after q lags and the autoregressive processes, such as the lag-one Markov process given in Eq.(1.4), which are characterized by attenuation of the acf. Thus the acf is a useful tool in deciding whether the process is pure autoregressive or pure moving average.

The acf can be used more precisely by considering its estimation properties. As was stated by Carlson et al (1970) a useful simple approximation is to assume that the estimate $\hat{\rho}_k$ is normally distributed with zero mean and variance $1/n$. If all $\hat{\rho}_k$ values beyond lag q lie within, say, 95% confidence limits of $1.96/n^{1/2}$ then it can be decided that the process is moving average of order q . In some situations it may not be entirely clear whether the acf terminates or attenuates.

In such a case the partial acf will be helpful as a complementary tool because of the fact that the acf of an autoregressive process attenuates and the partial acf of the same process truncates, the reverse is valid for the moving average processes. If both the samples acf and partial acf attenuate, then a mixture of moving average and autoregressive process, such as the ARIMA(1,0,1) process results. Hence, the model whose theoretical acf is very close to the estimated acf of a given sample will be identified as the best model that suits the data. The paucity of hydrological data limits the application of acf in identifying the underlying generating mechanism.

However, the acf mentioned above tends to fit a model to historic data in the high frequency domain only, whereas most of the low frequency effects are ignored. Therefore, to have a more realistic identification, long-term persistence measured by the Hurst coefficient h , must also be preserved in the generated sequences. In the case of $h > 0.5$ the processes that might be used to simulate annual streamflow sequences are the approximations to dfGn. In practical situations, if cumbersome and expensive computer simulations are not to be undertaken then simple approximations to dfGn can be used; among such simple models is the ARIMA(1,0,1) process.

In order to distinguish between short-term and long-term persistence processes, a new test referred to as the R_n/S analysis proposed by Mandelbrot and Wallis (1969e). A detailed study of R_n/S analysis is deferred to chapter 2.

1.4 Summary

The properties of hydrological processes and the assumptions necessary to initiate a generating scheme by using the theory of stochastic processes are reviewed. Moreover, various types of short-term and long-term persistent processes currently employed in synthetic hydrology

and their advantages and disadvantages are discussed. In addition to this, the methods for identifying a suitable process are briefly reviewed.

Chapter 2

RANGE, RESCALED RANGE AND HURST PHENOMENON

2.1. General

One of the fundamental problems in the design of a reservoir storage capacity is the investigation of its behaviour when subject to a given input sequence and a demand pattern. It is one of the purposes of this chapter to treat this problem in detail. The demand will always be assumed to be equal to either the population mean value, μ , or the sample mean, $\bar{x}$, of the input sequence. The case where μ is known, is an idealized situation.

The first treatment of such a problem to appear in the hydrological literature is due to Rippl (1883) who analyzed the cumulative mass curve of a historical record to find the reservoir size which would cope with the sequence of flows without causing any spillages or water shortages over the period of the historical record. The results obtained from such a mass curve provide useful information in the case of low yield reservoir coupled with the within year storages, but unfortunately its use is questionable for over-year storage calculations. The deficiencies of mass curve analysis have been cited by Fiering (1967) as follows,

- (1) Mass curve analysis has solely been based on the historical record whose recurrence in the future is highly unlikely.
- (2) Mass curve analysis does not provide any information as to the risks of water shortages during periods of low flows, and
- (3) Mass curve analysis yields a reservoir size which increases with increasing length of historical record considered. Moreover, the length of historical record is highly likely to differ from the actual projected life of the structure; therefore, the calculated reservoir size may not be compatible with the actual life of the reservoir.

The use of synthetic hydrology overcomes these deficiencies by generating an ensemble of synthetic flow sequences with length, equal to the economic life of reservoir. Each one of these sequences, can be routed through a mass curve analysis so that a set of representative reservoir sizes is obtained. From such a set, various statistical properties of reservoir size can be calculated. In this chapter the long-run average of reservoir size will be analytically treated for various currently used stochastic processes in hydrology, whereas for the verification of the derived analytical results, techniques of synthetic hydrology together with the sequent peak algorithm which is a sophisticated form of mass curve analysis will be used extensively. The formulation for evaluating the reservoir size using this approach has been developed by Wallis and Matalas (1972). Let the sequence of flows and demands be denoted by $x_1, x_2, \dots, x_n$ and $d_1, d_2, \dots, d_n$ respectively. Hence, the total inflow $I = \sum_{i=1}^n x_i$ and the total demand $D = \sum_{i=1}^n d_i$ are related to each other as

$$D = \alpha_1 I \quad (2.1)$$

where $0 < \alpha_1 < 1$ denotes the level of development. In the case of $\alpha_1 = 1$ there is full development, otherwise partial development is to be considered. Let minimum capacity required to meet all demands be denoted by C_m which corresponds to $\alpha_1 = 1$. If C_0 denotes the initial storage necessary to avoid storage deficiencies, then the magnitude of the storage at the end of the i -th time period, C_i , may be calculated as

$$C_i = \min \left[C_m, (x_i - d_i + C_{i-1}) \right] \quad (2.2)$$

and the waste water ω_i at time i is given by

$$w_i = m \alpha_i \times \left[0, (x_i - d_i - C_m^* C_{i-1}) \right] \quad (2.3)$$

Let the difference between the sum of inflows and demands at a time instant k , be given as

$$Z_k = \sum_{i=1}^k (x_i - d_i) \quad (2.4)$$

where Z_k represents water surplus when $Z_k > 0$, and in the case of $Z_k < 0$ a water deficit occurs. Hence, the sequence, $Z_1, Z_2, \dots, Z_n$ consisting of a mixture of surpluses and deficits is obtained. If the largest surplus and the greatest deficit are notationally represented by P_n and Q_n respectively, then

$$P_n = \max_{0 \leq k \leq n} Z_k \quad (2.5)$$

$$Q_n = \min_{0 \leq k \leq n} Z_k \quad (2.6)$$

The range of cumulative departures from the given demand pattern is defined as

$$R_n = P_n - Q_n \quad (2.7)$$

As noted earlier for $\alpha_i = 1$, $R_n = C_m$ and $C_0 = Q_n$ in which case there occurs no spillages or deficits over n years period.

The minimum design capacity, C_m , may be obtained by the sequent peak algorithm proposed by Thomas and Burden (1963) as follows. For the sake of operational convenience both the input and demand sequences are assumed to be extended over a second cycle which has no physical significance at all. Therefore, the operational time period becomes $2n$ instead of n , with

the extension of $x_i = x_{n+1}$ and $d_i = d_{n+1}$ where $i = 1, 2, \dots, n$. The necessary steps for the solution for C_m are as follows,

- (1) The differences $x_j - d_j$ are calculated for all $j = 1, 2, \dots, n, \dots, 2n$.
- (2) The net cumulative inflows, Z_j , are calculated for all $j = 1, 2, \dots, n, \dots, 2n$.
- (3) The sequent peaks $P_1, P_2, \dots, P_m$ are located such that $P_1 < P_2 < \dots < P_m$.
- (4) The sequence of troughs $T_1, T_2, \dots, T_s$ is located between sequent peaks, where $s = 1, 2, \dots, (m - 1)$.
- (5) The sequence of $P_s - T_s$ is formed. The minimum design capacity, C_m , is given

$$C_m = \max_s (P_s - T_s) \quad (2.8)$$

The random nature of inflow sequence causes C_m to be a random variable. The expectation of C_m for various stochastic models is one of the objectives of this chapter. Throughout the chapter, the level of development α_1 is assumed to be equal to one and the demand is considered as being equal to either the sample mean or the population mean value in which case C_m becomes identical to the range R_n . R_n has a very attractive interpretation in terms of hydrological design which involves the concept of ideal reservoir storage. The definition of the ideal reservoir has been given by Mandelbrot and Wallis (1969b) as a storage which fulfils the following four conditions,

- (1) that the outflow is uniform
- (2) that the reservoir ends the period as full as it started it
- (3) that the reservoir never overflows
- (4) that the capacity is the smallest compatible with conditions (1), (2), and (3).

Such an ideal reservoir storage for a future n years can suitably be designed when the data necessary to design it, are already available for n

years. The combined effect of all four conditions is that an ideal reservoir must get almost dry at some point in time and full at some other point in time.

In sections (2.2) - (2.3) all of the hydrological literature related to the adjusted range and the population range are reviewed. The definitions of Hurst's law, the Hurst coefficient and Hurst phenomenon are given in section (2.4). In this context, the estimates of Hurst coefficient are reviewed. A general analytical expression for the small sample expectation of the adjusted range is derived in section (2.5) by using a new methodology and in the subsequent sections the applications of this formula are illustrated for the case of the NIP, the lag-one Markov, the ARIMA(1,0,1) and the dfGn processes. Section (2.6) accounts for the general analytical expression of expected value of population range.

Section (2.7) gives the analytical expectations of both the rescaled adjusted range, R_n/S and rescaled population range R_p/S for various currently employed stochastic processes in the case of small sample sizes as well as in the case of the population situation. The relationships between the expectations of the population and the adjusted ranges and their rescaled counterparts are obtained and moreover it is shown that, asymptotically, the expected values for the ranges and rescaled ranges are identical whatever the underlying generating process is.

In section (2.9) firstly a literature review concerning R_n/S analysis is given and then on the basis of analytical expressions derived for $E(R_n/S)$ various deductions proposed by Mandelbrot and Wallis (1969e) are analytically confirmed; in this way it is shown quantitatively whether a hydrological process is short-term or long-term persistent.

Finally, in section (2.10) the analytical pdf of rescaled population range is asymptotically derived for the case of a NIP.

2.2 The Adjusted Range

For the purpose of theoretical considerations, the demand from a reservoir over any one year period can be assumed to be equal to the sample mean, $\bar{x}_n$, of the available sequence of input volumes of length n years. If this is the situation the total demand over the first k -year period can be written as

$$D_k = k\bar{x}_n \quad (2.9)$$

or, more explicitly

$$D_k = \frac{k}{n} \sum_{i=1}^n x_i \quad (2.10)$$

On the other hand, the amount of water flowing into the reservoir over the same period can be expressed as,

$$I_k = \sum_{i=1}^k x_i \quad (2.11)$$

Or, alternatively in terms of the mean, $\bar{x}_k$, of the first k flow values Eq.(2.11) can be re-written as,

$$I_k = k\bar{x}_k \quad (2.12)$$

where $\bar{x}_k$ is a function of the time period, k , and will be referred to as the successive mean.

Now, the amount of water that remains in the reservoir after the passage of the first k -year period is given as,

$$Z_k = I_k - D_k$$

or, by substituting Eq.(2.10) and Eq.(2.11) into this last expression, the following equation is obtained, that is

$$Z_k = \sum_{i=1}^k x_i - \frac{k}{n} \sum_{i=1}^n x_i \quad (2.13)$$

Or, by considering Eq.(2.9) and Eq.(2.12) a more revealing form of Eq.(2.13) can be written as

$$Z_k = k \bar{x}_k - k \bar{x}_n$$

or

$$Z_k = k(\bar{x}_k - \bar{x}_n) \quad (2.14)$$

which is a convenient form for later work in this chapter. It is apparent from Eq.(2.14) that the amount of water remaining in the reservoir after any time period, say k , is a function of the successive means, $\bar{x}_k$, of the inflow sequence. Since the inflow sequence is a random sample, its function $\bar{x}_k$, is also a random variable whose statistical properties can be related to the statistical properties of this inflow sequence. In the same way statistical behaviour of z_k can be evaluated in terms of the random variable $\bar{x}_k$.

With the notion used in this section the new forms of Eq.(2.5) and Eq.(2.6) are as

$$P_n = \max_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \quad (2.15)$$

and

$$Q_n = \min_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \quad (2.16)$$

Hence, by substituting Eq.(2.15) and Eq.(2.16) into Eq.(2.7) the range of cumulative departures from the sample mean, denoted by R_n , becomes,

$$R_n = \max_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] - \min_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \quad (2.17)$$

where R_n is referred to as the adjusted range. The statistical properties of R_n were first investigated by Feller (1951) and Hurst (1951) independently

for the case of NIP and the following asymptotic results were obtained,

$$E(R_n) = \sqrt{\frac{\pi}{2}} n \sigma \quad (2.18)$$

$$V(R_n) = \left(\frac{\pi^2}{6} - \frac{\pi}{2} \right) n \quad (2.19)$$

Furthermore, the asymptotic pdf of R_n was also derived by Feller (1951) by considerations from the theory of Brownian motion. Unfortunately, this pdf and Eq.(2.18) and Eq.(2.19) do not provide any practical applications because of the finite length of hydrological records. Therefore, the small sample properties of R_n needed to be evaluated. The original study towards this end is due to Solari and Anis (1957) who derived the small sample expectations of R_n in the case where the underlying process is normal and independent with zero mean and unit variance as

$$E(R_n) = \sqrt{\frac{n}{2\pi}} \sigma \sum_{i=1}^{n-1} i^{-1/2} (n-i)^{-1/2} \quad (2.20)$$

which has the asymptotic value given by Eq.(2.18) as derived by Hurst (1951) and Feller (1951) independently.

In a later section of this chapter Eq.(2.18) and Eq.(2.20) have been rederived using a different methodology. Moreover, the expectation of R_n in both asymptotic and small sample situations have been derived for a number of normal stationary stochastic processes currently used in hydrological literature.

2.3 The Population Range

In an ideal situation when the population mean, μ , is known in advance then a theoretical formulation of the range emerges where the demand can be taken to be equal to this population mean value. In such a situation the underlying r.v's can be transformed by applying a shifting operator,

to have a zero expectation. Throughout this chapter such a transformation is assumed to be performed on the original r.v. in the case of known population mean only, hence a zero population value introduces no loss of generality.

However, the amount of inflows over the first k-year period remains the same as it is provided by Eq.(2.11) or Eq.(2.12). Hence the amount of water that is impounded in the reservoir is given from Eq.(2.13) as

$$Z_k = \sum_{i=1}^k x_i$$

or, shortly

$$Z_k = k \bar{x}_k \quad (2.21)$$

The maximum and the minimum of variables Z_k over n time period are denoted by M_n and m_n respectively whose expressions are given as

$$M_n = \max_{0 \leq k \leq n} (k \bar{x}_k) \quad (2.22)$$

$$m_n = \min_{0 \leq k \leq n} (k \bar{x}_k) \quad (2.23)$$

In fact, in relation to the work of Anis and Lloyd (1953) M_n and m_n can be interpreted as the maximum and minimum of partial sums of random variables respectively. Finally, the definition of the population range, denoted as R_p , parallels the definition of the adjusted range and emerges as

$$R_p = M_n - m_n \quad (2.24)$$

or more explicitly

$$R_p = \max_{0 \leq k \leq n} (k \bar{x}_k) - \min_{0 \leq k \leq n} (k \bar{x}_k) \quad (2.25)$$

where R_p is the range of cumulative departures from the known population mean and is sometimes referred to as the population range.

According to Feller (1951) the sampling stability of R_p is better than that of R_n . Again the same author has studied the asymptotic statistical properties in great detail and he has given asymptotic expectation and the variance of R_p as

$$E(R_p) = 2 \sqrt{\frac{2}{\pi}} n \sigma \quad (2.26)$$

and

$$V(R_p) = 4n \left(\log 2 - \frac{2}{\pi} \right) \quad (2.27)$$

A theoretical derivation of the asymptotic pdf of R_p was also provided by him, but unfortunately none of these asymptotic results are directly applicable to hydrology where there are small samples available only.

However, in relation to the small sample properties of R_p the first study in this area was conducted by Anis and Lloyd (1953) for the case where the underlying variates are independent and normally distributed. They have shown the mean value of R_p for finite ^{sample} and independent variates with a common normal distribution is

$$E(R_p) = \sqrt{\frac{2}{\pi}} \sigma \sum_{i=1}^n i^{-1/2} \quad (2.28)$$

which for very large n values assumes its asymptotic value given in Eq.(2.26). A very nice interpretation of Eq.(2.28) is that it is the expected ideal reservoir size in the case where the demand is equal to the known population mean.

By manipulating Eq.(2.28) together with the incorporation of a clever

guess, Yevjevich (1967) succeeded in deriving the expected population range value for normal stationary processes in general. In particular $E(R_p)$ for the lag-one Markov process is given by him as,

$$E(R_n) = \sqrt{\frac{2}{\pi}} \sigma \sum_{i=1}^n i^{-1/2} \left[\frac{1+\rho_1}{1-\rho_1} - \frac{2\rho_1(1-\rho_1^i)}{i(1-\rho_1)^2} \right]^{1/2} \quad (2.29)$$

where ρ_1 is the first order serial correlation coefficient. When $\rho_1 = 0$ is substituted in Eq.(2.29) then it reduces to Eq.(2.28) which is the NIP case. Yevjevich has verified Eq.(2.29) through the use of Monte Carlo techniques. A logical derivation of Eq.(2.29) will be presented in section (2.6.2)

2.4 Hurst's Law, Hurst Coefficient and Hurst Phenomenon

Out of the original works of Hurst (1951,1956) in relation to the long-term storage requirements on the Nile river, emerged one of the most important statistics of streamflow which is given from Eq.(1.11) as

$$\frac{R_n}{S} \sim n^h$$

where S denotes the sample standard deviation of the time series of length n, R_n has been previously defined by Eq.(2.17) and h is a constant power. The expression given by Eq.(1.11) has been empirically developed on the basis of 900 annual time series comprising streamflow and precipitation records, stream and lake levels, tree rings, mud varves, atmospheric pressure and sunspots.

In general, the statistic R_n/S has been referred to as the rescaled range by Mandelbrot and Wallis (1969b) without making any distinction whether adjusted or population range is used in Eq.(1.11). Such a distinction will be strictly observed in this thesis. Hence, if the adjusted range of Eq.(2.17) is adopted, then the corresponding rescaled range will be called rescaled adjusted range, whose explicit form can be obtained by dividing both sides

of Eq.(2.17) by S which gives,

$$\frac{R_n}{S} = \frac{1}{S} \left\{ \max_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \right\} - \frac{1}{S} \left\{ \min_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \right\}$$

or, equivalently

$$\frac{R_n}{S} = \max_{0 \leq k \leq n} \left[k \frac{\bar{x}_k - \bar{x}_n}{S} \right] - \min_{0 \leq k \leq n} \left[k \frac{\bar{x}_k - \bar{x}_n}{S} \right] \quad (2.30)$$

On the other hand if Eq.(2.25) is adopted as the definition of the range, then, the resulting rescaled range will be referred to as the rescaled population range of which the general expression convenient for later work in this study is given as

$$\frac{R_D}{S} = \max_{0 \leq k \leq n} \left(k \frac{\bar{x}_k}{S} \right) - \min_{0 \leq k \leq n} \left(k \frac{\bar{x}_k}{S} \right) \quad (2.31)$$

The general expression given in Eq.(1.11) is usually referred to as Hurst's law, and in addition to this, the constant power, h, has been named as the Hurst coefficient. This coefficient, h, is a function of the rescaled range which is a random variable, therefore h itself is a random variate. As a result, theoretically h should have its own pdf and in turn, descriptive parameters such as the expectation, variance etc. Moreover, like any other statistical coefficient in stochastic hydrology, an estimator for h is inevitable. So far in the hydrological literature there exists two types of estimates of h. One was proposed by Hurst (1951), where he wrote Eq.(1.11) in an equality form by assuming the proportionality factor to be equal to $1/2^K$. Hence,

$$\frac{R_n}{S} = \left(\frac{n}{2}\right)^K \quad (2.32)$$

where K denotes the resulting estimate of the population coefficient, h.

When Eq.(2.32) is solved for K, then the estimate of h for a given historic record becomes,

$$K = \frac{\log(R_n/S)}{\log(n/2)} \quad (2.33)$$

or

$$K = \frac{\log R_n - \log S}{\log n - \log 2} \quad (2.34)$$

For the 900 annual time series investigated by Hurst (1951) the mean value of K was found to be 0.73 and the standard deviation as 0.092 over all series. The lengths of these time series varied from 30 to 2000 years. Mandelbrot and Wallis (1968, 1969d) have pointed out the shortcoming of Hurst's estimator K, that is, only the full length of sample is used to estimate h, and assumes that the line defining the estimate of h always passes through the point of abscissa log2 and ordinate log1 which is equal to zero. A more general alternative estimation procedure of h has been given by Mandelbrot and Wallis (1969b,d). This estimate is denoted by H and the method of estimation has been given by the same authors. Wallis and Matalas (1970) have made a series of extensive computer simulation experiments to assess the small sample properties of both K and H. According to their work, for

a Gaussian independent process both K and H are biased estimators; although K has relatively greater bias than H , its variance is found to be smaller than H . However, the effect of bias on both estimators slightly decreases with increasing sample size.

According to the value that h assumes the hydrological processes can broadly be divided into two classes. Those processes which yield a value of h equal to 0.5, are said to be short-term persistent and processes giving rise to a value of h in the range $0.5 < h < 1$, are said to be long-term persistent. The tendency of natural time series to yield h values greater than 0.5 is referred to as Hurst phenomenon. The difference between the average value of K , 0.73, and the exponent in Eq.(2.18), 0.5, puzzled many hydrologists and statisticians at the time, and following this a number of suggestions were put forward as to the cause of discrepancy. In general, three explanations of the Hurst phenomenon have been offered

- (1) marginal distribution
- (2) transience
- (3) autocorrelation

The first of these explanations can be discarded in the light of a work by Matalas and Huzzzen (1967) who carried out an exhaustive computer experiment to obtain the expected value of estimate of K , $E(K)$, for the normal and, then, a group of log-normal distribution function with different skewness coefficient γ in the range $0.2 \leq \gamma \leq 2$. They found that skewness had not a significant influence on the values of $E(K)$ and therefore concluded that skewness does not explain the Hurst phenomenon, hence confirming the deduction by Langbein (1956) who stated that skewness was an unlikely explanation as was serial correlation of a Markovian nature. Moreover, Langbein suggested that the autocorrelation structure of data used by Hurst was of a very complex nature.

However, the theoretical works, concerning the small sample properties of adjusted and population ranges, by Anis and Lloyd (1953), Anis (1955, 1956)

and Solari and Anis (1957) are not sufficient to explain the Hurst phenomenon, because all of these theoretical studies were based on an underlying normal independent processes for which $h=0.5$. In addition, Hurst had considered the statistic R_n/S rather than R_n or R_p , whose expectations are not the same at least for small samples. This assertion will be analytically confirmed later in this chapter. It was deduced by Moran (1964) that for moderate n values, the Hurst phenomenon could be explained by a highly skewed distribution with very large second order moments about the mean value. Unfortunately, he was concerned with R_p , therefore the same conclusions could not be valid for R_n/S . Both Lloyd and Moran have long been considering the behaviour of either R_n or R_p for a possible explanation of Hurst's original law, where, in fact Hurst worked with the rescaled range, R_n/S only.

However, to the best of the author's knowledge there has not yet appeared any analytical expression as to the expectation of R_n/S in the hydrologic literature so far. In later sections, the expected values of both rescaled^{and} adjusted range will be analytically derived for models currently used in hydrology for simulating annual streamflow sequences. The analytical expressions obtained for $E(R_n/S)$ will replace the expensive data generating method for explaining the behaviour of $E(R_n/S)$. The same expressions make it possible to apply R_n/S analysis analytically to determine whether a process is short-term or long-term persistent.

2.5 Small Sample Expectation of the Adjusted Range. (R_n)

The general expression for the adjusted range has already been given in section (2.2) by Eq.(2.17) as

$$R_n = \max_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] - \min_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)]$$

For a normally distributed underlying random variable, $\bar{x}_k$ and $\bar{x}_n$ are also

normally distributed, in general, as $N(0, \sigma_{\bar{x}_k}^2)$ and $N(0, \sigma_{\bar{x}_n}^2)$ respectively. The mean values of these random variates, namely, $\bar{x}_k$ and $\bar{x}_n$ are equal to zero because of the assumption that the underlying random variate is shifted to the origin. In this section interest lies with the expected value of the adjusted range, therefore taking expectations of both sides of Eq.(2.17) leads to,

$$E(R_n) = E\left\{ \max_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \right\} - E\left\{ \min_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \right\} \quad (2.35)$$

or, because of the symmetric property of normal pdf the following equality, in expectation, is valid, that is,

$$E\left\{ \max_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \right\} = -E\left\{ \min_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \right\}$$

Hence, by substituting this expression into Eq.(2.35)

$$E(R_n) = 2E\left\{ \max_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \right\} \quad (2.36)$$

It is clear from this last expression that when the expected value of r.v. $\max_{0 \leq k \leq n} k(\bar{x}_k - \bar{x}_n)$ over the time period (0-n) is known, then $E(R_n)$ follows immediately. Let this r.v. be denoted by y_n , and hence

$$y_n = \max_{0 \leq k \leq n} [k(\bar{x}_k - \bar{x}_n)] \quad (2.37)$$

where subscript n, in y_n , emphasizes the fact that y_n is the maximum cumulative departure from sample mean over n-year period for a given time

series. Furthermore, let the r.v. whose maximum is sought be denoted by,

$$z_{k,n} = k(\bar{x}_k - \bar{x}_n) \quad (2.38)$$

where subscripts k and n denote the dependence of r.v. $z_{k,n}$ on two time instances. A first glance at Eq.(2.38) reveals that both of the terms, $k\bar{x}_k$ and $k\bar{x}_n$, on the right-hand side are normally distributed variables whatever the values of k and n are, but if the underlying r.v. is non-Gaussian then in the light of the central limit theorem the two terms are again normally distributed but for large values of k and n only. As a result of this fact, their difference $z_{k,n}$ is normally distributed, hence the mean and the variance of z will suffice to describe this normal distribution.

Therefore, by taking the expectation of Eq.(2.38)

$$E(z_{k,n}) = k[E(\bar{x}_k) - E(\bar{x}_n)]$$

where $E(\bar{x}_k)=0$ and $E(\bar{x}_n)=0$, hence

$$E(z_{k,n}) = 0$$

and the variance, in general, turns out to be

$$V(z_{k,n}) = k^2 \left[V(\bar{x}_k) - 2\text{Cov}(\bar{x}_k, \bar{x}_n) + V(\bar{x}_n) \right] \quad (2.39)$$

where $V(\bar{x}_k)$, $V(\bar{x}_n)$ and $\text{Cov}(\bar{x}_k, \bar{x}_n)$ are all dependent on the nature of the

hydrological process adopted.

The methodology applied here for finding the expectation of the maximum of r.v. $z_{k,n}$ over the n-year period is as follows. The two most important facts which constitute the basis of the methodology pursued are,

(1) that the discrete r.v.'s, can be thought of being sampled at regular intervals from the corresponding continuous r.v. Therefore, the discrete cumulative variable $z_{k,n}$ is obtained from the continuous cumulative variable, $z(k,n)$ by sampling at regular intervals as shown in Figure (2.1). Moreover, as the length of interval goes to zero, the discrete variable converges to its continuous counterpart.

(2) that the situation of the maximum of $z_{k,n}$ for a given historic record has associated with it two uncertainties. One of these uncertainties is connected with the time instant at which the maximum occurs whereas the other source of uncertainty is the magnitude of the maximum.

As a result there are two jointly distributed r.v.'s for completely describing the position of the maximum. Let this joint pdf be denoted as $f(m',t')$ where m' is the r.v. associated with the uncertain magnitude of maximum and t' is the r.v. associated with the time instant. Hence, the expected value of this joint pdf can be written as,

$$E(M) = \int_0^{\infty} \int_0^{\infty} m' t' f(m',t') dm' dt'$$

or, alternatively

$$E(M) = \int_0^{\infty} E(M|t') f(t') dt' \quad (2.40)$$

where $E(M|t')$ denotes the conditional expectation of the maximum occurring at

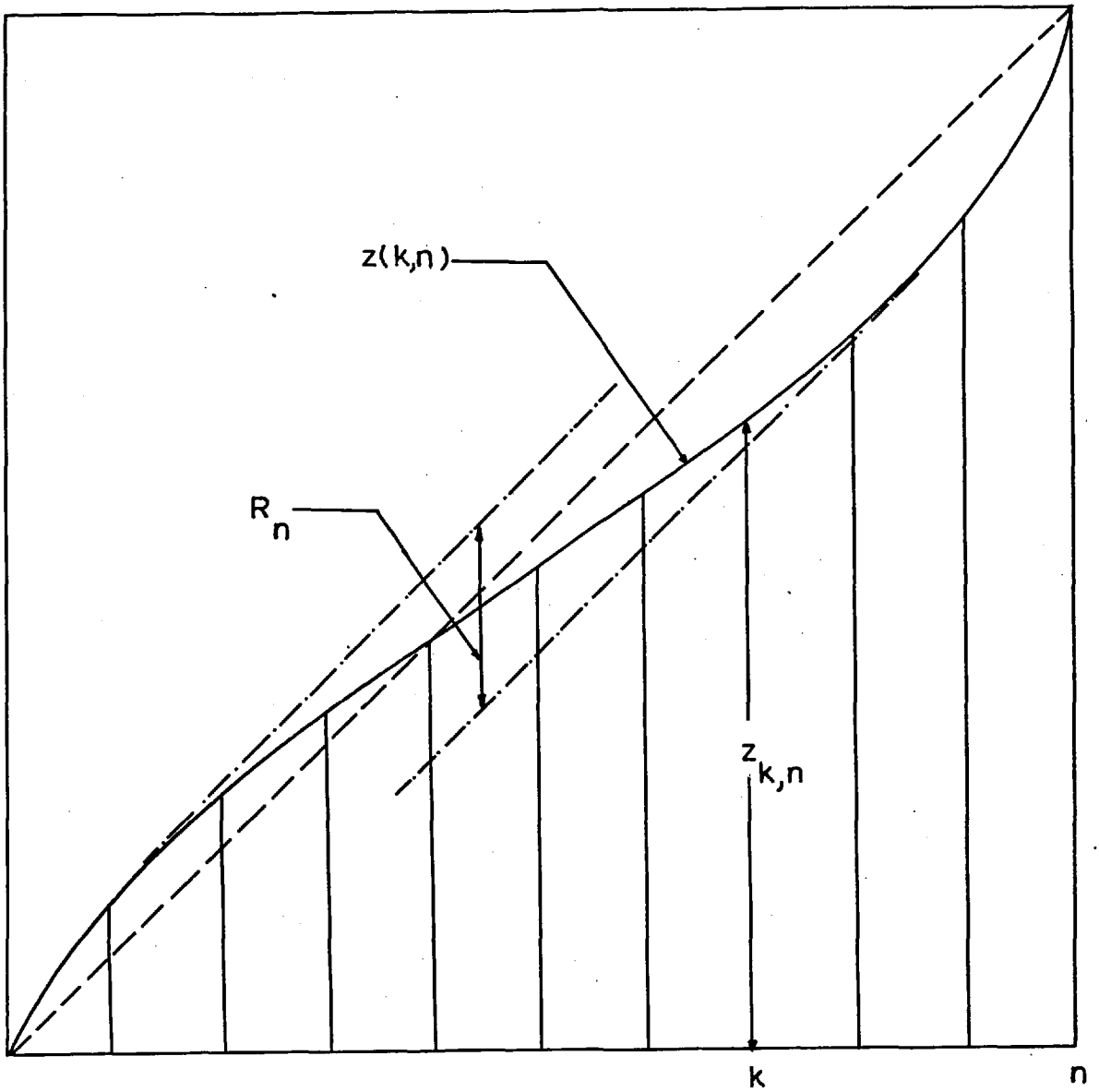


Figure (2.1) - The sampling of discrete cumulative variables, $z_{k,n}$ from a continuous cumulative variable $z(k,n)$ at regular time intervals.

a given time instant. Furthermore, $E(M|t')$ is the expectation obtained from the cumulative departures of a continuous r.v. In this section, this continuous r.v. is considered to be given by Eq.(2.38) whose pdf can be written as

$$f(z_{k,n}) = \frac{1}{\sqrt{2\pi V(z_{k,n})}} \exp \left\{ - \frac{1}{2V(z_{k,n})} z_{k,n}^2 \right\} \quad (2.41)$$

For any fixed value of n , this pdf represents a kind of Brownian motion, with zero mean and variance given by Eq.(2.39) for successive k values. According to Lévy (1965) the following sequence of equalities are valid for any symmetric distribution function, that is

$$\Pr(M > \xi) = 2\Pr(M > \xi, x > \xi) = 2\Pr(x > \xi) = \Pr(|x| > \xi) \quad (2.42)$$

where x represents a Brownian motion and M is the maximum of this variable over a definite period of time. By considering the first and last terms in this string of equalities, it can be concluded that the pdf of maxima is the same as the pdf of the absolute value of the r.v. In this section only the normal distribution is considered therefore the pdf of maxima is the absolute normal distribution which is given by Lévy (1965) as

$$f(x) = \frac{2}{\sqrt{2\pi t}} \exp \left(\frac{-x^2}{2t} \right) \quad 0 < x < \infty \quad (2.43)$$

An assumption regarding Eq.(2.43) is that it is considered as being a conditional pdf of a given time period t moreover it does not alter the quantitative value of Eq.(2.43) whether t is thought as a time instant or time period. Therefore, the above referred conditional pdf can be written

from Eq.(2.43) as

$$f(x|t) = \frac{2}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right) \quad (2.44)$$

In general, the relationship between a normal r.v. x and absolute normal r.v. z is given as

$$f(z) = 2f(x) \quad (2.45)$$

where $-\infty < x < \infty$ and $0 < z < \infty$. Therefore, the absolute pdf corresponding to Eq.(2.41) turns out to be,

$$f(|z|_{k,n}) = \frac{2}{\sqrt{2\pi V(z_{k,n})}} \exp\left\{-\frac{1}{2V(z_{k,n})} z_{k,n}^2\right\} \quad (2.46)$$

The moments of $z_{k,n}$ have been given by Papoulis (1965) as

$$E(|z|_{k,n}^l) = \begin{cases} 1.3 \dots (n-1) [V(z_{k,n})]^{1/2} & \text{for } l=2m \\ \sqrt{\frac{2}{\pi}} 2^m m! [V(z_{k,n})]^{(2m+1)/2} & \text{for } l=2m+1 \end{cases} \quad (2.47)$$

Hence, the expectation of Eq.(2.46) can be given from Eq.(2.47) with $l=1$, as

$$E(|z|_{k,n}) = \sqrt{\frac{2}{\pi}} V(z_{k,n}) \quad (2.48)$$

This is a conditional expectation given k and n . So, at each time instant $k < n$ there is an expectation associated with it given by Eq.(2.48). One assumption at this stage is that, the occurrence of the maximum associated

with different time points are independent from each other. Therefore, the probability of occurrence of maxima at each time instant can be constructed as follows. At the first time instant there can occur two events only, one is a maximum and the other is the minimum and these can be represented as in figure (2.2)

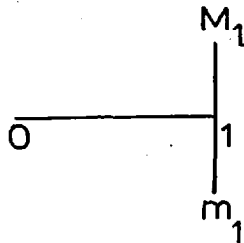


Figure (2.2)

Therefore, the probability associated with the maximum at time instant $k=1$ is

$$P_1 = \frac{1}{2.1}$$

when two time instances are considered then figure (2.3) results

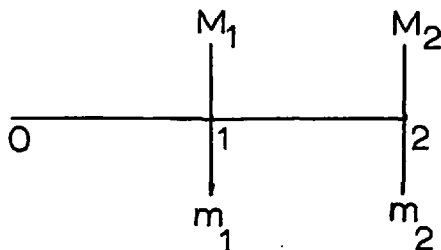


Figure (2.3)

In this situation there are four possible events and the probability of maximum occurring at $k=2$ is given as

$$P = \frac{1}{2.2}$$

In general, when n time instances are considered as illustrated in figure (2.4)

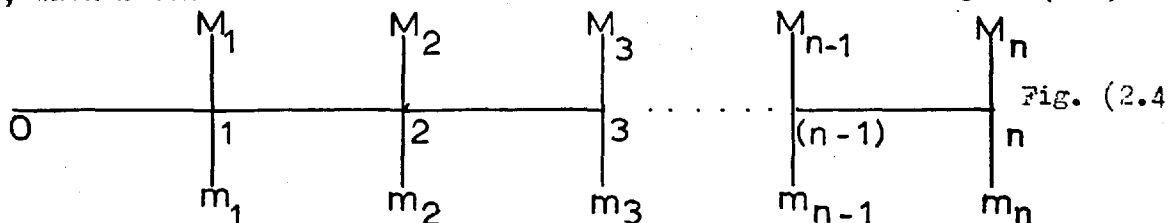


Fig. (2.4)

then there are $2n$ equally likely events and the occurrence of the greatest maximum at time instant n , has a probability of

$$P_n = \frac{1}{2 \cdot n}$$

On the other hand Eq.(2.40) can be written for discrete variables and new variables as

$$E(M) = \sum_{k=1}^n E(|z_{k,n}|) P_k \quad (2.49)$$

after substituting Eq.(2.48) and the above calculated probabilities into Eq.(2.49) leads to

$$E(M) = \frac{1}{2} \sqrt{\frac{2}{\pi}} \sum_{k=1}^n \sqrt{\frac{V(z_{k,n})}{k^2}} \quad (2.50)$$

or, because of the fact that

$$E(R_n) = 2E(M) \quad (2.51)$$

insertation of Eq.(2.50) into Eq.(2.51) yields,

$$E(R_n) = \sqrt{\frac{2}{\pi}} \sum_{k=1}^n \sqrt{\frac{V(z_{k,n})}{k^2}} \quad (2.52)$$

This is the general expression for the expectation of adjusted range for small samples and it is valid for any type of hydrological process provided that it is stationary and normal irrespective of its autocorrelation structure because in the derivation of Eq.(2.52) no assumption as to whether the

underlying process is independent or dependent has been made. So far, to the best of the author's knowledge, no such general expression for the expectation of the adjusted range has appeared in the hydrological literature. As an explanation of Eq.(2.52) results quoted earlier given by Eq.(2.28) and Eq.(2.29) will be obtained in the following sections.

A final remark at this point is that the sum of probabilities, P_k , in Eq.(2.49) does not add up to one when overall averaging is considered but if figures (2.2) - (2.4) are considered individually then one can see that the probabilities, P_k , are properly defined.

2.5.1 The Normal Independent Process. (NIP)

In relation to the Hurst phenomenon, all of the theoretical studies of the range, in general, have been based on the NIP which corresponds to the value of $h=0.5$. If the inflows into a reservoir have the nature of NIP then the expected adjusted range in this special case can easily be worked out from the general expression of Eq.(2.52). To do this first of all the variance in Eq.(2.39) has to be found. For the case of NIP the following equations are valid

$$V(\bar{x}_k) = \frac{\sigma^2}{k}$$

$$V(\bar{x}_n) = \frac{\sigma^2}{n}$$

and

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{\sigma^2}{n}$$

where σ^2 is the variance of the input sequence. The substitution of these last three equations into Eq.(2.39) leads to

$$V(z_{k,n}) = k \frac{n-k}{n} \sigma^2 \quad (2.53)$$

and furthermore the substitution of Eq.(2.53) into Eq.(2.52) gives the required result which is expressed as,

$$E(R_n) = \sqrt{\frac{2}{\pi}} \sigma \sum_{k=1}^n \left(\frac{n-k}{nk} \right)^{1/2} \quad (2.54)$$

or by considering the following equality

$$2 \sum_{k=1}^n \left(\frac{n-k}{nk} \right)^{1/2} = \sqrt{n} \sum_{k=1}^{n-1} [k(n-k)]^{-1/2} \quad (2.55)$$

then Eq.(2.54) becomes,

$$E(R_n) = \sqrt{\frac{n}{2\pi}} \sigma \sum_{k=1}^{n-1} k^{-1/2} (n-k)^{-1/2} \quad (2.56)$$

This is the expected value of the adjusted range for small samples of NIP which was derived by Solari and Anis (1957) as was previously given by Eq.(2.20). The same authors have also shown that the asymptotic value of Eq.(2.56) i.e. its behaviour for large n becomes the asymptotic adjusted range given by Feller (1951). In order to obtain this conclusion the summation in the Eq.(2.56) is replaced by the approximate integral, that is,

$$\int_1^{n-1} k^{-1/2} (n-k)^{-1/2} dk$$

On making the substitution $n\theta = k$ this integral becomes,

$$\int_{1/n}^{1-1/n} \theta^{-1/2} (1-\theta)^{-1/2} d\theta$$

which converges to $B(\frac{1}{2}, \frac{1}{2}) = \pi$ as n tends to infinity. Thus the expectation of the asymptotic adjusted range is given as,

$$E(R_n) = \sqrt{\frac{\pi}{2}} n \sigma$$

This last expression shows the rederivation of an already known formula due to Feller (1951) and it is given by Eq.(2.18). When a short time span is considered to be the economic life of a reservoir then Eq.(2.56) will give the size of the ideal reservoir with independent sequences.

2.5.2 The Lag-One Markov Process.

The simplest model used in hydrology to preserve the interdependence among the successive observations of streamflow is the lag-one Markov process. Interdependence among observations exerts great effect on the final product, such as the reservoir size, in any water resource system design. It will be analytically proved later in this chapter that in the case of correlated inputs the capacity tends to be greater than the one where the inflows are independent. As a result of this property a simulation process of streamflow might result in an overdesign or underdesign relative to an optimal design.

The autocorrelation function of the lag-one Markov process is given as

$$\rho_k = \rho^k \quad (2.57)$$

With this autocorrelation function the terms necessary to find the expected

adjusted range through employing the general formula given in Eq.(2.52) are derived in appendix (2.1) as

$$V(\bar{x}_k) = \frac{\sigma^2}{k^2} \left\{ k + \frac{2\rho}{(1-\rho)^2} \left[k(1-\rho) - (1-\rho^k) \right] \right\} \quad (2.58)$$

$$V(\bar{x}_n) = \frac{\sigma^2}{n^2} \left\{ n + \frac{2\rho}{(1-\rho)^2} \left[n(1-\rho) - (1-\rho^n) \right] \right\} \quad (2.59)$$

and

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{\sigma^2}{nk} \sum_{j=1}^k \sum_{i=1}^n \rho^{|i-j|} \quad (2.60)$$

By inserting these three terms into Eq.(2.39), $V(z_{k,n})$ is derived which in turn may be inserted Eq.(2.52) to give the expected adjusted range in small samples for the lag-one Markov process. The verification of this analytic solution has been achieved by computer simulation techniques, and very good agreement between the analytical and simulation results have been observed as is evident from figure (2.5). The same analytical result gives the expected value of the ideal reservoir size in the case of Markovian inflows with an unknown population mean value. It can also be observed in figure (2.5) that the $E(R_n)$ of the lag-one Markov process is always greater than that of the NIP case. When $\rho=0$ is substituted in $E(R_n)$ then Eq.(2.20) which is valid for the NIP case is obtained.

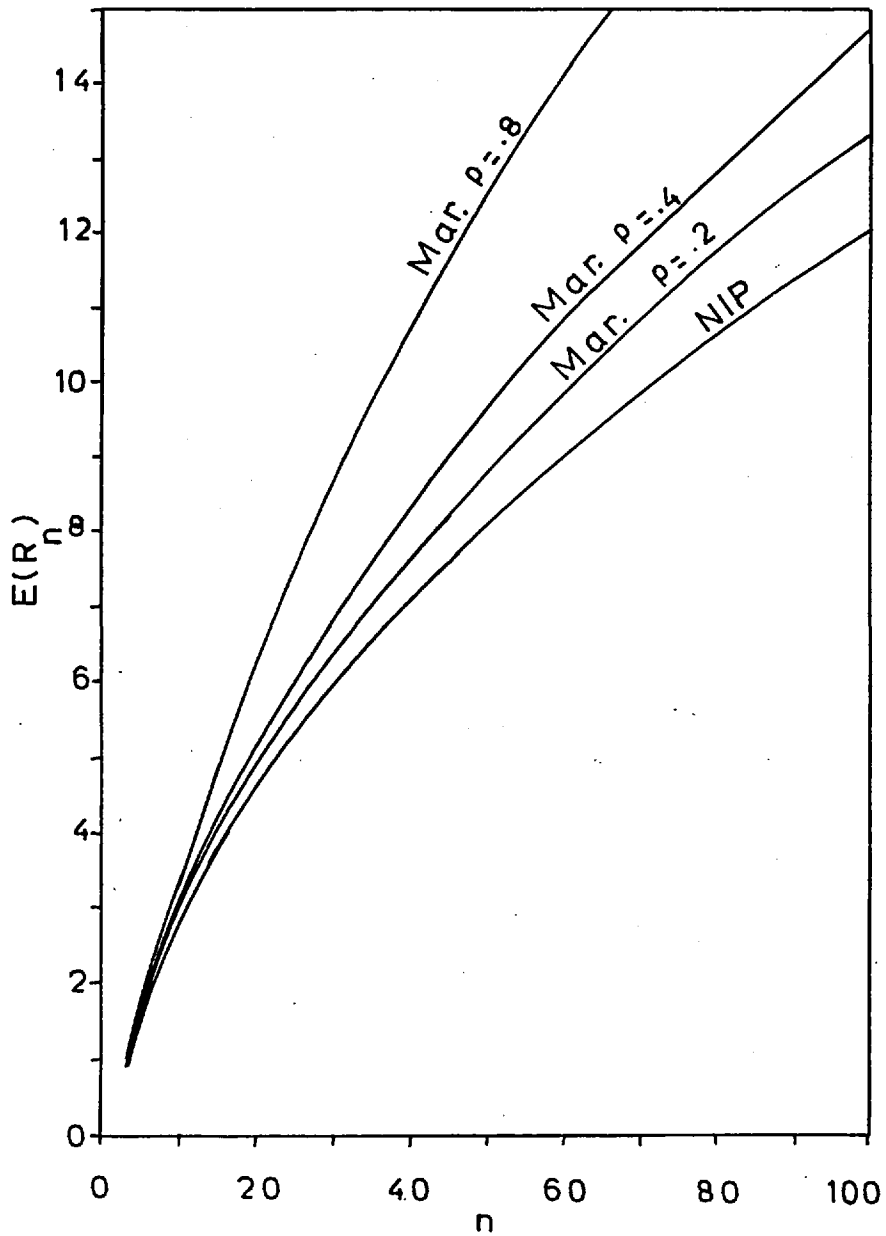


Figure (2.5) - The expected adjusted range of various processes.

2.5.3 The ARIMA(1,0,1) Process.

Many of the existing processes which have appeared recently in the hydrological literature have originated from the desire to preserve Hurst's coefficient exactly or approximately. The ARIMA(1,0,1) process which has been proposed by O'Connell (1971) for simulating annual streamflow sequences provides such an approximation and simple process for modelling Hurst's law.

By using Eq.(1.21) and Eq.(1.22) in the algebraic development of appendix (2.2) the following terms necessary for finding the expected adjusted range for the ARIMA(1,0,1) process are obtained.

$$V(\bar{x}_k) = \frac{\sigma^2}{k^2} \left\{ k + \frac{2\rho}{(1-\phi)^2} \left[k(1-\phi) - (1-\phi^k) \right] \right\} \quad (2.61)$$

$$V(\bar{x}_n) = \frac{\sigma^2}{n^2} \left\{ n + \frac{2\rho}{(1-\phi)^2} \left[n(1-\phi) - (1-\phi^n) \right] \right\} \quad (2.62)$$

and

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{\sigma^2}{nk} \left\{ k + \rho \sum_{i=1}^k \sum_{j=1}^n \phi^{|i-j|} \right\} \quad (2.63)$$

$i \neq j$

The substitution of these final equations into Eq.(2.52) yields the desired result for $E(R_n)$. Hence, it is clear that $E(R_n)$ is obtained merely when $\theta = 0$ and information on the autocorrelation structure is available only. When $\phi = \rho$ is substituted in the expression for $E(R_n)$ then it reduces to that for lag-one Markov process case and the substitution of $\phi = \theta$ gives the NIP case for which $E(R_n)$ is given by Eq.(2.20).

2.5.4. Fractional Gaussian Noise

For the theoretical fractional Gaussian process the following equations are given by Mandelbrot and Wallis (1969b)

$$V(\bar{x}_k) = C_h k^{2h-2} \quad (2.64)$$

$$V(\bar{x}_n) = C_h n^{2h-2} \quad (2.65)$$

In the light of these expressions the following quantity can easily be evaluated

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{C_h}{k n} E \left[\left(\sum_{i=1}^k x_i \right) \left(\sum_{i=1}^n x_i \right) \right] \quad (2.66)$$

or

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{C_h}{k n} E \left[\left(\sum_{i=1}^k x_i \right)^2 + \left(\sum_{i=1}^k x_i \right) \left(\sum_{i=k+1}^n x_i \right) \right]$$

or finally

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{C_h}{2kh} \left[n^{2h} (n-k)^{2h} + k^{2h} \right] \quad (2.67)$$

The incorporation of Eq.(2.64), Eq.(2.65) and Eq.(2.67) into Eq.(2.52) yields

$$E(R_n) = \sqrt{\frac{2}{\pi}} C_h \sum_{k=1}^n \left[n^{2h-2} - \frac{n^{2h-1}}{k} + \frac{(n-k)^{2h}}{k n} - \frac{k^{2h-1}}{n} + k^{2h-2} \right]^{1/2} \quad (2.68)$$

where C_h is the standard deviation of fGn. For $h=0.5$ Eq.(2.68) reduces to the case of NIP for which $E(R_n)$ has been given earlier by Eq.(2.56).

2.6 Small Sample Expectation of the Population Range. (R_p)

An idealized situation occurs when the population mean, μ , of a time series is assumed to be known. In such a case the assumption that the population mean is equal to zero causes no loss of generality and the general expression of R_p is given by Eq.(2.25) as

$$R_p = \max_{0 \leq k \leq n} (k \bar{x}_k) - \min_{0 \leq k \leq n} (k \bar{x}_k)$$

where $\bar{x}_k$ is distributed according to $N(0, \sigma_{\bar{x}_k}^2)$. A first glance at Eq.(2.17) shows that Eq.(2.25) is a special case of it if $\bar{x}_n$ is set equal to zero. Hence, all of the formula obtained for $E(R_p)$ in the previous sections will be valid when $\bar{x}_n = 0$ is inserted into them. Therefore, the expectation of R_p can be obtained from Eq.(2.36) as

$$E(R_p) = 2E \left[\max_{0 \leq k \leq n} (k \bar{x}_k) \right] \quad (2.69)$$

In the same manner the counterpart of Eq.(2.39) in this case will be given as

$$V(z_k) = k^2 V(\bar{x}_k) \quad (2.70)$$

the subscript n has been dropped out because of the fact that the sample mean is assumed to be known and independent of the sample length n . The same methodology applied in section (2.5) remains applicable in this situation. Hence, from Eq.(252) the general expected value of R_p can be obtained simply by replacing $V(z_{k,n})$ by $V(z_k)$. Hence, in general,

$$E(R_p) = \sqrt{\frac{2}{\pi}} \sum_{k=1}^n \sqrt{\frac{V(z_k)}{k^2}}$$

Or, substitution of Eq.(2.70) in the above expression yields

$$E(R_p) = \sqrt{\frac{2}{\pi}} \sum_{k=1}^n \sqrt{V(\bar{x}_k)} \quad (2.71)$$

This result is applicable to any stationary normal hydrological process. The same equation has been given by Yevjevich (1967) on the basis of a clever guess and manipulating Eq.(2.28) which was first derived by Anis and Lloyd (1953).

2.6.1 The Normal Independent Process

After having got the general expression in Eq.(2.71) it is , then, straightforward to make an application of it to various hydrological processes. The only parameter is the variance of the subsample mean, $V(\bar{x}_k)$, which for the NIP is given in section (2.5.1) by Eq.(2.53) and by inserting this equation into Eq.(2.71), $E(R_p)$ becomes

$$E(R_p) = \sqrt{\frac{2}{\pi}} \sigma \sum_{k=1}^n k^{-1/2} \quad (2.72)$$

which has been derived by Anis and Lloyd (1953) and given in Eq.(2.28).

However, the asymptotic expected value of R_p has been given by Feller (1951). In order to obtain the same result, the summation in Eq.(2.72) is replaced by the appropriate integral given as

$$\int_0^n k^{-1/2} dk = \left| -2k^{1/2} \right|_0^n$$

which becomes equal to $2\sqrt{n}$ and therefore

$$E(R_p) = 2 \sqrt{\frac{2}{\pi}} \sigma \sqrt{n}$$

This is the expression given in Eq.(2.26).

2.6.2 The Lag-One Markov process.

The variance of the mean of a lag-one Markov process sample of length , k, has been given in Eq.(2.58), the substitution of which into Eq.(2.71) gives

$$E(R_p) = \sqrt{\frac{2}{\pi}} \sigma \sum_{k=1}^n k^{-1/2} \left[\frac{1+p}{1-p} - \frac{2p(1-p^k)}{k(1-p)^2} \right]^{1/2} \quad (2.73)$$

This expression has been derived previously by Yevjevich (1967) who has shown its correctness by exhaustive computer experiments.

The asymptotic expected population range can be obtained from Eq.(2.73) by expanding the second factor under the summation sign, into a Binomial series, hence

$$E(R_p) = \sqrt{\frac{2(1+p)}{\pi(1-p)}} \sigma \sum_{k=1}^n k^{-1/2} \left[1 - \frac{2p(1-p^k)}{k(1-p)^2} \right]^{1/2}$$

or

$$E(R_p) = \sqrt{\frac{2(1+p)}{\pi(1-p)}} \sigma \sum_{k=1}^n k^{-1/2} \left[1 - \frac{p(1-p^k)}{k(1-p)^2} + \frac{4p^2(1-p^k)^2}{3!k^2(1-p)^2} \dots \right]$$

by ignoring second and subsequent terms in the brackets, the following formula is obtained for very large n values,

$$E(R_p) = \sqrt{\frac{2(1+p)}{\pi(1-p)}} \sigma \sum_{k=1}^n k^{-1/2}$$

by applying the appropriate integral to the summation in this last expression as applied in the previous section, the following asymptotic value results,

$$E(R_p) = 2 \sqrt{\frac{2(1+\rho)}{\pi(1-\rho)}} n \sigma \quad (2.74)$$

A simple relationship between the asymptotic expected population ranges of the lag-one Markov process and NIP can be given by comparing Eq.(2.74) and Eq.(2.28) as,

$$E(R_{p_M}) = \sqrt{\frac{1+\rho}{1-\rho}} E(R_{p_I}) \quad (2.75)$$

where subscripts M and I refer to the lag-one Markov and NIP respectively. The factor is always greater than unity provided that $\rho > 0$, as a result $(1+\rho) > 1$ and $(1-\rho) < 1$. Hence,

$$\sqrt{\frac{1+\rho}{1-\rho}} > 1$$

Therefore, Eq.(2.75) analytically confirms the fact that the size of a reservoir in the case of dependent inflow sequences is greater than that of independent inflow sequence.

2.6.3 The ARIMA(1,0,1) Process

Again the asymptotic $E(R_p)$ for this process will be obtained from the general expression of Eq.(2.71) after substituting Eq.(2.64) into it, which yields

$$E(R_p) = \sqrt{\frac{2}{\pi}} \sigma \sum_{k=1}^n k^{-1/2} \left[\frac{1-\phi+2\rho}{1-\phi} - \frac{2\rho(1-\phi^k)}{k(1-\phi)^2} \right]^{1/2} \quad (2.76)$$

which is valid for small samples. The behaviour of Eq.(2.76) can be approximately found by the application of the Binomial expansion to the brackets on the right-hand side, hence,

$$E(R_p) = \sqrt{\frac{2}{\pi}} \sigma \sqrt{\frac{1-\phi+2\rho}{1-\phi}} \sum_{k=1}^n k^{-1/2} \left[1 - \frac{2\rho(1-\phi^k)}{k(1-\phi+2\rho)(1-\phi)} \right]^{1/2}$$

or, after ignoring second and subsequent terms in the expansion and applying the appropriate integral as in the previous section, the asymptotic expected population range of the ARIMA(1,0,1) process is obtained as

$$E(R_p) = 2 \sqrt{\frac{2(1-\phi+2\rho)}{\pi(1-\phi)}} k \sigma \quad (2.77)$$

The comparison of this equation with Eq.(2.28) yields

$$E(R_{pA}) = \sqrt{1 + \frac{2\rho}{1-\phi}} E(R_{pI}) \quad (2.78)$$

where the factor on the right-hand side is always greater than unity provided that $\phi > 0$, which is the case employed in hydrology. A comparison of the lag-one Markov process and the ARIMA(1,0,1) process on the basis of the asymptotic expected population ranges can be achieved through the use of Eq.(2.74) and Eq.(2.77) which yields

$$E(R_{pA}) = \sqrt{\left(1 + \frac{2\rho}{1-\phi}\right) \frac{1-\rho}{1+\rho}} E(R_{pM}) \quad (2.79)$$

for $\rho = \phi$, $E(R_{pA}) = E(R_{pM})$. The factor assumes the following values

$$\sqrt{\left(1 + \frac{2\rho}{1-\rho}\right) \frac{1-\rho}{1+\rho}} \begin{cases} > 1 & \text{for } \rho < \varphi \\ < 1 & \text{for } \rho > \varphi \end{cases}$$

Hence, as a result when $\varphi > \rho$, then the size of ideal reservoir in the case of an ARIMA(1,0,1) input sequence will be larger than in the case of lag-one Markov process but if $\varphi < \rho$ then the converse will be valid.

2.6.4 Fractional Gaussian Noise

The substitution of Eq.(2.64) into Eq.(2.71) directly gives the expected population range for the small sample in the case of fGn as,

$$E(R_p) = \sqrt{\frac{2}{\pi} C_h} \sum_{k=1}^n k^{h-1} \quad (2.80)$$

For $h=0.5$ Eq.(2.80) reduces to the NIP case as given by Eq.(2.26).

The asymptotic $E(R_p)$ for theoretical fGn can be obtained by substituting the summation in Eq.(2.80) by an appropriate integral as follows,

$$\int_0^h k^{h-1} dk = \frac{1}{h} \left| k^h \right|_0^h$$

which equals $\frac{1}{h} k^h$ hence Eq.(2.80) becomes for the asymptotic case

$$E(R_p) = \frac{1}{h} \sqrt{\frac{2}{\pi} C_h} k^h \quad (2.81)$$

which for $h=0.5$ gives Eq.(2.26) of the NIP case.

2.7 Small Sample Expectation of the Rescaled Range of Various Hydrological Processes

In section (2.4) two types of rescaled range have been introduced; one is associated with the unknown mean which is estimated by the sample mean value and the other one is associated with the known mean or the population mean value.

2.7.1 The Rescaled Adjusted Range

When the range of cumulative departures from sample mean is substituted in the original Hurst's empirical formula of Eq.(1.11), then the statistic, R_n/S is obtained, which has already been referred to as the rescaled adjusted range whose general expression for a single time series is given in Eq.(2.30) as

$$\frac{R_n}{S} = \max_{0 \leq k \leq n} \left[k \frac{\bar{x}_k - \bar{x}_n}{S} \right] - \min_{0 \leq k \leq n} \left[k \frac{\bar{x}_k - \bar{x}_n}{S} \right]$$

where S is the estimate of population standard deviation. In this case the r.v. of which the maximum over range $(0,n)$ is sought, is denoted by z and given as

$$z = k \frac{\bar{x}_k - \bar{x}_n}{S}$$

or, in abbreviated form,

$$z = \frac{z_{k,n}}{S} \quad (2.82)$$

where $z_{k,n}$ is a new r.v. given by

$$z_{k,n} = k(\bar{x}_k - \bar{x}_n)$$

The r.v. $z_{k,n}$, is normally distributed with mean equal to zero and variance given by Eq.(2.39) which can be written as

$$V(z_{k,n}) = k^2 \left[V(\bar{x}_k) - 2 \text{Cov}(\bar{x}_k, \bar{x}_n) + V(\bar{x}_n) \right]$$

or, because of zero mean

$$V(z_{k,n}) = k^2 \left[E(\bar{x}_k^2) - 2E(\bar{x}_k, \bar{x}_n) + E(\bar{x}_n^2) \right] \quad (2.83)$$

The standardized r.v., u' , is obtained by dividing u by its standard deviation, that is

$$u' = \frac{z_{k,n}}{\sqrt{V(z_{k,n})}}$$

where u' is distributed according to $N(0,1)$. In term of this standardized r.v. u' , Eq.(2.82) becomes,

$$z = \sqrt{V(z_{k,n})} \cdot \frac{u'}{S}$$

or by further manipulation,

$$z = \sqrt{V(z_{k,n})} \cdot \frac{u'}{\frac{\sqrt{n} \cdot S}{\sigma}} \cdot \frac{\sqrt{n}}{\sigma}$$

and finally

$$z = \frac{\sqrt{V(z_{k,n})}}{\sigma} \cdot \frac{u\sqrt{n}}{v} \quad (2.84)$$

where

$$v^2 = \frac{ns^2}{\sigma^2}$$

and v^2 is distributed according to χ -(Chi-square) distribution with n degrees of freedom. A very useful theorem due to Hoel (1965) is brought into the argument at this stage which states that if u' is normally distributed with zero mean and unit variance and v^2 is a χ -distribution with n degrees of freedom and in addition u' and v are independently distributed then the random variate,

$$t_n = \frac{u'\sqrt{n}}{v} \quad (2.85)$$

has a Student's t -distribution with n degrees of freedom of which the functional form is given by

$$f(t) = \frac{1}{\sqrt{\pi n}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad (2.86)$$

The variate t_n has a symmetric distribution which becomes normal for large values of n . With this new variable, t_n , Eq.(2.30) can be re-written as

$$\frac{R_n}{S} = \max_{0 \leq k \leq n} \left(\frac{\sqrt{V(z_{k,n})}}{\sigma} t_k \right) - \min_{0 \leq k \leq n} \left(\frac{\sqrt{V(z_{k,n})}}{\sigma} t_k \right) \quad (2.87)$$

where t_k is the t -distributed r.v. associated with time instant k . The

assumption of linear independence of the mean and the standard deviation for any autoregressive process is shown to be valid in appendix (2.3).

By considering that the t-distribution is symmetric and then taking the expectation of both sides of Eq.(2.87) and assuming that

$$E \left[\max_{0 \leq k \leq n} \left(\frac{\sqrt{V(z_{k,n})}}{\sigma} t_n \right) \right] = -E \left[\min_{0 \leq k \leq n} \left(\frac{\sqrt{V(z_{k,n})}}{\sigma} t_n \right) \right] \quad (2.88)$$

then, Eq.(2.87) reduces to

$$E \left(\frac{R_n}{S} \right) = 2E \left[\max_{0 \leq k \leq n} \left(\frac{\sqrt{V(z_{k,n})}}{\sigma} t_n \right) \right] \quad (2.89)$$

Hence, the only problem now, is to find the expectation of the maximum of random variate

$$z_k = \frac{\sqrt{V(z_{k,n})}}{\sigma} t_n \quad (2.90)$$

over a time interval (0,n). In order to find the expectation on the right-hand side of Eq.(2.89) a logical assumption deduced from the case of a normally distributed r.v. is that the distribution of the maximum of t_k is an absolute t-distribution given as

$$f(|t|) = \frac{2}{\sqrt{\pi n}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \left(1 + \frac{t^2}{n} \right)^{-\frac{n+1}{2}} \quad 0 \leq t \leq \infty \quad (2.91)$$

Eq.(2.91) can be written as a consequence of a result derived by Papoulis(1965) which states that if the pdf of any r.v. is even then the absolute pdf of the same r.v. is twice the original pdf provided that the absolute value of the

r.v. is considered to vary from zero to infinity. For any given time instant, k , the expected value of $|t_k|$ becomes

$$E(|t_n|) = \frac{2}{\sqrt{\pi n}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \frac{n}{n-1}$$

or by taking the expectation of Eq.(2.90)

$$E(|z_k|) = \frac{2\sqrt{V(z_{k,n})}}{\sigma\sqrt{\pi n}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \frac{n}{n-1} \quad (2.92)$$

When Eq.(2.92) is substituted into Eq.(2.89) for the expectation of the maximum the following result is obtained.

$$E\left(\frac{R_n}{S} | k\right) = \frac{4}{\sigma} \sqrt{\frac{n V(z_{k,n})}{\pi}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \frac{1}{n-1} \quad (2.93)$$

Eq.(2.93) is a conditional expectation on the time instant k and use of the averaging scheme described in section (2.5) by Eq.(2.49) leads to

$$E\left(\frac{R_n}{S}\right) = \frac{2}{\sqrt{\pi} \sigma} \cdot \frac{\sqrt{n}}{(n-1)} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \sum_{k=1}^n \sqrt{\frac{V(z_{k,n})}{k^2}} \quad (2.94)$$

This is the general expression for the expected value of the rescaled adjusted range. On the other hand the relationship between the adjusted range and rescaled adjusted range can be found simply by the comparison of Eq.(2.52) with that of Eq.(2.94) which yields

$$E\left(\frac{R_n}{S}\right) = \frac{\sqrt{2n}}{(n-1)\sigma} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} E(R_n) \quad (2.95)$$

2.7.2 The Rescaled Population Range

In the case of known population mean, the range associated with Hurst's fundamental expression of Eq.(1. 11) is referred to as the population range whose definition is given by Eq.(2.25) for a time series and the resulting statistic R_p/S is termed the rescaled population range. The general expression for R_p/S is provided by Eq.(2.31) as

$$\frac{R_p}{S} = \max_{0 \leq k \leq n} \left(k \frac{\bar{x}_k}{S} \right) - \min_{0 \leq k \leq n} \left(k \frac{\bar{x}_k}{S} \right)$$

In the light of manipulations performed in section (2.7.1) the general expression for the rescaled population range can be derived by setting $\bar{x}_n=0$ in each one of the expressions of section (2.7.1). Hence, straightforwardly,

$$E\left(\frac{R_p}{S}\right) = \frac{2}{\sqrt{\pi}\sigma} \cdot \frac{\sqrt{n}}{(n-1)} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sum_{k=1}^n \sqrt{V(\bar{x}_k)} \quad (2.96)$$

The comparison of Eq.(2.71) with that of Eq.(2.96) reveals that the rescaled population range is related to the population range through a constant as follows

$$E\left(\frac{R_p}{S}\right) = \frac{\sqrt{2n}}{(n-1)\sigma} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot E(R_p) \quad (2.97)$$

2.8 Application of Rescaled Range to Hydrological Processes

In this section the application of the rescaled range expressions derived in the previous section will be performed. To do this the only information required is the autocorrelation function of the process concerned.

2.8.1 The Normal Independent Process

a) The Rescaled Adjusted Range

In order to find the expectation of the rescaled adjusted range for small samples in the case of NIP Eq.(2.83), which gives the equivalent value of $V(z_{k,n})$, is substituted into Eq.(2.94), hence,

$$E\left(\frac{R_n}{S}\right) = \frac{2}{\sqrt{\pi}(n-1)} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \sum_{k=1}^n \sqrt{\frac{n-k}{k}} \quad (2.98)$$

In addition, the relationship between the adjusted range and rescaled adjusted range for NIP is given by Eq.(2.97). It has been shown in appendix (2.4) that the product term of Eq.(2.97) converges to one as n goes to infinity hence confirming the correctness of the deduction by Mandelbrot and Wallis (1969e) that for very large values of n $E(R_n/S)$ and $E(R_n)$ are virtually equal. Therefore,

$$E\left(\frac{R_n}{S}\right) = E(R_n) \quad (\text{for large } n)$$

Eq.(2.98) has been verified through Monte Carlo techniques and the results of both the theoretical and experimental approaches are presented in figure (2.6). To obtain each one of the experimental points in figure (2.6) 500 samples were generated. A very good agreement between the theoretical and experimental results can be observed from these figures.

b) The Rescaled Population Range

Two ways exist of deriving the expectation of the rescaled population range. It can either be obtained by substituting $V(\bar{x}_k)$ which is σ^2/k into the general expression of Eq.(2.96) or by substituting the already known formula for expected population range in the case of NIP given by Eq.(2.72) into Eq.(2.97). Hence, the result becomes,

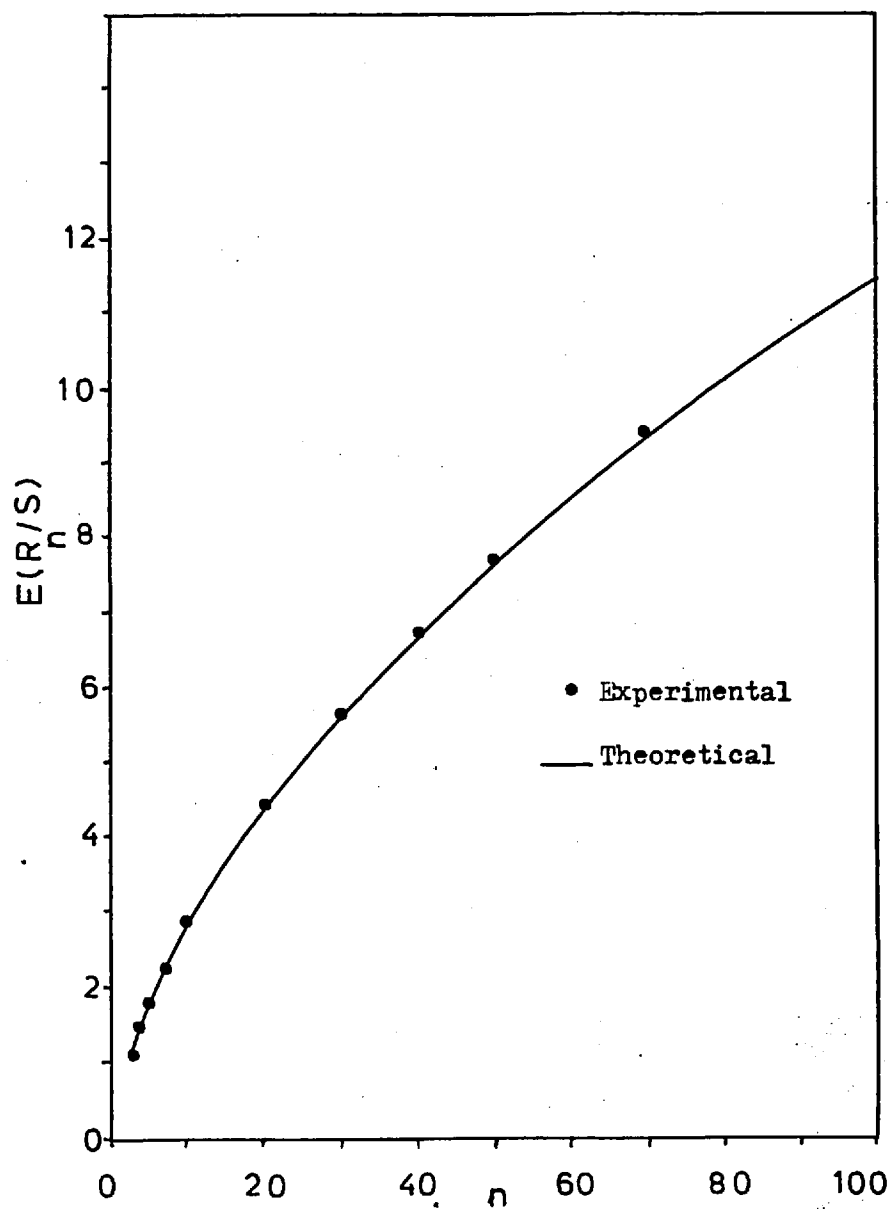


Figure (2.6) - The expected rescaled adjusted range versus sample length for NIP.

$$E\left(\frac{R_p}{S}\right) = \frac{2\sqrt{n}}{\sqrt{\pi}(n-1)} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sum_{k=1}^n k^{-1/2} \quad (2.99)$$

For very large n values, Eq.(2.99) reduces to Eq.(2.26) which is the expected value of the population range associated with the NIP. For small sample sizes, $E(R_p)$ and $E(R_p/S)$ are slightly different from each other in the case of normal distribution. Again Eq.(2.99) has been verified by the Monte Carlo method and the results are presented in figure (2.7) and a good agreement is observed between the theoretical and experimental methods.

2.8.2 The Lag-One Markov Process

a) The Rescaled Adjusted Range

In the light of general formula given in Eq.(2.94), it is only a matter of expressing $V(z_{k,n})$ in terms of the lag-one Markov correlation function, to derive the value of $E(R_p/S)$ for the same process. The elements of $V(z_{k,n})$ have already been given by Eq.(2.58), Eq.(2.59) and Eq.(2.60); therefore substitution of these equations into Eq.(2.83) to find $V(z_{k,n})$ and in turn, the incorporation of $V(z_{k,n})$ into Eq.(2.94) results in a very large expression, which has been numerically evaluated and presented in figures (2.8) - (2.9). The results obtained from 500 synthetic samples prove a good agreement between the experimental and theoretical approach

Furthermore the same relationship between the adjusted range and rescaled adjusted range given by Eq. (2.97) remains valid and asymptotically they become equal as shown in appendix (2.4)

b) The Rescaled Population Range

In a similar way the expectation of the rescaled population range for

small samples and for the lag-one Markov process is obtained by employing either Eq.(2.96) or Eq.(2.97). The expression of $V(\bar{x}_k)$ for this process is given by Eq.(2.58), which on substitution into Eq.(2.96) yields

$$E\left(\frac{R_p}{S}\right) = \frac{2\sqrt{n}}{\sqrt{\pi(n-1)}} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \sum_{k=1}^n k^{-1/2} \left[\frac{1+\rho}{1-\rho} - \frac{2\rho(1-\rho^k)}{k(1-\rho)^2} \right]^{1/2} \quad (2.100)$$

This expression, apart from a constant, is similar to the corresponding expected population range given in Eq.(2.73) and when $\rho = 0$ is substituted it reduces to Eq.(2.99) which is the NIP case. The numerical evaluation of Eq.(2.100) and its verification by generating 500 samples for each particular n value on the computer, is given in figure (2.10) - (2.11).

2.8.3 The ARIMA(1,0,1) Process

a) The Rescaled Adjusted Range

The incorporation of Eq.(2.61), Eq.(2.62) and Eq.(2.63) into Eq.(2.83) gives the necessary expression for $V(z_{k,n})$, the substitution of which into Eq.(2.94) yields a very long formula for the expected rescaled adjusted range. Although this huge expression can explicitly be written, this has been avoided in this section but the numerical evaluation of this huge expression can easily be accomplished on a digital computer.

b) The Rescaled Population Range.

The substitution of Eq.(2.61) into the general expression of Eq.(2.96) results in,

$$E\left(\frac{R_p}{S}\right) = \frac{2\sqrt{n}}{\sqrt{\pi(n-1)}} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \sum_{k=1}^n k^{-1/2} \left[\frac{1-\phi+2\rho}{1-\phi} - \frac{2\rho(1-\phi^k)}{k(1-\phi)^2} \right]^{1/2} \quad (2.101)$$

where $E(R_p/S)$ denotes the expectation of rescaled population range for

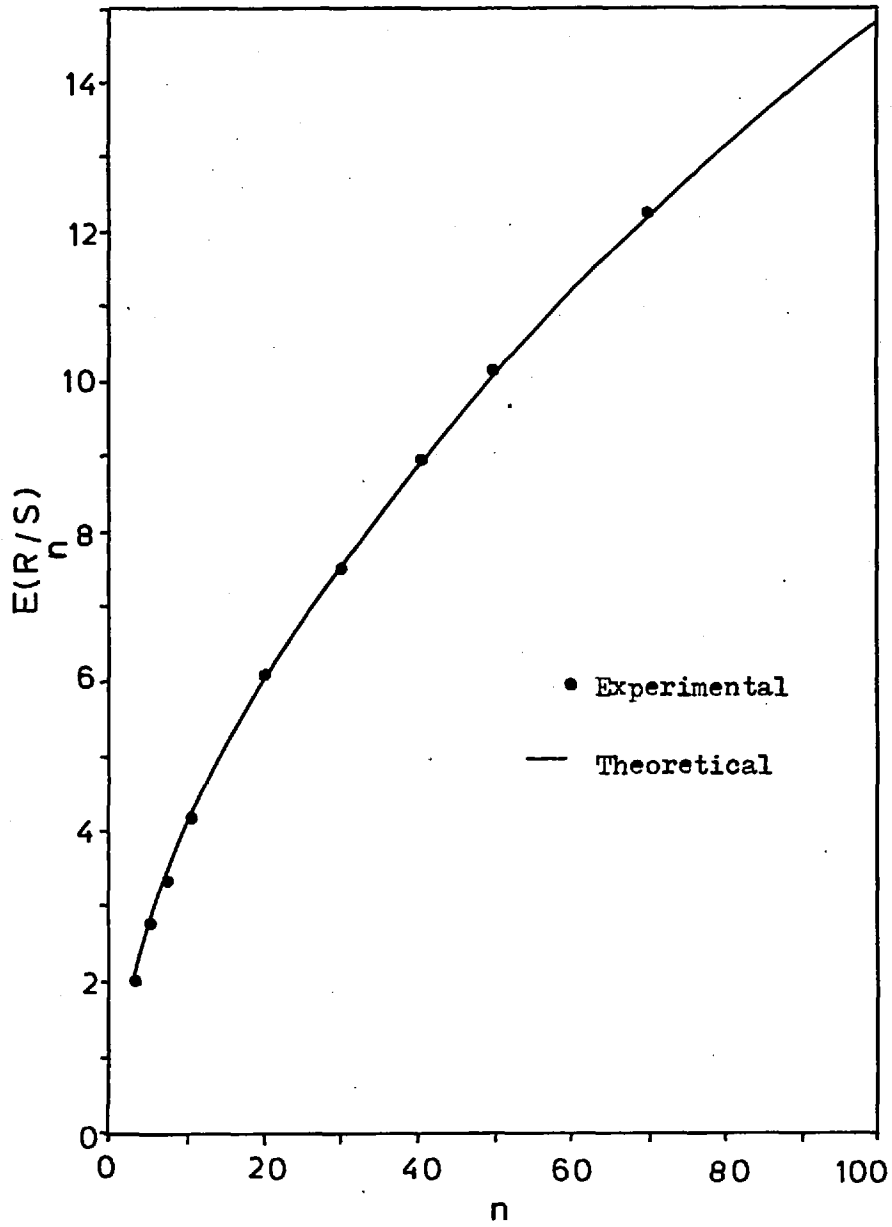


Figure (2.7) - The expected rescaled population range versus sample length for NIP.

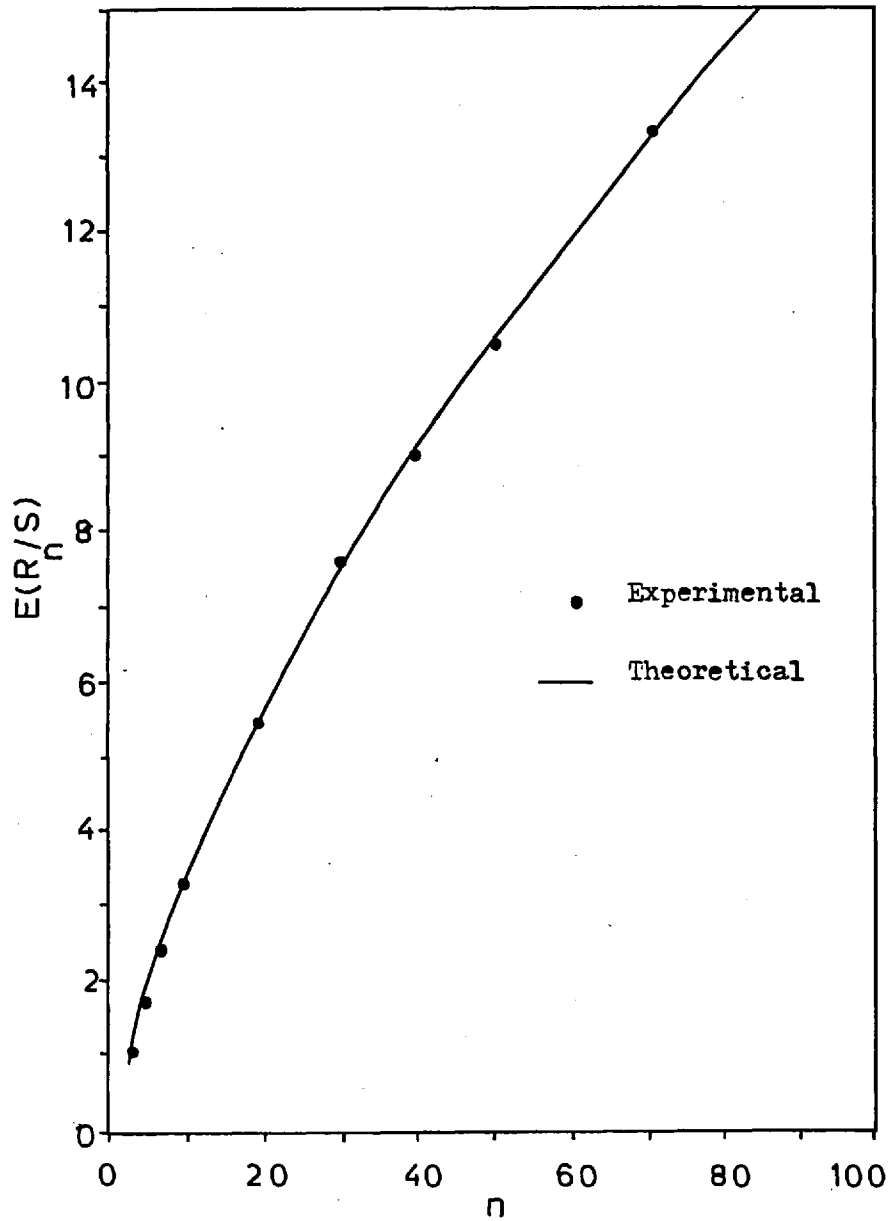


Figure (2.8) - The expected rescaled adjusted range versus sample length for lag-one Markov process with $\rho = 0.4$.

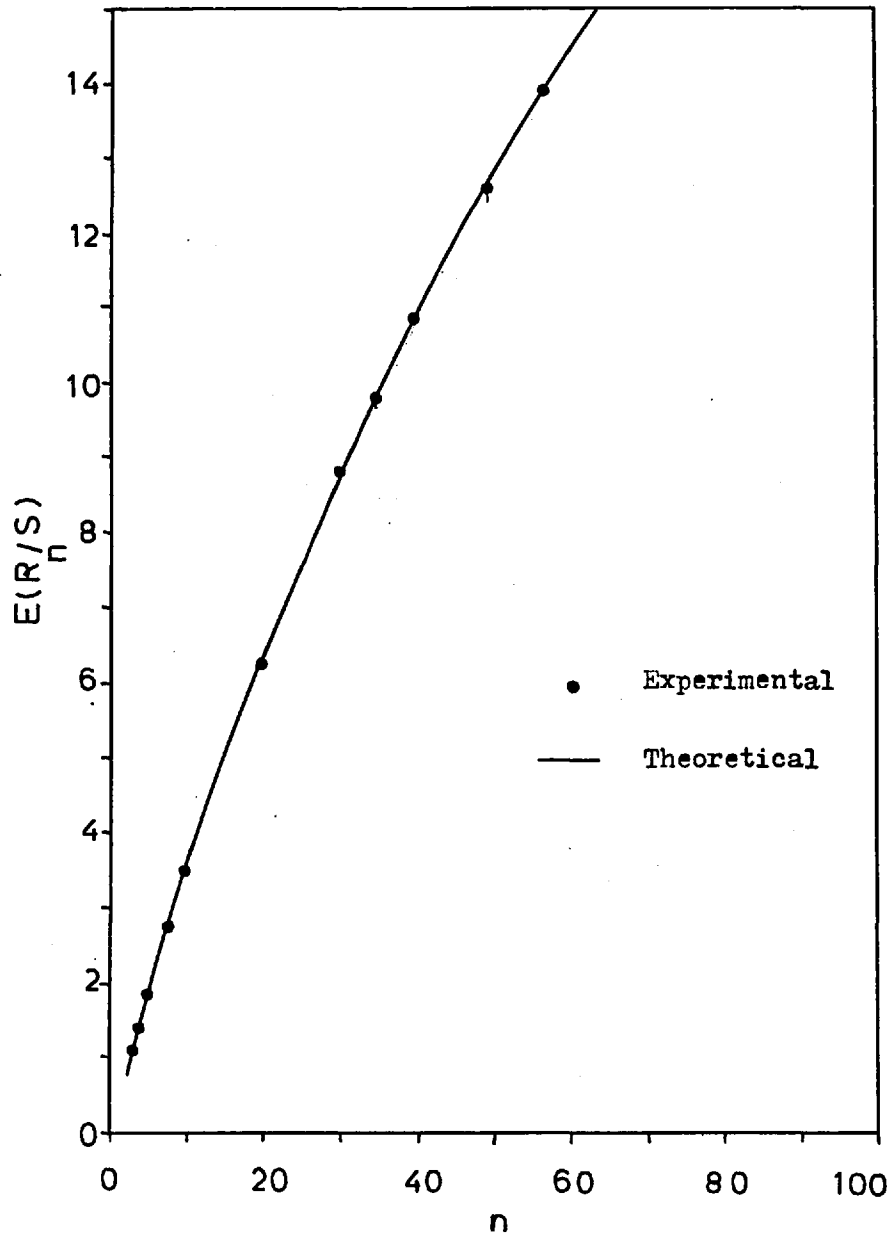


Figure (2.9) - The expected rescaled adjusted range versus sample length for lag-one Markov process with $\rho = 0.6$.

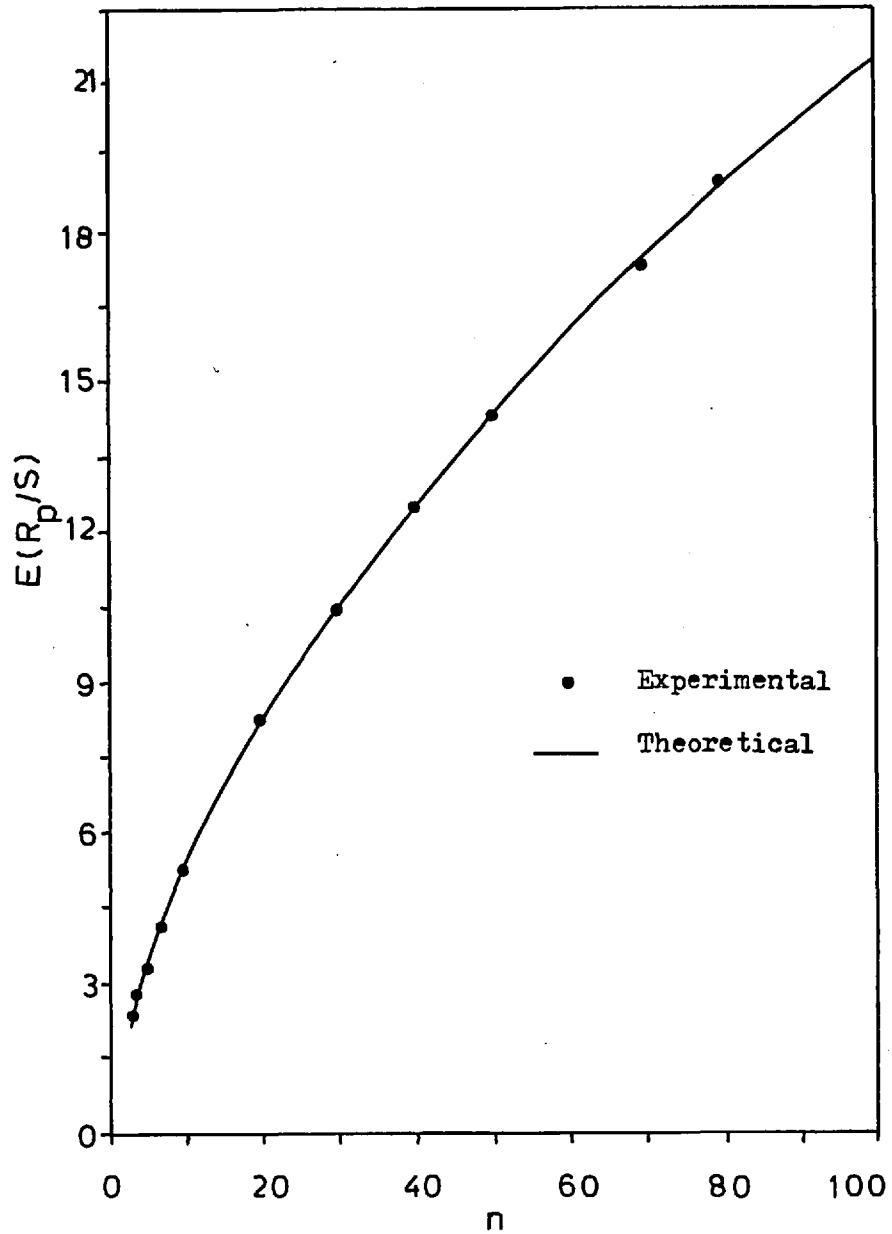


Figure (2.10) - The expected rescaled population range versus sample length for lag-one Markov process with $\rho = 0.4$.

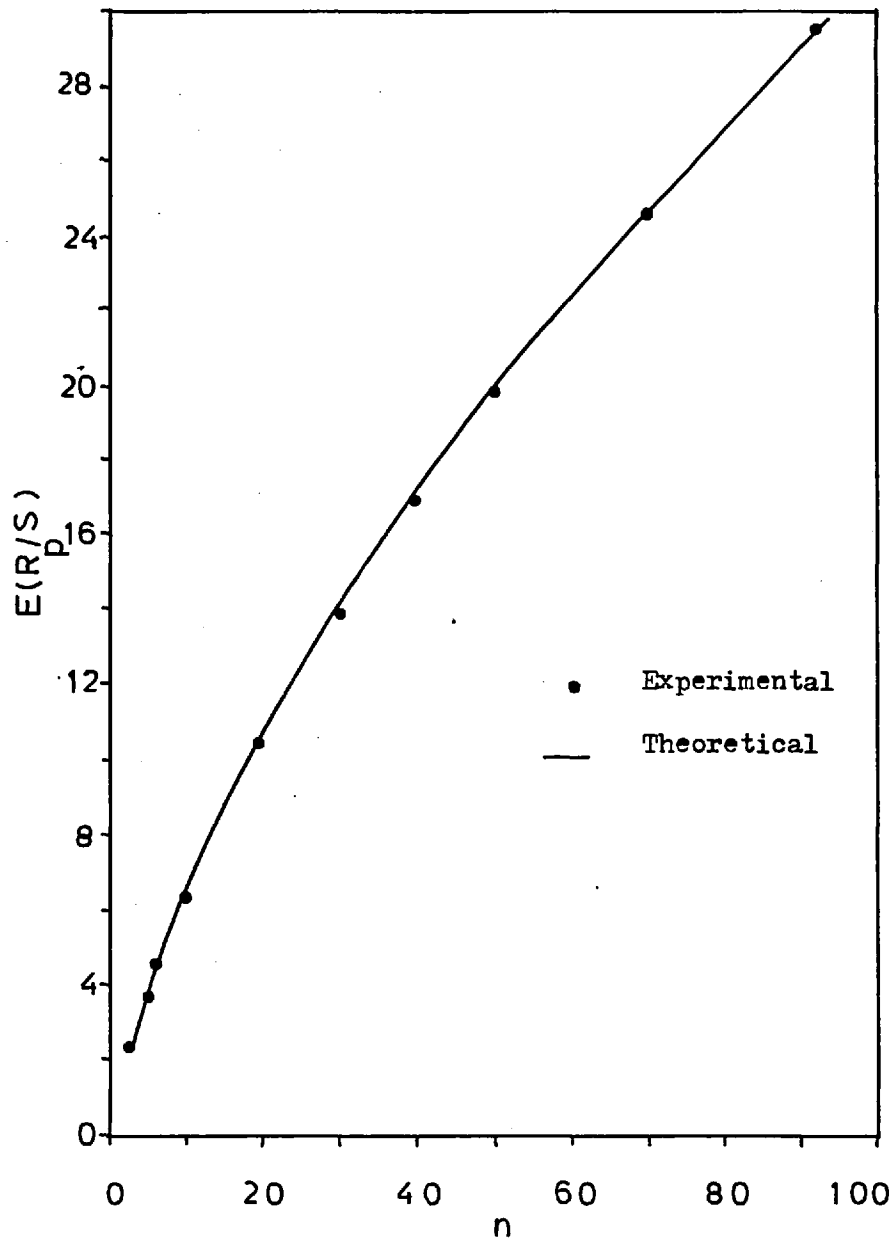


Figure (2.11) - The expected rescaled population range versus sample length for lag-one Markov process with $\rho = 0.7$.

small samples in the case of the ARIMA(1,0,1) process . When $\varphi = \rho$ is substituted into Eq.(2.101) then the result for the lag-one Markov process given by Eq.(2.100) is obtained and for $\rho=0$ and $\varphi=0$ Eq.(2.101) reduces to the NIP case.

2.8.4 Fractional Gaussian Noise

a) The Rescaled Adjusted Range

Again, by substituting Eq.(2.68) in the general expression of Eq.(2.97) provided that R_p is replaced by R_n which yields,

$$E\left(\frac{R_n}{S}\right) = \frac{2\sqrt{n}}{\sqrt{\pi}(n-1)} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \sqrt{C_h} \sum_{k=1}^n \left[k^{2h-2} - \frac{n^{2h-2}}{k} + \frac{(n-k)^{2h}}{kn} - \frac{k^{2h-1}}{k} + n^{2h-2} \right]^{1/2} \quad (2.102)$$

When $h=0.5$ is inserted into Eq.(2.102) then the expected rescaled adjusted range for the NIP given in Eq.(2.98) is obtained.

b) The Rescaled Population Range

A straightforward derivation of the expectation of the rescaled population range is obtained by the substitution of Eq.(2.80) into Eq.(2.97) which leads to

$$E\left(\frac{R_p}{S}\right) = \frac{2\sqrt{n}}{\sqrt{\pi}(n-1)} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \sqrt{C_h} \sum_{k=1}^n k^{h-1} \quad (2.103)$$

which also reduces to Eq.(2.99) for $h=0.5$.

2.9 R_n/S Analysis

One of the fundamental problems in the analysis of historic records is the scientific measurement and investigation of noncyclic very long run

statistical dependence that may exist in these records. A number of techniques for this purpose and their relative merits have been assessed by Mandelbrot (1972). One of these techniques, to be called ' R_n/S Analysis' very effective in this context, has been proposed by Mandelbrot and Wallis (1968; 1969a,c,d). It has also been shown by them that the dependence of $E(R_n/S)$ upon n can introduce the concept of R_n/S intensity, which is one precise and useful measure for the intensity of noncyclic very long-run statistical dependence. The measure of R_n/S intensity has been provided by the quantity $h - 0.5$ with $0 < h < 1$ and particularly $h = 0.5$ corresponds to the absence of noncyclic very long run dependence.

However, in relation to the long-term persistence of a record two new concepts, namely, the Noah and Joseph effect, have been introduced into hydrology by Mandelbrot and Wallis (1968). The Joseph effect describes the fact that a long period of high or low precipitation or streamflow can be extremely long whereas the Noah effect designates the fact that extreme precipitation can be very extreme indeed. It is the Joseph effect which is intimately related to the noncyclic very long-run dependence, so in a way R_n/S may possibly measure how strongly the Joseph effect is present in a given hydrological phenomenon.

The superiority of R_n/S analysis over other techniques, such as the variance-time curves introduced by G.I. Taylor in 1921, the spectral, correlogram and range analysis, in the same context, is because of the fact that the statistic R_n/S is comparatively more robust than other statistics as was pointed out by Mandelbrot and Wallis (1969e). On the other hand, as claimed by Mandelbrot and Wallis (1969d) the unique virtue of R_n/S analysis lies in its being 'blind' to the Noah effect, that is, in its being equally applicable to ^{non-}Gaussian records and to records with a strong Noah effect. The same authors made a definition in order to quantitatively declare a process as, either being long-term persistent or short-term persistent, as

follows. A random process will be said to satisfy the $R_n/S \sim n^{0.5}$ law in the mean if the expression

$$\lim_{n \rightarrow \infty} \left\{ n^{-0.5} E(R_n/S) \right\} \quad (2.104)$$

exists (that is , is well defined) and is positive and finite. In more intuitive terms this $n^{0.5}$ law means that the graph of the function of $\log E(R_n/S)$ versus $\log n$ is asymptotically a straight line with slope 0.5. On the other hand, if the $n^{0.5}$ law in the mean fails to hold then two cases can occur,

(1) the limit operation given in Eq.(2.104) tends to no limit

(2) the same limit tends to zero or to infinity.

These two possibilities amount to saying that the above mentioned graph does not possess a linear asymptotic slope of 0.5. As a result, those processes for which the $n^{0.5}$ law in the mean holds are short-term persistent otherwise they are long-term persistent i.e. exhibit noncyclic long-term statistical dependence.

The application of the definition given above can be undertaken by using the analytical expressions derived for $E(R_n/S)$ in previous sections.

2) The Normal Independent Process.

The limit operation given in Eq.(2.104) can be performed for both the rescaled adjusted range and the rescaled population range. Firstly, let it be applied to $E(R_n/S)$ given in Eq.(2.98) , hence,

$$\lim_{n \rightarrow \infty} \left\{ n^{-0.5} \frac{2}{\sqrt{\pi}} \cdot \frac{1}{(n-1)} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \sum_{k=1}^n \sqrt{\frac{n-k}{k}} \right\} = \sqrt{\frac{\pi}{2}}$$

Therefore the result which confirms an already known fact is that the NIP

is short-term persistent because the limit is positive and finite.

On the other hand, a similar result is obtained when the rescaled population range is employed in the limit operation. Hence, by using Eq.(2.99)

$$\lim_{n \rightarrow \infty} \left\{ n^{-0.5} \left[\frac{2\sqrt{n}}{\sqrt{\pi(n-1)}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \sum_{k=1}^n k^{-1/2} \right] \right\} = 2\sqrt{\frac{2}{\pi}}$$

which also confirms that NIP is short-term persistent, that is $h = 0.5$.

b) The Lag-One Markov Process

In this case the limit operation by the use of rescaled population range yields,

$$\lim_{n \rightarrow \infty} \left\{ n^{-0.5} \frac{2\sqrt{n}}{\sqrt{\pi(n-1)}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \sum_{k=1}^n k^{-1/2} \left[\frac{1+\rho}{1-\rho} - \frac{2\rho(1-\rho^k)}{k(1-\rho)^2} \right]^{1/2} \right\} = 2\sqrt{\frac{2(1+\rho)}{\pi(1-\rho)}}$$

which is positive for any ρ value, hence shows that the lag-one Markov process is short term persistent. The same conclusion could be arrived at by the use of rescaled adjusted range but this has not been tackled here because of the tedious algebraic calculations.

c) The ARIMA(1,0,1) Process

This time Eq.(2.101) is substituted into the expression given in Eq.(2.104) which leads to

$$\lim_{n \rightarrow \infty} \left\{ n^{-0.5} \frac{2\sqrt{n}}{\sqrt{\pi(n-1)}} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \sum_{k=1}^n k^{-1/2} \left[\frac{1-\varphi+2\rho}{1-\varphi} - \frac{2\rho(1-\varphi^k)}{k(1-\varphi)^2} \right]^{1/2} \right\} \\ = 2\sqrt{\frac{2(1-\varphi+2\rho)}{\pi(1-\varphi)}}$$

which is always positive and definite hence the ARIMA(1,0,1) process is also short-term persistent, therefore $h = 0.5$.

d) Fractional Gaussian Process

The application of Eq.(2.104) to Eq.(2.103) yields

$$\lim_{n \rightarrow \infty} \left\{ n^{-0.5} \frac{2\sqrt{n}}{\sqrt{\pi}(n-1)} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \sum_{k=1}^n k^{h-1} \right\} = \frac{1}{h} \sqrt{\frac{2}{\pi}} n^{h-0.5}$$

for values of $0.5 < h < 1.0$ this expression tends to infinity, which means that fGn does not comply with the $n^{0.5}$ law as expected. Therefore, this process is long-term persistent except for $h = 0.5$, in which case the same limit tends to a finite value, reducing the process to the NIP case.

2.10 The Asymptotic Distribution of Rescaled Population Range for NIP.

Through a computer experiment carried out by Mandelbrot and Wallis (1969d) it has been confirmed that the distribution of the rescaled range R_n/S is highly skewed in the case of small samples. Because of this fact it does not suffice to know the lower order moments only, such as $E(R_n/S)$ and $E(R_n/S)^2$, to use R_n/S as a test of statistical independence; the whole distribution needs to be known. No analytical results apparently exist for the distribution of R_n/S , although it can be approximately found through Monte Carlo simulation techniques for any desired process and finite sample length. This approach has been undertaken by Wallis and O'Connell (1973).

However, the asymptotic pdf of statistic R_p/S , has been derived in this section, by making use of the asymptotic pdf of the population range given by Feller (1951), and an assumption which states that the population range, R_p , which has been shown to be a function of the sample mean, $\bar{x}_k$, in Eq.(2.25) and the standard deviation are independently distributed from each

other. This assumption is so important that without it a solution of the problem would be very complicated. Another point is that because the pdf of the asymptotic population range exists for NIP only, the pdf derived below will also be for NIP. As a result, it can be deduced that, in order to derive the pdf of R_p/S for small samples, the first step must be to derive the pdf of R_p for small samples, then in a similar fashion, the small sample distribution of R_p/S could perhaps be attained.

The main theoretical development of the pdf of R_p/S has been given in appendix (2.5). The final result is transferred from appendix (2.5) as,

$$f(y) = \frac{8\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} n^n \sum_{k=1}^{\infty} (-1)^{k-1} k^{2(n^2+k^2y^2)} \frac{1}{2} \frac{1}{2^{(n+1)}} \quad (2.105)$$

where y represents the rescaled population range, R_p/S , and n is the length of time period considered. This expression is valid for any value of n provided that the underlying r.v. is continuous. Otherwise, when the underlying r.v. is discrete, it remains valid asymptotically that is for very large values of n . A partial verification of Eq.(2.105) may be achieved by finding the expected value of the r.v., y , which should be equal to Feller's celebrated result for the expected value of asymptotic population range rederived as Eq.(2.26). The expected value of Eq.(2.105) is,

$$E(y) = \frac{4n}{\sqrt{\pi}(n-1)} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}$$

which at a first glance does not reveal the equivalence to Eq.(2.26), but after a limit operation by extending n to infinity this expression becomes,

$$E(y) = 2 \sqrt{\frac{2}{\pi} n}$$

which is the expected rescaled population range for large n values given by Eq.(2.26). The higher order moments of y , could be worked out from Eq.(2.105) but this has not been attempted here. Unfortunately, the distribution function derived is not very helpful to hydrologists who are mostly concerned with small samples, an unknown population mean and discrete variables. In such a case the distribution function would be very helpful as this would allow a suitable test of significance for long-term persistence to be formulated.

2.11 Summary

A general methodology for deriving the small sample expectations of the range and rescaled range have been presented and its applications to various processes currently used in hydrology have also been shown. The asymptotic equivalence of expected range and rescaled range variates has quantitatively been proven. The R_n/S test for statistical independence in a process has first been reviewed and subsequently an analytical framework of this test for a variety of processes has been proposed; in this way the NIP, lag-one Markov and the ARIMA(1,0,1) process have analytically been shown to be short-term persistent processes characterized by $h = 0.5$.

Finally, the asymptotic pdf of the rescaled population range has been derived. This represents a step further towards formulating a suitable test of significance for values of R_n/S obtained from finite hydrological time series.

Chapter 3

SMALL SAMPLE PROPERTIES OF PARAMETERS OF STATIONARY PROCESSES

3.1. General

The small sample properties of stationary stochastic processes occupy a very important place in synthetic hydrology because of the paucity of hydrological data. Ideally, hydrologists would like to repeat the experiments from which they derive their data under the same conditions, until stable estimates of all the parameters are obtained. Unfortunately, this is not possible in the case of streamflow or precipitation and hydrologists therefore, have to be satisfied with the data collected up to the time of analysing any particular hydrological event which in most cases consists of a short sequence of hydrological observations. Moreover, most of the hydrological data on annual streamflow have a length on average of about 40 to 50 observations; in developing countries even shorter sequences are usually available. This lack of data is a major drawback both in deducing the precise pdf associated with the generating mechanism and in identifying the model which is most suitable to represent this underlying generating mechanism. Even if the pdf and the generating mechanism were known accurately, then the estimates of model parameters, such as the variance, the standard deviation, the serial correlation coefficients etc., based on a small sample, will not, in general be equal to their respective population values.

However, the estimates of model parameters have to be effectively obtained from a historic sequence before a generating process may be applied to any water resources design. The method of moments has been employed by Matalas (1967) for estimation purposes where the small sample moments of any parameter, say, $\hat{\alpha}_n$, which is obtained from a historic record of finite length, n , is equated to its population counterpart, α . The underlying assumption of such a procedure is that,

$$E(\hat{\alpha}_n) = \alpha \quad (3.1)$$

which means that the estimate $\hat{\alpha}_n$ is statistically unbiased; this is a desired property of a parameter estimation procedure. Otherwise, when an estimation procedure does not satisfy Eq.(3.1) then it is said to be statistically biased. Wallis and Matalas (1972) have shown that in the presence of persistence, the moment estimate of the variance is biased and this can be eliminated by using procedures that will be outlined in section (3.5). Wallis and Matalas (1970,1971) have shown that Eq.(3.1) is not justified with respect to estimates of lag-one serial correlation coefficient and the Hurst coefficient. The bias affects the resemblance that is to be maintained between historic and synthetic sequences and as a result is likely to influence the system design in an undesirable manner unless some small sample bias corrections are applied to the estimates. In order to illustrate this point the small sample properties of generating processes must be considered; this is the main objective of this chapter. The processes considered in this chapter are assumed to be stationary stochastic processes with Gaussian r.v's and thus, the generating mechanism is assumed to be known.

In section (3.2) the effect of bias on some very important parameters will be mentioned and it will be shown quantitatively for the moment estimate of the variance in the case of NIP which is the simplest case of all with no persistence. Section (3.3) gives an account of the resemblance that is desirable between a historic and synthetic sequence whereas in section (3.4) a review of the various estimating procedures have been given in general. In addition to this again a review of the small sample properties^{of} the variance of various types of processes has been briefly presented for processes currently used in hydrology. In section (3.6) a new contribution is presented; that is some small sample properties of the standard deviations of different processes, which should be useful to hydrologists who are concerned with the effects of bias in estimates of this parameter. The small sample properties of ρ have been analytically derived, in section (3.7) and in particular

the expressions for small sample estimates of ρ for the ARIMA(1,0,1) process have been given.

3.2. The Bias Effect

An assumption which is a prerequisite in any simulation scheme is that the historic data collected in the field represents a finite random sample from an underlying population. It is not possible to have full information on a r.v. unless all of the values of this r.v. are considered. Since this is not the case for hydrological sequences, the information in each sequence will be lacking in some way and as a consequence of this the estimates of all the parameters describing the distribution function concerned will be operationally biased, that is

$$\hat{\alpha}_n \neq \alpha \quad (3.2)$$

where $\hat{\alpha}_n$ is the estimate and α is the corresponding population value. This bias can sometimes affect important hydrological variables, such as the size of a reservoir, which, in the case of an ideal reservoir is equal to the expected value of range as shown in chapter 2. For instance, the ideal reservoir size, corresponding to a given Gauss-Markov inflow sequence of zero mean and unit variance is given by Eq.(2.73) which is a function of ρ only, for a given n . Therefore as a result, the substitution of an estimate, $\hat{\rho}$, which is different from ρ , will yield an estimate of the ideal reservoir size which in turn be different from the one given by Eq.(2.73).

However, the effect of bias in ρ on the reservoir size for various levels of development in the case of a Gauss-Markov input sequence has been investigated by Wallis and Matalas (1972) by employing Monte Carlo computer experiments. They have shown that without correcting ρ for bias the reservoir size to meet a given demand is underestimated. A similar study was carried out by Şen (1972) for the ARIMA(1,0,1) process for which the effect

of bias in ρ on reservoir size is more pronounced than in the case of the Markov process. So as a consequence of this argument, one can conclude that if the estimates of model parameters were not corrected for bias, then in general any hydrological design variable based on a model with biased parameters will, on average, either be overestimated or underestimated relative to the design associated with the unbiased parameters.

Correspondingly, the water resource system design will be overdesigned or underdesigned, both cases of which incur losses in the sense that if it is overdesigned, the maintenance and structuring costs of the system will be too large and in the case of underdesign insufficient water will be available to meet the demand and water will have to be obtained from other sources.

A very informative example of the bias effect can be given as the sample variance, which is the second order central moment, proposed as an estimate of the population variance. By definition

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.3)$$

where n is the length of the sample and $\bar{x}$ is the estimate of the population mean value, given as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.4)$$

where x is a normal r.v. with mean μ and variance σ . For repeated random samples of size n from the NIP, there is, so to speak, a transformed r.v. that is S^2 given in Eq.(3.3). Hence, S^2 has itself a pdf, the expectation of which is hopefully σ^2 . However, expanding the square in Eq.(3.3) and then taking expectations leads directly to the result,

$$E(S^2) = \frac{n-1}{n} \sigma^2 \quad (3.5)$$

The striking fact is that the expected value of the estimate of the second central moment, $E(S^2)$, is not equal to its population counterpart. On the average over many different samples of fixed size, S^2 will not equal the value of σ^2 , and as a result a definition of biased estimators can be given after Benjamin and Cornell (1970) "Estimates, like S^2 , whose expected values are not equal to the parameter they estimate, are said to be biased estimates".

In order to eliminate the bias in S^2 given by Eq.(3.3) it is convenient to introduce the concept of a bias correction factor, of which frequent mention will be made in the following sections. In this thesis it will, in general, be defined as the ratio of the expected value of the estimate for small samples to its population value. Hence for the example given above this turns out to be,

$$f(n) = \frac{E(S^2)}{\sigma^2} = \frac{(n-1)}{n} \quad (3.6)$$

where $f(n)$ denotes the bias correction factor and the argument n emphasizes that it is a function of the sample length only. Therefore, to find the bias correction factor associated with any estimate, it is necessary first to find the expected value of this estimate for small samples and then divide it by its population value. The division of the biased estimate by corresponding bias correction factor gives an unbiased estimate; for the variance of NIP from Eq.(3.3) and Eq.(3.6) this becomes

$$S^{\star 2} = \frac{S^2}{f(n)} = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.7)$$

To see that this latter estimate is unbiased all that is needed is to expand the square in Eq.(3.7) and take expectations which reduces to

$$E(S^{\star 2}) = \sigma^2 \quad (3.8)$$

which is equal to the population variance whatever the sample size is. The procedure mentioned so far is applicable to any kind of estimator for various parameters such as the standard deviation, σ and the lag-one serial correlation coefficient, ρ . Sometimes it may be necessary to consider the amount of bias which is a measure of the deviation of the biased estimate from its population value. For the case of the variance this can be written as

$$B = \sigma^2 - E(S^2) \quad (3.9)$$

where B denotes the amount of bias. In the case of positive B values, the estimate is said to be biased downwards otherwise it is biased upwards. Some estimates encountered in hydrology have been found to have a downward bias such as the variance for the lag-one Markov process as observed by Wallis and Matalas (1972). It is obvious that Eq.(3.9) is relevant only before the bias correction has actually been applied; when the estimator has been corrected for bias then B will always be equal to zero, whatever the sample length, n , is.

Although, the elimination of bias in an estimate might appear desirable, the unbiased estimate may not necessarily be a 'better' estimate than the biased one. In order to compare a biased estimate with an unbiased estimate

or two estimates that have different amounts of bias, a useful criterion is the mean square error, MSE, of the estimate, measured from the population value which is being estimated. In general, if $\hat{\theta}$ is the biased estimate and $E(\tilde{\theta})$ is the population mean value of r.v., $\tilde{\theta}$, then the MSE is defined as $E[\hat{\theta} - E(\tilde{\theta})]^2$. A meaningful interpretation of MSE will be provided in chapter 5 in connection with the Bayesian estimation of model parameters. More explicitly, MSE can be written as

$$\begin{aligned} E[\tilde{\theta} - E(\tilde{\theta})]^2 &= E\left\{\left[\tilde{\theta} - \hat{\theta}\right] + \left[\hat{\theta} - E(\tilde{\theta})\right]\right\}^2 \\ &= E(\tilde{\theta} - \hat{\theta})^2 + 2\left[\hat{\theta} - E(\tilde{\theta})\right]E(\tilde{\theta} - \hat{\theta}) + \left[\hat{\theta} - E(\tilde{\theta})\right]^2 \\ &= E(\tilde{\theta} - \hat{\theta})^2 + \left[\hat{\theta} - E(\tilde{\theta})\right]^2 \end{aligned} \quad (3.10)$$

Hence, the cross product term vanishes since $E(\tilde{\theta} - \hat{\theta}) = 0$. Or, Eq.(3.10) can be expressed as

$$MSE = (\text{Variance of } \tilde{\theta}) + (\text{Bias})^2 \quad (3.11)$$

So, by employing Eq.(3.10) the comparison of any two estimates can be achieved on the basis of minimum MSE. Unfortunately, a bias correction procedure does create a new estimate and hence in turn a new r.v. and it can be such that the MSE for the biased estimate is smaller than for the unbiased one. Therefore, the conclusion that can be drawn is that the unbiased estimate is desirable only when it is accompanied by a smaller MSE.

However, this chapter is confined to finding unbiased estimates and hence to the bias correction factors, thus leaving the question of the MSE out of its framework.

3.3. Parameter Estimates

Synthetic hydrology suffers from the affliction of inadequate data, resulting frequently in biased parameter estimates and in addition large associated standard errors. By no means is it possible to extract the exact value of the population parameters of the underlying pdf from a finite sample. Therefore, the transition from observed raw data to the necessary parameters is achieved only by an assumption which states that the sample estimates of the parameters are assumed to be equal to their population counterparts. If the estimate obtained from a single finite set of data is not equal to its population counterpart then it is referred to as being operationally biased. There is a gross error involved in the above assumption the magnitude of which diminishes as the length of the sample increases, but for small sample sizes it always exists and it is impossible to get rid of it entirely.

However, operational bias can, by some means, such as regionalization techniques proposed by Benson and Matalas (1967), be reduced to a minimal value. The basis of the regionalization method is to derive a general relationship between the statistical parameters of events at various sites and the hydrological characteristics of the drainage basin such as the drainage area, the slope of the main stream, climatic variables etc. The multiple regression method is usually employed in the derivation of such a relationship. Quite distinct from operational bias which is due to sampling variability, is the statistical bias associated with parameter estimates. If the underlying generating mechanism of the hydrological variable concerned were known a priori, then the estimates of parameters drawn from the finite historic sequence could first be corrected for bias and then these corrected estimates could be assumed to be the equivalents of corresponding parent population parameters.

However, the central idea in synthetic hydrology is to generate

equally likely synthetic sequences such that they will resemble the historic observations with respect to certain parameters. As suggested by Matalas and Wallis (1973) two types of statistical resemblance can be achieved between historic and synthetic sequences. The first type of resemblance is asymptotically satisfied for large samples; although for small samples bias may exist in some estimates, and small sample resemblance may not be satisfied. If $\hat{\theta}$ denotes the estimate of any desired parameter derived from the historic data, then $\hat{\theta}$ plays the role of a population parameter and if $\tilde{\theta}$ is the estimate of the same parameter from a generated synthetic sequence, then the historic and synthetic sequences are said to resemble each other asymptotically provided that $\tilde{\theta} \longrightarrow \hat{\theta}$ as the length of synthetic sequence $n \longrightarrow \infty$. On the contrary, when dealing with a finite historic sample, it would appear more logical to maintain the resemblance for small samples. This kind of resemblance is defined in terms of an ensemble of m synthetic sequences of length n ; from each one of these finite-length sequence an estimate of the parameter is extracted and the estimates are then averaged over the ensemble to give an overall estimate denoted by $\tilde{\theta}_n^*$; then $\tilde{\theta}_n^* \longrightarrow \hat{\theta}$ as the number of sequence $m \longrightarrow \infty$. The first type of resemblance will be referred to as type A and the second type as type B after O'Connell (1974). In this chapter the resemblance for small sample sizes has been fully considered, whereas all of the forthcoming analytical expressions derived for bias corrections remain valid for large sample sizes although this is not necessary.

Moreover, the asymptotic value of the bias correction factor has to be equal to unity, which ensures that the historic and synthetic sequences resemble each other asymptotically as well as for the small samples. The bias is referred to as the statistical bias, which, contrary to the operational bias can be completely eliminated through suitable analytical expressions, in fact by the concept of a bias correction factor.

The view taken in this thesis is that wherever possible an estimate will be corrected for bias.

The magnitude of bias in estimates of parameters of a stochastic process is dependent on the following factors.

- (1) the marginal pdf
- (2) the autocorrelation structure
- (3) the type of estimate employed
- (4) the length of sequence
- (5) whether or not the mean of the process is known or has to be estimated.

Here, only the normal pdf will be considered as a marginal distribution. The relationships between the bias and the autocorrelation structures of various stationary stochastic processes will be studied in detail. As far as the types of parameter estimates are concerned, there exist three widely used alternatives which are

- (1) The moment estimate
- (2) The maximum likelihood estimate, and
- (3) The Bayesian estimate

The moment and maximum likelihood estimates of parameters are the classical statistical methods based on historic data only, whereas the Bayesian method of parameter estimation incorporates information obtained from both the historical data and other resources related to these data. A very brief description of these estimates will be given in the following sections whereas a detailed study of the Bayesian estimate is deferred to chapter 5. In the case of both the mean and the variance, the moment and the maximum likelihood methods are equivalent when the underlying generating mechanism is NIP. On the other hand, three of the most commonly used estimates of the serial correlation coefficient will be presented in section (3.7).

As the length of the sample increases, the amount of bias will decrease, giving a closer estimate to the population counterpart. Finally, if the mean of the underlying generating mechanism is known the amount of bias in the estimate of the variance, the standard deviation and the serial correlation coefficient will be smaller than that of the unknown mean case, because an estimate of the mean will introduce a further bias.

3.4. Methods of Estimation

After having hypothesized the type of model the next problem of the hydrologist is the estimation of the parameters from the available data in such a way that the model fits the observed pattern in some statistical sense. In general, a parameter estimate can be defined as being a function of the sample. Therefore, the estimates themselves represent transformed r.v's of the observed sequence.

However, because of their unique role in fitting a model to the historical sequence they deserve to be examined more closely. Generally, two types of uncertainties are recognized in stochastic hydrology. Firstly there is the uncertainty that the underlying generating mechanism of the phenomenon is unknown and secondly the failure to estimate the model parameters with precision which is referred to as statistical uncertainty and is treated in detail in the following sections on the basis of the assumption that the underlying generating process is known.

3.4.1. The Method of Moments

This is the most widely used and simple method of estimating parameters where the central idea is that the moments calculated from a given sample are assumed to be equal to the moments of the parent pdf. The number of sample moments employed is equal to the number of unknown parameters and in this way a unique solution for the parameters in terms of the sample moments is possible. As stated by Cramer (1966) under fairly

general conditions, the distribution of an estimate of this kind will be asymptotically normal for large n and the mean of the estimate will tend to be different from the true value by a quantity of order n^{-1} , and the standard deviation will be asymptotically of the form $c/\sqrt{n}$. In general, by a simple operation the bias in moment estimates can be removed and an unbiased estimate obtained. Although the method of moments yield asymptotically consistent estimates that is the estimate converges to its population counterpart as the length of sample increases, they do not have the smallest possible variance.

3.4.2. The Maximum Likelihood Method

Theoretically, this is the most important general method applied to estimate parameters of a given stochastic process. If a sequence of ^{independent} r.v's $x_1, x_2, x_3, x_4, \dots, x_n$ were given which are characterized by a set of unknown parameters, then, a kind of multivariate pdf of the r.v. is given as

$$f(x_1, x_2, x_3, \dots | \theta) = f(x_1) \cdot f(x_2) \cdot f(x_3) \cdot \dots \cdot f(x_n) \quad (3.12)$$

where the left hand side of this expression reads as the multivariate pdf of the r.v. given θ if θ were known. However, this is not the situation encountered in practice because of the fact that by observing a random phenomenon one gets a sample of which the parameters are unknown and to be estimated. Therefore, the r.h.s. of Eq.(3.12) can be considered as a function of θ only, for a given sample and the likelihood function is defined as,

$$l(\theta, x_1, x_2, x_3, \dots, x_n) = \prod_{i=1}^n f(x_i | \theta) \quad (3.13)$$

The method of maximum likelihood now consists of choosing, as an estimate of the unknown population value of θ , the particular value that renders l as large as possible. Since $\log l$ attains its maximum for the same value of θ , as l , it is thus only necessary to solve the likelihood equation

$$\frac{\delta \log l}{\delta \theta} = 0 \quad (3.14)$$

with respect to θ . Any solution of the likelihood equation through Eq.(3.14) will then be called a maximum likelihood estimate of θ . The importance of the method is clearly shown by Cramer (1966) who states:

(i) If an efficient estimate $\hat{\theta}$ of θ exists, the likelihood equation will have a unique solution equal to $\hat{\theta}$.

(ii) If a sufficient estimate $\hat{\theta}$ of θ exists, any solution of the likelihood equation will be a function of $\hat{\theta}$.

(iii) It is also shown by Cramer that under certain conditions, the likelihood equation given in Eq.(3.13) has a solution which converges in probability to the true value of θ , as $n \rightarrow \infty$. This solution is an asymptotically normal and efficient estimate of θ . In addition, the MSE is theoretically found to be asymptotically equal to

$$\text{Var}(\theta) = - \frac{1}{nE \left[\frac{\delta^2 \log l(x|\theta)}{\delta \theta^2} \right]}$$

which diminishes as n increases.

3.4.3. Bayesian Estimation

This alternative method of parameter estimation basically requires three essential ingredients, namely, the prior distribution of the parameters

concerned, the data which give some additional information through the likelihood function and finally an objective function. The mixture of the prior pdf and the likelihood function is achieved by employing Bayes theorem, in order to yield a posterior distribution of the parameter. Bayes theorem states that,

$$P(\theta|x) \propto P(\theta) \cdot l(x|\theta) \quad (3.15)$$

where $P(\theta|x)$ is the posterior pdf; $P(\theta)$ is the prior pdf of the parameter, θ and $l(x|\theta)$ is the likelihood function which represents the data. The Bayesian estimate of any parameter is obtained by optimizing the loss function. For example, if the loss associated with an estimation error, $(\theta - \hat{\theta})$, is assumed proportional to the squared error $(\theta - \hat{\theta})^2$ then the optimal decision is to use the mean of the posterior pdf of θ as an estimate. The Bayesian estimation of the parameters of the ARIMA(1,0,1) process and thence the decision making procedure will be presented in greater detail in chapter 5.

3.5. Small Sample Expectation of the Population Variance. (σ^2)

In the literature the small sample estimation of σ^2 has apparently been explored to a greater extent than for some other parameters. An unbiased estimate of the variance as mentioned in section (3.2) for the case of NIP is given by Eq.(3.7), but in the case of dependent r.v's the same estimate yields a biased value because as explained in section (3.3) another source of bias, that is the autocorrelation structure, comes into play. Therefore, the bias correction function given in Eq.(3.6) will, apart from the size of sample, n , be a function of the autocorrelation function. In the case of the lag-one Markov process, it is a function of ρ which specifies the autocorrelation function completely and for the ARIMA(1,0,1) process ψ

and Θ which also specify the autocorrelation function of the process completely. The general expected value of S^2 for an autocorrelated process is found by manipulating Eq.(3.6) as in appendix (3.1) which gives

$$E(S^2) = \sigma^2 \left[1 - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n \rho_{|i-j|} \right] \quad (3.16)$$

Hence, the bias correction factor in general becomes, from Eq.(3.6)

$$f(n; \rho_1, \rho_2, \rho_3, \dots, \rho_n) = 1 - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n \rho_{|i-j|} \quad (3.17)$$

Moreover, the amount of bias can be found by substituting Eq.(3.16) into Eq.(3.9) which leads to

$$B(n; \rho_1, \rho_2, \rho_3, \dots, \rho_n) = \frac{\sigma^2}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n \rho_{|i-j|} \quad (3.18)$$

The double summation operates over the off-diagonal terms of the autocorrelation matrix, and because of the symmetry of this matrix, Eq.(3.16), Eq.(3.17) and Eq.(3.18) can be re-written in more compact forms as

$$E(S^2) = \sigma^2 \left[1 - \frac{2}{n(n-1)} \sum_{i=1}^{n-1} (n-i) \rho_i \right] \quad (3.19)$$

$$f(n; \rho_1, \rho_2, \rho_3, \dots, \rho_n) = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{n-1} (n-i) \rho_i \quad (3.20)$$

and finally

$$B(n; \rho_1, \rho_2, \rho_3, \dots, \rho_n) = \frac{2\sigma^2}{n(n-1)} \sum_{i=1}^{n-1} (n-i) \rho_i \quad (3.21)$$

Hence, the only information needed to work out various bias properties of the variance is the autocorrelation structure of the underlying generating mechanism. The expected value of the variance for small sample sizes, when the generating mechanism is a lag-one Markov process, has already been derived by Matalas (1967) as

$$E(S^2) = \sigma^2 \left\{ 1 - \frac{2\rho}{n(n-1)} \left[\frac{n(1-\rho) - (1-\rho^n)}{(1-\rho)^2} \right] \right\} \quad (3.22)$$

if $\rho = 0$, then $E(S^2) = \sigma^2$ which corresponds to the NIP case, while if $\rho > 0$ then the bias correction factor

$$f(n; \rho) = 1 - \frac{2\rho}{n(n-1)} \left[\frac{n(1-\rho) - (1-\rho^n)}{(1-\rho)^2} \right] \quad (3.23)$$

is positive and the magnitude of the bias obtained from Eq.(3.22) as,

$$B(n; \rho) = \frac{2\rho}{n(n-1)} \left[\frac{n(1-\rho) - (1-\rho^n)}{(1-\rho)^2} \sigma^2 \right] \quad (3.24)$$

which confirms the fact that the effect of bias increases as ρ increases for a fixed n , and as n decreases when ρ is fixed. In the case of the lag-one Markov process the unbiased estimate of the population variance becomes

$$S^* = \frac{S^2}{f(n;\rho)} \quad (3.25)$$

where ρ is the lag-one autocorrelation coefficient. By taking the expectation of both sides of Eq.(3.25)

$$E(S^*) = \frac{E(S^2)}{f(n;\rho)} \quad (3.25A)$$

so that S^* will vanish.

In the same manner, the small sample expectation of the variance of the ARIMA(1,0,1) process can be obtained from the general expression given in Eq.(3.19) by substituting the autocorrelation function of this process provided in Eq.(1.21) and Eq.(1.22) which leads to an expression already derived by O'Connell (1974) as

$$E(S^2) = \sigma^2 \left\{ 1 - \frac{2\rho}{n(n-1)} \left[\frac{n(1-\varphi) - (1-\varphi^n)}{(1-\varphi)^2} \right] \right\} \quad (3.26)$$

Moreover, the bias correction factor and the amount of bias for the ARIMA(1,0,1) process becomes,

$$f(n;\rho,\varphi) = 1 - \frac{2\rho}{n(n-1)} \left[\frac{n(1-\varphi) - (1-\varphi^n)}{(1-\varphi)^2} \right] \quad (3.27)$$

and

$$B(n;\rho,\varphi) = \frac{2\rho\sigma^2}{n(n-1)} \left[\frac{n(1-\varphi) - (1-\varphi^n)}{(1-\varphi)^2} \right] \quad (3.28)$$

where ρ and ϕ specify the autocorrelation function of the ARIMA(1,0,1) process. The amount of bias analytically provided in Eq.(3.28) decreases with increasing n , and increases with both increasing ρ and ϕ .

It has been noted by O'Connell (1971) that values of ϕ in the range $0.85 < \phi < 0.95$ are of interest in modelling long-term persistence. For such values of ϕ and values of ρ usually encountered for annual streamflow, the bias in S^2 is generally large. Bias correction factor for the dfGn process can be found from the general expression of Eq.(3.19) which is left outside of the framework of this thesis.

3.6. Small Sample Expectation of the Population Standard Deviation. (σ)

So far, in the hydrological literature, to the best of the author's knowledge no analytical small sample results for the standard deviation have appeared even though this parameter is more frequently used in the structure of any mathematical models than the variance. Hence, no facility exists for removing bias in estimates of the standard deviation. However, an approximate method of getting rid of the bias effect in the standard deviation is developed in this section, and it has been checked by the Monte Carlo simulation technique.

Firstly, although the standard deviation, S , is the square root of the variance, unfortunately the same transformation is not valid for the bias correction of the two parameters. That is to say, although,

$$E(S^2) = f(n;\rho)\sigma^2$$

but

$$E(S) \neq \sqrt{f(n,\rho)} \sigma$$

or

$$E(S) \neq \sqrt{E(S^2)}$$

Thus a separate procedure is necessary to correct S for bias. The question as to whether it is best to correct the variance rather than the standard deviation will not be resolved here. It is hoped here, that the bias correction procedure proposed in this chapter for the standard deviation will provide some useful information for hydrologists who attach an importance to a bias correction for S.

For a given sequence of observations $x_1, x_2, \dots, x_n$ the variance which is the measure of the spread of these observations about the mean value is expressed by Eq.(3.7). The r.h.s. of this equation can be expanded to give

$$S^2 = \frac{1}{(n-1)} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] \quad (3.29)$$

Although, S^2 is uniquely determined from a sequence of events, when an ensemble of the sequences of the same length n is considered, S^2 is obviously a r.v. with a distribution function for each value of n. Let Eq.(3.29) be written as,

$$S^2 = A \quad (3.30)$$

where

$$A = \frac{1}{(n-1)} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] \quad (3.31)$$

The pdf of the r.v. A can be centered round zero, by using the following linear transformation provided that the expected value of S^2 , $E(S^2)$, which is very simple to calculate and its general analytical form for various processes has been given in the previous section. The linear shifting procedure already referred to is given as

$$a = A - E(A) \quad (3.32)$$

Here, Q is a newly defined r.v. which has exactly the same pdf as A in shape but its moments are related to the moments of the r.v. A through Eq.(3.32). Therefore, the expected value of Q is zero. From Eq.(3.32) A is solved for and then substituted into Eq.(3.30) which leads to

$$S^2 = a + E(A) \quad (3.33)$$

By taking the square root of both sides, Eq.(3.33) becomes,

$$S = \sqrt{a + E(A)}$$

or,

$$S = \sqrt{E(A)} \left[1 + \frac{a}{E(A)} \right]^{1/2} \quad (3.34)$$

The second product term on the r.h.s. of this last expression may be expanded into a Binomial form as,

$$S = \sqrt{E(A)} \left[1 + \frac{a}{2E(A)} - \frac{a^2}{8E^2(A)} + \frac{a^3}{16E^3(A)} - \frac{5a^4}{128E^4(A)} + \dots \right] \quad (3.35)$$

By taking the expected value of both sides, Eq.(3.35) becomes,

$$E(S) = \sqrt{E(A)} \left[1 + \frac{E(a)}{2E(A)} - \frac{E(a^2)}{8E^2(A)} + \frac{E(a^3)}{16E^3(A)} - \frac{E(a^4)}{128E^4(A)} + \dots \right] \quad (3.36)$$

The second term in the brackets is zero because of the fact that $E(a) = 0$.

If it were easily possible to calculate the second and higher order central moments of the r.v., a , then Eq.(3.36) expression would yield an exact value of $E(S)$ and hence provide the amount of bias, that is $\sigma - E(S)$, for a given generating process as well.

In order to simplify the algebra, higher order moments greater than two will be neglected in the expansion; consequently the following simple approximation is obtained.

$$E(S) = \sqrt{E(A)} \left[1 - \frac{E(a^2)}{8E^2(A)} \right] \quad (3.37)$$

The only quantity that remains to be found is the value of $E(a^2)$ which is expressible in terms of the moments of the unshifted r.v., A , as follows.

From Eq.(3.32) by squaring both sides and then taking the expectations

$$E(a^2) = E(A^2) - 2E(A)E(a) + E^2(A)$$

or,

$$E(a^2) = E(A^2) - E^2(A)$$

where the r.h.s. is equal to the variance of A . Hence,

$$E(a^2) = V(A) \quad (3.38)$$

Thus the variance of A which is in turn the variance of S^2 satisfies the following equations

$$E(a^2) = V(A) = V(S^2) \quad (3.39)$$

On substitution, Eq.(3.37) takes its final form as

$$E(S) = \sqrt{E(S^2)} \left[1 - \frac{V(S^2)}{8E^2(S^2)} \right] \quad (3.40)$$

which is a general expression which is thought to be valid for every kind of stationary stochastic process. The only unknown term in Eq.(3.40) is $V(S^2)$, because $E(S^2)$ has been given in its general form by Eq.(3.16). On the other hand the general formula given by Eq.(3.40) can be applied to derive $E(S)$ for autocorrelated processes. The general form of the variance of the variance, $V(S^2)$ has been provided for any autoregressive process by Bayley and Hammersley (1946) as

$$V(S^2) = \frac{2}{(n^* - 1)} \sigma^4 \quad (3.41)$$

where n^* is given by the same authors in its general form for an autoregressive model as follows

$$n^* = n \left\{ \frac{1 + \frac{1}{(n-1)} \sum_{j=1}^{n-1} \rho_j^2}{1 + \frac{n}{(n-1)} \sum_{j=1}^{n-1} \rho_j^2} \right\} \quad (3.42)$$

A reasonable approximation to Eq.(3.42) has been also given by the same authors as

$$n^* = \frac{n}{\sum_{j=0}^{n-1} \rho_j^2} \quad (3.43)$$

where n^* is referred to as the effective number of independent observations in an autocorrelated series.

Hence, the derivation of the general expression for $E(S)$ can be carried out using the above expressions. To do this, the following abbreviation for the expressions of $E(S^2)$ given in Eq.(3.22) and Eq.(3.26) is employed

$$E(S^2) = \sigma^2 F \quad (3.44)$$

where F has a specific value for a specific autoregressive model. Substitution of Eq.(3.44) into Eq.(3.40) then gives,

$$E(S) = \sigma \sqrt{F} \left[1 - \frac{1}{4(n^*-1)F^2} \right] \quad (3.45)$$

a) The Normal Independent Process

The exact expression for $V(S^2)$ in the case of NIP has been given in almost all textbooks of statistics in relation to random sampling as,

$$V(S^2) = \frac{2}{(n-1)} \sigma^2 \quad (3.46)$$

which is a function of the sample length, n , only. Substitution of Eq.(3.46) and $E(S^2) = \sigma^2$ into Eq.(3.40) yields

$$E(S) = \sigma \left[1 - \frac{1}{4(n-1)} \right]$$

whence the bias correction factor and the amount of bias associated with the standard deviation can be written as,

$$f(n) = 1 - \frac{1}{4(n-1)} \quad (3.47)$$

and

$$B(n) = \frac{1}{4(n-1)} \sigma \quad (3.48)$$

These expressions reveal the two important facts one of which is that as the sample size increases the amount of bias, $B(n)$, tends to zero and therefore

$$\lim_{n \rightarrow \infty} E(S) = \sigma$$

which means that the estimate, S , is asymptotically efficient. The second fact is that the bias correction factor given in Eq.(3.47) is not the square root of the bias factor for the corresponding variance given by Eq.(3.6)

However, if an unbiased estimate S^* of S is required then by employing Eq.(3.7) the following form of estimate can be suggested in lieu of Eq.(3.7)

$$S^* = \frac{1}{\left[1 - \frac{1}{4(n-1)}\right]} \left[\frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2} \quad (3.49)$$

which, by taking the expectations of both sides yields

$$E(S^*) = \sigma$$

confirming the fact that the estimate proposed in Eq.(3.49) is unbiased. However, it should be noted that such an estimate yields a biased estimate of the variance, as by taking first the square and then the expectation of both sides of Eq.(3.49)

$$E(S^{*2}) = \frac{\sigma^2}{\left[1 - \frac{1}{4(n-1)}\right]^2}$$

Thus one very important conclusion that can be drawn is that it is not possible to have simultaneously unbiased estimates of S^2 and S in a generating process. When one is required to be unbiased, the other remains biased. This is true for small sample sizes only; as the sample length, n , tends to infinity the two estimates proposed in Eq.(3.7) and Eq.(3.49) converge to their population values, hence confirming that either equation gives unbiased estimates of the variance and standard deviation simultaneously.

The small sample expectation of S for the NIP given by Eq.(3.49) has been verified experimentally using the Monte Carlo technique and the result is plotted in figure (3.1). For production of this figure, 3000 synthetic sequences were generated independently from each other for each of a set of

sample sizes of 2,3,4,5,6,7,8,9,10,15,20,25,30,35,40,45 and 50. The agreement between the theoretical and experimental results is completely perfect as is seen on figure (3.1).

b) The Lag-One Markov Process

The only information needed to find the corresponding expression of Eq.(3.40) for any stationary stochastic process is its autocorrelation structure. The autocorrelation structure of the lag-one Markov process is the exponential form given in Eq.(2.57). Therefore, the summation terms in the expression for n^* given in Eq.(3.42) can be calculated in terms of this autocorrelation function as follows

$$\begin{aligned}\sum_{j=1}^{n-1} \rho_j^2 &= \rho_1^2 + \rho_2^2 + \rho_3^2 + \dots + \rho_{n-1}^2 \\ &= \rho^2 + \rho^4 + \rho^6 + \dots + \rho^{2(n-1)} \\ &= \rho^2 \left[1 + \rho^2 + \rho^4 + \dots + \rho^{2(n-2)} \right]\end{aligned}$$

The summation in brackets is a geometric progression; hence

$$\sum_{j=1}^{n-1} \rho_j^2 = \rho^2 \frac{1 - \rho^{2(n-1)}}{1 - \rho^2} \quad (3.50)$$

Substituting Eq.(3.50) into Eq.(3.42) gives

$$n^* = n \left\{ \frac{(n-1)(1 - \rho^2) + \rho^2 [1 - \rho^{2(n-1)}]}{(n-1) + \rho^2 [1 - n\rho^{2(n-1)}]} \right\} \quad (3.51)$$

If the approximate version of n^* given in Eq.(3.43) is considered, then

$$\bar{n}^* = \frac{n(1 - \rho^2)}{(1 - \rho^{2n})} \quad (3.52)$$

The value of F for the lag-one Markov process is given from Eq.(3.22) as

$$F = 1 - \frac{2\rho}{n(n-1)} \left[\frac{n(1 - \rho) - (1 - \rho^n)}{(1 - \rho)^2} \right] \quad (3.53)$$

Hence, after having found all the necessary terms of the general expression given by Eq.(3.40) the small sample expectation of S for this process and thence the bias correction factor and the amount of bias can easily be derived either by using the exact or approximate value of $\bar{n}^*$. The explicit form of these quantities have not been given here because the expressions are rather large.

On the other hand, Monte Carlo simulations have shown that the approximate form is very successful even for very small sample sizes such as $n = 2$. A series of experiments were carried out in order to check the analytical results for various ρ values; some of these results have been displayed in figures (3.2) - (3.5). The number of samples used in this experimental scheme is 4000. The general conclusion drawn from these figures is that for small values of ρ , Eq.(3.40) gives a perfect match with the experimental ones, even for very small sample sizes, n , but as ρ increases the results are not so consistent for very small sample sizes but they improve quickly as sample size increases. Considering the small sample sizes used in hydrology, that is mostly in the range $n = 10$ to $n = 40$ the bias correction factors found through the use of general analytical expressions give a satisfactory result even for large ρ values, which do not occur

widely for annual streamflow sequences. For example, the serial correlation coefficient of annual streamflow sequences are on average of the order of 0.2-0.3 as quoted by Yevjevich (1963) and as a result this hydrological fact makes the suggested analytical expressions quite attractive.

c) The ARIMA(1,0,1) Process

The autocorrelation structure of this process is given by Eq.(1.20) and Eq.(1.21) which are functions of ρ and φ only. By using this autocorrelation function the value of n^* for both the exact and approximate forms are

$$n^* = n \left\{ \frac{(n-1)(1-\varphi^2) + \rho^2 [1 - \varphi^{2(n-1)}]}{(n-1)(1-\varphi) + n\rho^2 [1 - \varphi^{2(n-1)}]} \right\} \quad (3.54)$$

and

$$n^* = n \frac{n(1-\varphi^2)}{\{1-\varphi-\rho^2 [1 - \varphi^{2(n-1)}]\}} \quad (3.55)$$

respectively. The expression of F can be obtained from Eq.(3.26) as

$$F = 1 - \frac{2\rho}{n(n-1)} \left[\frac{n(1-\varphi) - (1 - \varphi^n)}{(1-\varphi)^2} \right] \quad (3.56)$$

By substituting Eq.(3.54) or Eq.(3.55) and Eq.(3.56) into the general expression of Eq.(3.40), the value of $E(S)$ is obtained and has been plotted against n in Figures (3.6) and (3.7). For two sets of parameter values on the same figures the experimental results are also shown for the sake of comparison.

However, it was found for the ARIMA(1,0,1) process that the bias correction factors are not equally good for the entire range of the parameters φ and Θ , but it is fortunate that the analytical expressions provided are valid on the ranges of φ and Θ likely to be used in synthetic hydrology.

3.7. Small Sample Estimation of ρ

One of the most important features of hydrological time series is that the successive observations are serially correlated with each other. This property is known as persistence in hydrological processes. By a further consideration this persistence can be split up into two parts; short-term persistence which has so far been measured by the first order serial correlation coefficient, ρ , and long-term persistence a measure of which is achieved by the Hurst coefficient, h . From a sequence only a single estimate of both ρ and h can be obtained and they have been shown to be biased in small samples by Wallis and Matalas (1970).

It is the main objective of this section to derive an approximate analytical expression for the small sample expectations of the estimate of lag- k serial correlation coefficient and in turn the amount of bias associated with the serial correlation coefficients of stationary stochastic processes in general but in particular of the ARIMA(1,0,1) process. The most commonly used estimating procedures of ρ and associated bias affects have been reviewed by Wallis and O'Connell (1972). They have considered the lag-one Markov process only and proposed a bias correction procedure which with some constraints gave a satisfactory solution. Apart from this O'Connell (1974) found the expected value of ρ in small samples for the ARIMA(1,0,1) process by employing Monte Carlo techniques which require a large computer investment, whereas analytical expressions for the same purpose have been derived in the following sections.

In hydrology the most commonly used estimate of ρ_k for a given sequence of observations is given by

$$\hat{\rho}_k = \frac{\sum_{i=1}^{n-k} x_i x_{i+k} - \frac{1}{(n-k)} \sum_{i=1}^{n-k} x_i \sum_{i=1}^{n-k} x_{i+k}}{\left[\sum_{i=1}^{n-k} x_i^2 - \frac{1}{(n-k)} \left(\sum_{i=1}^{n-k} x_i \right)^2 \right]^{1/2} \left[\sum_{i=1}^{n-k} x_{i+k}^2 - \frac{1}{(n-k)} \left(\sum_{i=1}^{n-k} x_{i+k} \right)^2 \right]^{1/2}} \quad (3.57)$$

This expression is an open-series approach in the sense that the first part of the series of length $n - k$, is used for x_i and the last part of length $n - k$ is used for x_{i+k} . Another way of estimating ρ_k is by the circular-series approach, in which the end of a sample series is assumed to be followed by its beginning, and the general form of which is given by Kendall (1954)

$$\hat{\rho}_k = \frac{\sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{i=1}^n x_i \right) \left(x_{i+k} - \frac{1}{n} \sum_{i=1}^n x_i \right)}{\sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{i=1}^n x_i \right)^2} \quad (3.58)$$

Finally, a third alternative method of estimating ρ_k is provided by Jenkins and Watts (1968) and Box and Jenkins (1970) as

$$\hat{\rho}_k = \frac{\sum_{i=1}^{n-k} \left(x_i - \frac{1}{n} \sum_{i=1}^n x_i \right) \left(x_{i+k} - \frac{1}{n} \sum_{i=1}^n x_i \right)}{\sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{i=1}^n x_i \right)^2} \quad (3.59)$$

which is claimed to be a more efficient estimate of ρ_k than the two previous estimates. This last estimate is also referred to as the maximum likelihood estimate of ρ_k . In their work Wallis and O'Connell (1972) have shown through computer experiments for the lag-one Markov process only that Eq.(3.59)

yields estimates of ρ_1 that have greater bias and smaller variance than those provided by Eq.(3.57)

3.7.1. Small Sample Expectation of ρ_k .

So far, in the literature on time series analysis to the best of the author's knowledge no exact results have been derived for the amount of bias in estimates of ρ_1 . However, a few approximate bias correction procedures exist that are currently employed.

One of the earliest approximate procedures was presented by Quenouille (1956) but it is only a rough method of removing bias in the estimate of the first order correlation coefficient and is expressed as,

$$\hat{\rho}_1^Q = 2\hat{\rho}_1 - \frac{1}{2}(\hat{\rho}_{1,1} + \hat{\rho}_{1,2}) \quad (3.60)$$

To apply the correction, the time series is divided into two equal halves each providing an estimate $\hat{\rho}_{1,1}$ and $\hat{\rho}_{1,2}$ of the first order serial correlation. If $\hat{\rho}_1$ is the estimate of ρ_1 for the whole series, then $\hat{\rho}_1^Q$ is supposed to be an unbiased estimate of the lag-one serial correlation coefficient. This bias correction procedure is independent of the underlying process. As stated by Wallis and O'Connell (1972), on the average, the variance of $\hat{\rho}_1^Q$ is greater than the variance of an unbiased $\hat{\rho}_1$ estimate by an alternative procedure proposed by Kendall (1954) which is not universally applicable to any process. However, if many short realizations of the same process were available then the mean value of $\hat{\rho}_1^Q$ would provide an acceptable estimate of ρ_1 . The bias correction procedure given by Kendall (1954) requires the time series considered to be normal and stationary. He also assumed the estimates of ρ_1 to be normally distributed and was able to derive general analytical results to order n^{-1} . It is through using this latter method that unbiased estimates of ρ_1 for the ARIMA(1,0,1) process are derived in the following sections.

One justification for the assumption that the estimate of ρ_1 is normally distributed is that the mathematical procedures employed are rendered into comparatively simple forms.

All of the estimates proposed in the previous section for the estimation of the autocorrelation coefficient from a given finite length of sample can in general be written as

$$\hat{\rho}_k = \frac{A}{\sqrt{B} \sqrt{C}} \quad (3.61)$$

where $\hat{\rho}_k$ denotes the small sample estimate of the k-th order autocorrelation coefficient, A is the k-th order covariance function and B and C are the variances of the two overlapping subsections of the actual series. It is obvious that A, B and C are all r.v's for a given length of historic sequence; hence they have their own pdf's and associated moments. Let the first order moments of A, B, and C be represented by E(A), E(B) and E(C) respectively. The terms on the r.h.s. of Eq.(3.61) can be written in another form which incorporates the first order moments of A, B and C and is given by

$$\hat{\rho}_k = \frac{E(A) + a}{\sqrt{E(B) + b} \sqrt{E(C) + c}} \quad (3.62)$$

where a, b and c are new r.v's obtained by measuring the original r.v's about their respective mean values. Therefore, the r.v's a, b and c will all have expectations equal to zero. However, the r.h.s. of Eq.(3.62) can be expanded into an infinite summation form by applying the Binomial expansion formulae to each term both in the numerator and denominator so that it will be easy and straightforward to take the expectation of both sides.

First of all Eq.(3.62) can be rewritten as

$$\hat{\rho}_k = [E(A) + a] [E(B) + b]^{-1/2} [E(C) + c]^{-1/2}$$

or

$$\hat{\rho}_k = \frac{E(A)}{\sqrt{E(B)E(C)}} \left[1 + \frac{a}{E(A)} \right] \left[1 + \frac{b}{E(B)} \right]^{-1/2} \left[1 + \frac{c}{E(C)} \right]^{-1/2}$$

When the Binomial expansion is applied to the last two product terms on the r.h.s. and the necessary algebra is performed then the following result is obtained

$$\begin{aligned} \hat{\rho}_k = & 1 - \frac{c}{2E(C)} + \frac{c^2}{8E^2(C)} - \dots - \frac{b}{2E(B)} + \frac{bc}{4E(B)E(C)} \\ & + \frac{3b^2}{8E^2(B)} - \frac{3b^2c}{16E^2(B)E(C)} + \frac{9b^2c^2}{64E^2(B)E^2(C)} + \frac{a}{E(A)} - \frac{ac}{2E(A)E(B)} \\ & + \frac{3ac^2}{8E(A)E^2(C)} - \frac{ab}{2E(A)E(B)} + \frac{abc}{4E(A)E(B)E(C)} - \frac{3abc^2}{16E(A)E(B)E^2(C)} \\ & + \frac{3ab^2}{8E(A)E^2(B)} - \frac{3ab^2c}{8E(A)E^2(B)E(C)} + \frac{9ab^2c^2}{64E(A)E^2(B)E^2(C)} + \dots \end{aligned}$$

(3.63)

A further simplification of Eq.(3.63) can be made by assuming that $B = C$. In the case of the maximum likelihood estimate given by Eq.(3.58) and the circular-series estimate given by Eq.(3.59), this equivalence is perfectly valid but it is an approximation when the estimate is provided by the open-series approach given by Eq.(3.57)

$$E(B) = E(C) , \quad E(b^2) = E(c^2) = E(bc) = \text{Var}b$$

Therefore, by taking expectations of both sides of Eq.(3.63), by making the necessary substitutions and by neglecting third and higher order terms the following result is obtained.

$$E(\hat{\rho}_k) \approx \frac{E(A)}{E(B)} \left[1 - \frac{E(b^2)}{E^2(B)} + \frac{E(a,b)}{E(A)E(B)} \right]$$

which yields,

$$E(\hat{\rho}_k) \approx \frac{E(A)}{E(B)} - \frac{\text{Cov}(a,b)}{E^2(B)} + \frac{E(A)\text{Var}b}{E^3(B)} \quad (3.64)$$

This last expression gives the approximate small sample expectation of $\hat{\rho}_k$ for any kind of normal stationary process. The same result has been given by Kendall (1954) without a derivation and he has applied it to lag-one and higher order autoregressive processes and to simple moving average processes. Kendall (1954) presented a bias correction to order n^{-1} based on the definition of $\hat{\rho}_k$ given in Eq.(3.57) which for the lag-one Markov process becomes,

$$E(\hat{\rho}_k) = \rho^k - \frac{1}{(n-k)} \left[\frac{1+\rho}{1-\rho} (1-\rho^k) + 2k\rho^k \right] \quad (3.65)$$

which for the first order autocorrelation coefficient, ρ , reduces to

$$E(\rho) = \rho - \frac{1}{(n-1)} (1+3\rho) \quad (3.66)$$

The bias correction to order n^{-1} for the circular series definition given in Eq.(3.58) has also been given as

$$E(\hat{\rho}_k) = \rho^k - \frac{1}{n} \left[\frac{1+\rho}{1-\rho} (1-\rho^k) + 3k\rho^k - k\rho^{n-k} \right] \quad (3.67)$$

and in particular for $k = 1$ and ignoring higher order terms,

$$E(\hat{\rho}) = \rho - \frac{1}{n} (1+4\rho) \quad (3.68)$$

A bias correction algorithm has been proposed by Wallis and O'Connell (1972) by rearranging Eq.(3.68) as

$$\rho = \frac{E(\hat{\rho}) + (1/n)}{1 - (4/n)} \quad (3.69)$$

When the sample estimate of ρ obtained through Eq.(3.59) is substituted for $E(\hat{\rho})$ in Eq.(3.69) then an unbiased estimate of ρ , denoted by $\hat{\hat{\rho}}$ is

obtained. For those who postulate that the generating mechanism of a hydrological process can be adequately approximated by the lag-one Markov model, the following steps are recommended for generating synthetic sequences which yield unbiased lag-one serial correlation coefficients

- (1) Estimate ρ from a given historical record by using Eq.(3.58).
- (2) Find an unbiased $\hat{\rho}$ value from Eq.(3.69) by replacing $E(\hat{\rho})$ by the sample estimate found in step (1)
- (3) Generate synthetic traces using lag-one Markov process given in Eq.(1.4). As a result generated synthetic traces will be unbiased as far as the lag-one serial correlation coefficient is concerned. When Eq.(3.57) is employed as estimate of ρ then the following bias correction is applied first:

$$\rho = \frac{E(\hat{\rho}) + 1/(n-1)}{1 - 3/(n-1)} \quad (3.70)$$

and the steps given above will be repeated for generating unbiased synthetic sequences.

In short, it can be said that if the generating mechanism is known to be the lag-one Markov process then for maintaining resemblance between historic and synthetic sequences in terms of ρ , bias correction procedures given by Eq.(3.69) and Eq.(3.70) will be considered whichever one serves for the purpose.

3.7.2. Small Sample Expectation of ρ for the ARIMA(1,0,1) Process.

The general expression given in Eq.(3.64) for finding the small sample expectation of $\hat{\rho}_k$ can also be applied to the ARIMA(1,0,1) process. First the open-series definition of estimation given in Eq.(3.57) is taken into consideration; hence, the small sample expectation of $\hat{\rho}_k$ to order n^{-1} is

found in appendix (3.2) to be

$$E(\hat{\rho}) = \rho \varphi^{k-1} - \frac{1}{(n-k)} \left[\frac{(1-\varphi+2\rho)(1+\varphi-2\rho^2\varphi^{k-1}) - \rho(1-\varphi^2)\varphi^{k-1}}{(1-\varphi^2)} + 2k\rho^2\varphi^{k-2} \right] \quad (3.71)$$

for $k = 1$ this expression reduces to

$$E(\hat{\rho}) = \rho - \frac{1}{(n-1)} \left[\frac{(1-\varphi+2\rho)(1+\varphi-2\rho^2)}{(1-\varphi^2)} + \rho \right] \quad (3.72)$$

where n is the size of sample considered, φ and Θ are the model parameters for the ARIMA(1,0,1) process. When $\varphi = \rho$ is substituted in Eq.(3.72) then the lag-one Markov process case given by Eq.(3.66) results.

However, if the circular definition given in Eq.(3.59) or Box and Jenkins definition given in Eq.(3.58) are taken as estimates of ρ then the small sample expectation derived in appendix (3.2) becomes,

$$E(\hat{\rho})_k = \rho \varphi^{k-1} - \frac{1}{n} \left[\frac{(1-\varphi+2\rho)(1-\rho\varphi^{k-1})}{(1-\varphi)} + k\rho(\varphi+2\rho)\varphi^{k-2} - k\rho\varphi^{n-k-1} - \frac{4\rho^2\varphi^{k-1}}{(1-\varphi^2)}(\varphi-\rho) \right] \quad (3.73)$$

which for $\hat{\rho}$ reduces to

$$E(\hat{\rho}) = \rho - \frac{1}{n} \left[\frac{(1-\varphi+2\rho)(1-\rho)}{(1-\varphi)} + 3\rho + \frac{4\rho^2(\varphi-\rho)}{(1-\varphi^2)} \right] \quad (3.74)$$

Furthermore, when $\varphi = \rho$ is substituted in Eq.(3.74) it then reduces to Eq.(3.68) which is corresponding expression for the case of the lag-one Markov process. The bias correction procedure for the ARIMA(1,0,1) process becomes rather complicated partly because of the fact that there are now two

parameters, namely, φ and ρ , which control the autocorrelation function and partly because of the non-linear appearances of Eq.(3.72) and Eq.(3.74). This difficulty can be eased to a greater extent by assuming that the population value of φ is known or its estimate $\hat{\varphi}$ extracted from a given historic record is assumed to be the population value. If this is the case then the following bias correction procedure to obtain an unbiased ρ value can be pursued.

Although it is not possible to solve Eq.(3.72) or Eq.(3.74) analytically in terms of ρ , their numerical solutions can easily be found on a digital computer; hence, tables for given values of a set consisting φ , θ , ρ and n can easily be prepared. One example of such a table is given for $\varphi=0.85$ and various values of θ , ρ and n , in Table (3.1). In order to show how to use the tables to obtain an unbiased estimate of $\hat{\rho}$ when population φ value is known a numerical example will be discussed. Let the estimate of ρ , φ and θ from a given historical record, using the estimation methods given by O'Connell (1974) for the ARIMA(1,0,1) process be $\hat{\rho} = 0.218$, $\hat{\varphi} = 0.850$ and automatically because of the relationship given in Eq.(1.24) $\hat{\theta} = 0.680$ for a given sequence of length $n = 40$. By assuming $\hat{\varphi} = 0.850$ to be the population value, in Table (3.1) and column 5 $E(\hat{\rho}) = 0.218$ is observed correspondingly from column 1 the estimate $\hat{\rho} = 0.360$ and from column 2 $\theta = 0.600$ are obtained; hence an unbiased estimate $\hat{\rho}$ is obtained. As a result by employing these newly found set of parameters in the generating scheme synthetic sequences which resemble historical sequence in terms of ρ can be obtained. This bias correction procedure can be applied only when the population φ value is known. Otherwise, $\hat{\varphi}$ must be estimated from the given historic data as the estimate of φ may be biased. The above bias correction procedure may not yield an unbiased estimate of ρ . On the other hand, if the population value of θ is known then tables similar to Table (3.1) can be prepared and similar kind of bias correction procedure be

applied.

The comparison of the bias corrections of the ARIMA(1,0,1) process with that of the lag-one Markov process, on the grounds that both have the same ρ will confirm the fact that the former model suffers from larger bias effects than the latter one. Another important point which has been already presented by Wallis and O'Connell (1972) for the lag-one Markov process is that sampling error in the uncorrected estimates may lead to corrected values of the lag-one serial correlation coefficient which is greater than one. As a result there appears to be upper and lower constraints for uncorrected serial correlation coefficient; Another alternative to avoid such a result is to include higher order terms in the bias correction algorithms. The similar situation has been encountered in the case of the ARIMA(1,0,1) process but the derivation of such a constraint has not been undertaken in this thesis.

3.8. Summary

The small sample properties of estimates of three parameters, namely, the variance, the standard deviation and the lag-one serial correlation coefficients are analytically investigated. First of all a literature review pertaining to the small sample properties of stationary normal processes has been given and the definition and the sources of bias are presented. Then general analytical expressions for bias corrections of standard deviation and the lag-one serial correlation coefficient are derived and bias correction procedures of these parameters are proposed.

ρ	θ	$\varphi = .85$				
		n=20	n=30	n=40	n=50	n=100
1	2	3	4	5	6	7
.203	.700	.005	.071	.104	.124	.163
.218	.680	.011	.080	.115	.135	.177
.250	.670	.023	.099	.137	.159	.205
.298	.640	.043	.128	.171	.196	.247
.329	.620	.058	.149	.194	.221	.275
.360	.600	.075	.170	.218	.246	.303
.406	.570	.102	.203	.254	.284	.345
.503	.500	.184	.295	.350	.383	.450
.619	.400	.286	.397	.452	.486	.552
.706	.300	.396	.499	.551	.582	.644

Table (3.1) - $E(\hat{\rho})_n$ values for the ARIMA(1,0,1) process

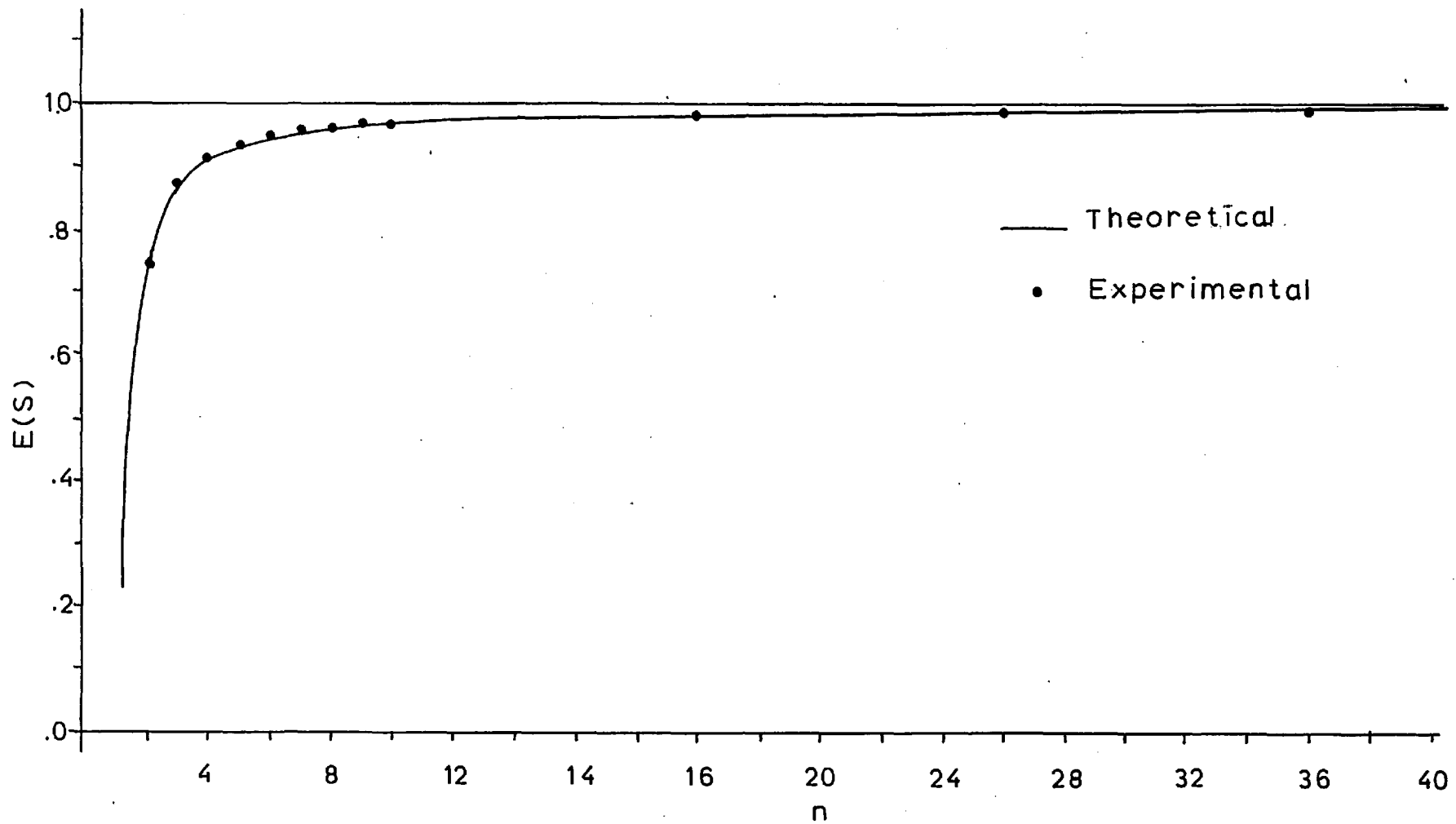


Figure (3.1) - Small sample expectation of S - NIP-

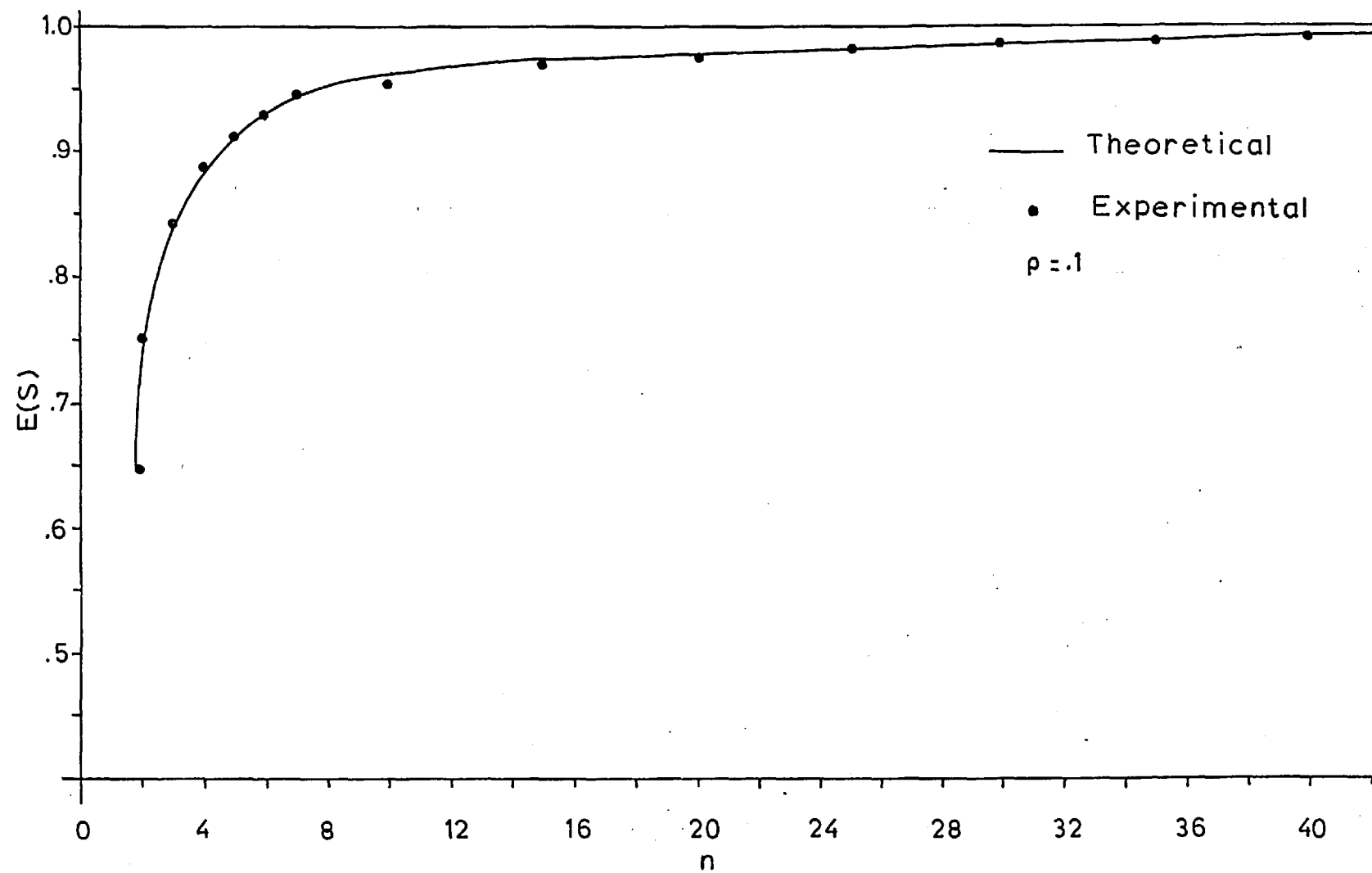


Figure (3.2) - Small sample expectation of S - Lag-One Markov process.

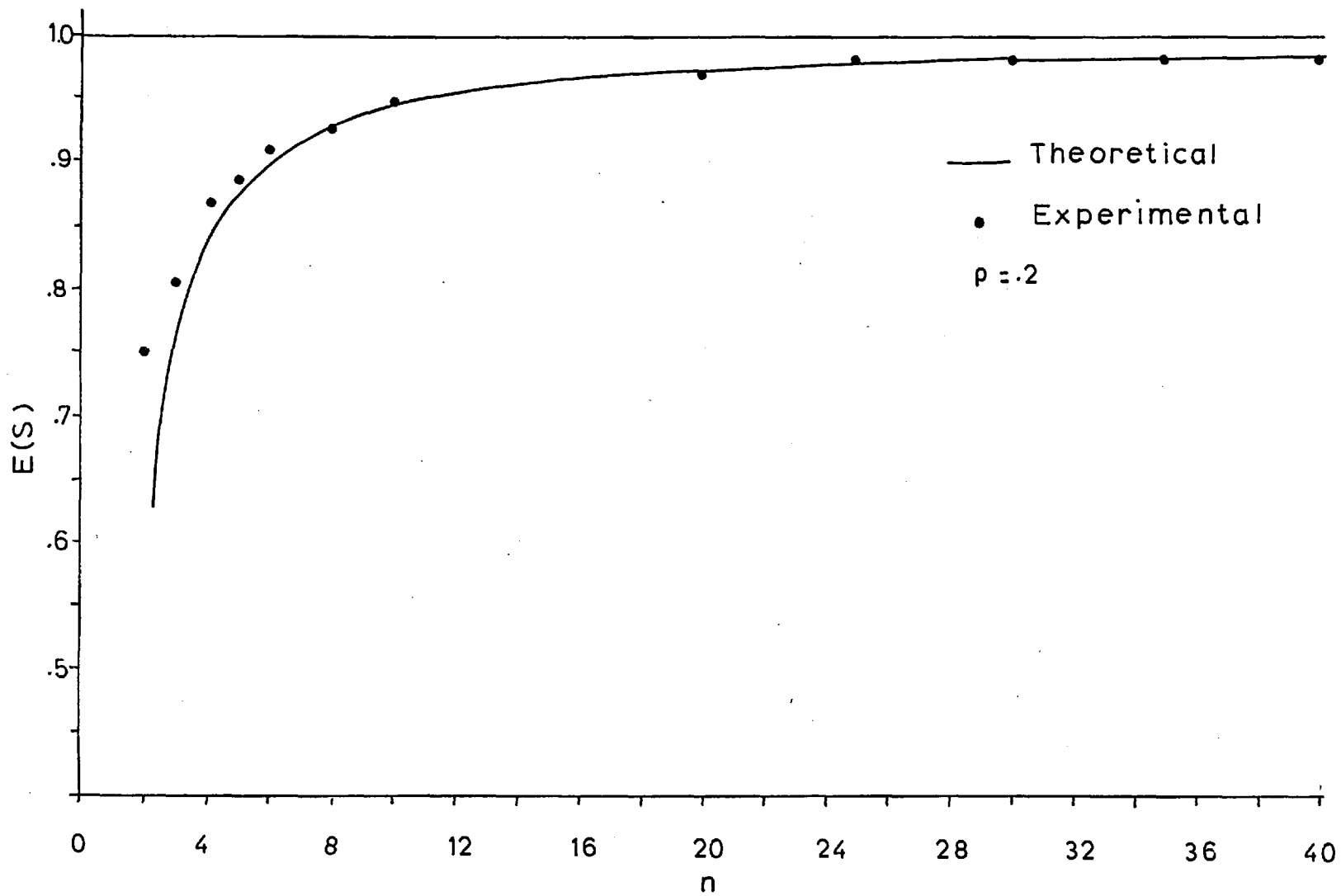


Figure (3.3) - Small sample expectation of S - Lag-One Markov process.

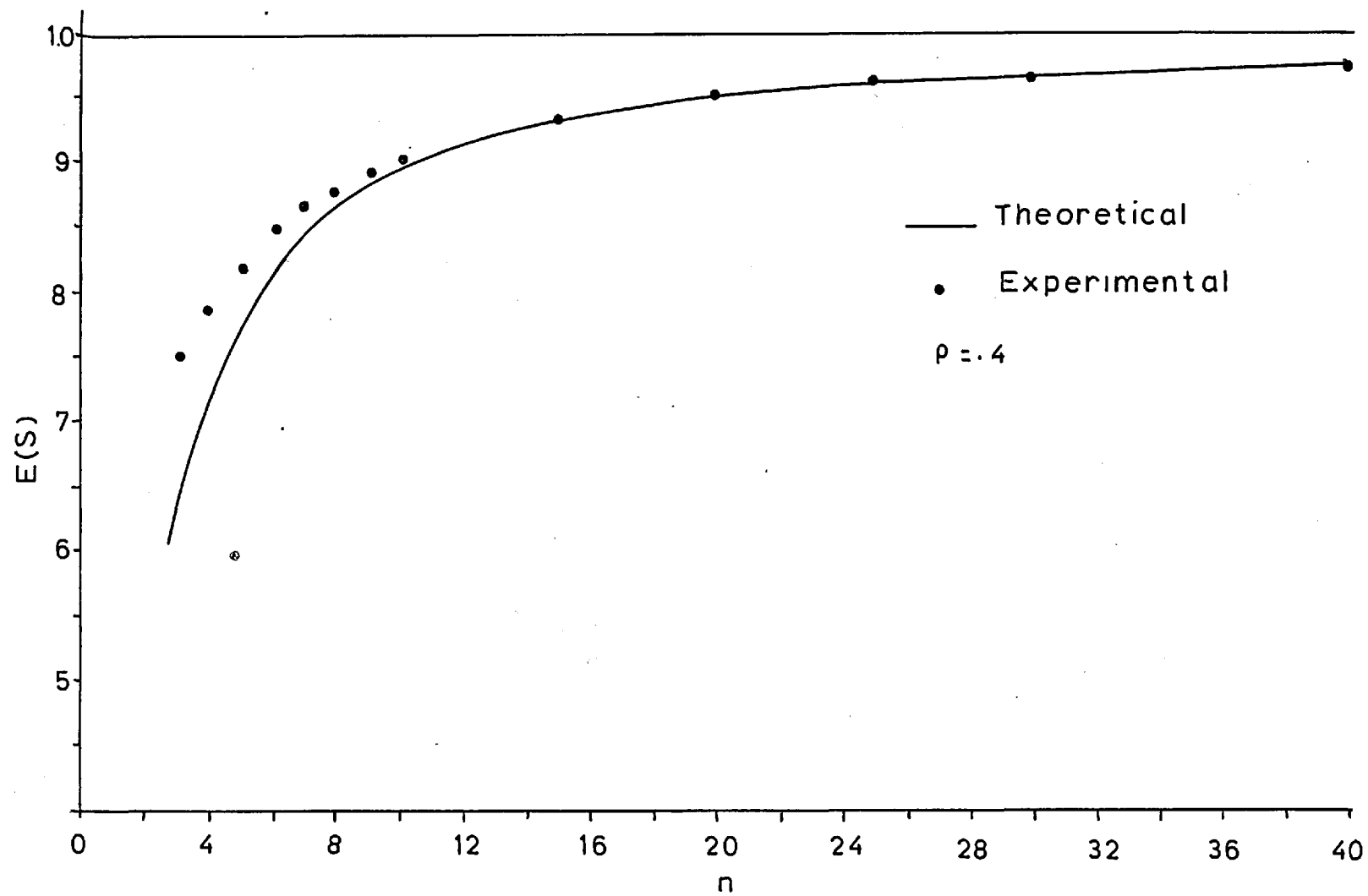


Figure (3.4) - Small sample expectation of S -Lag-One Markov process

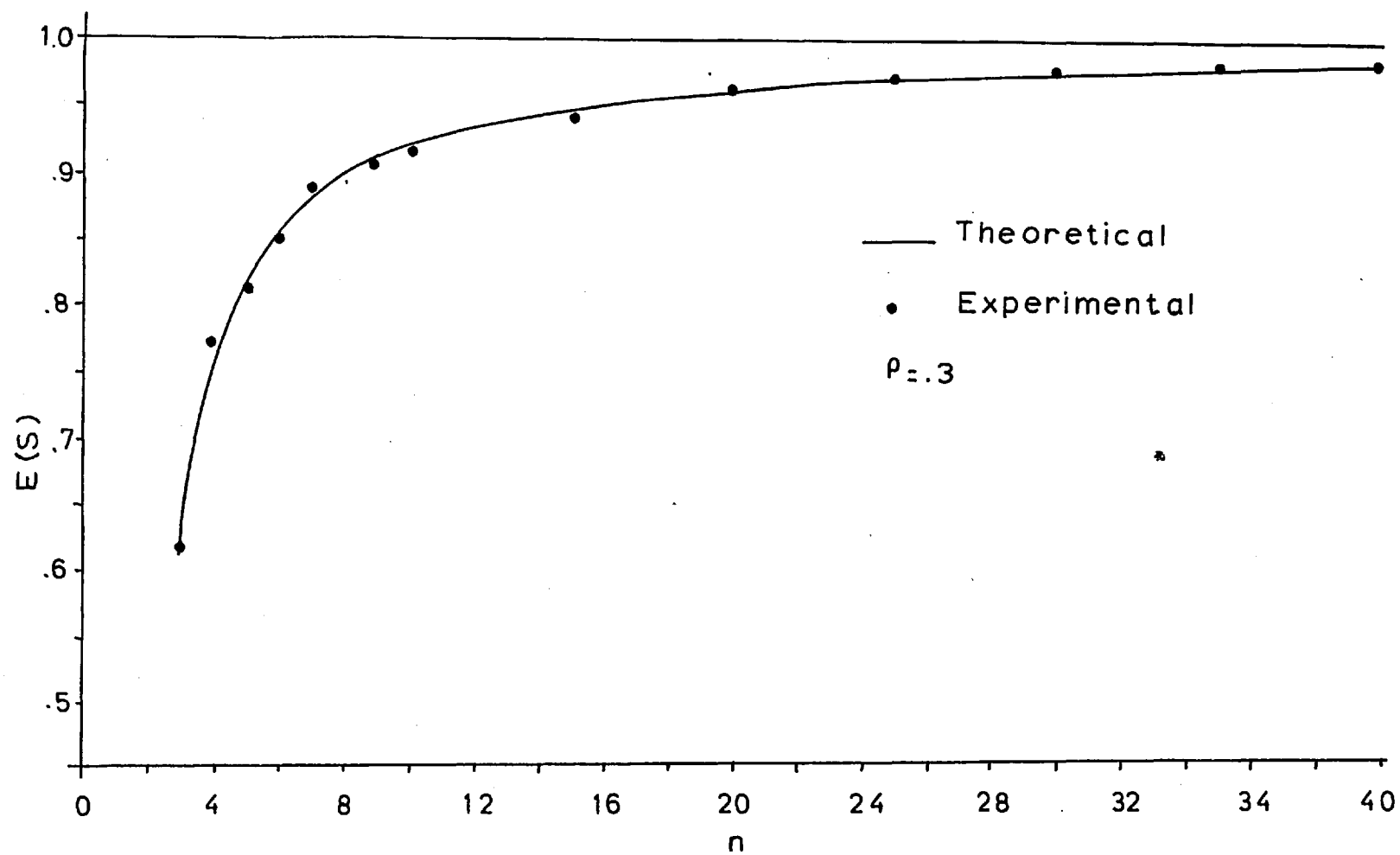


Figure (3.5) - Small sample expectation of S - Lag-One Markov process-.

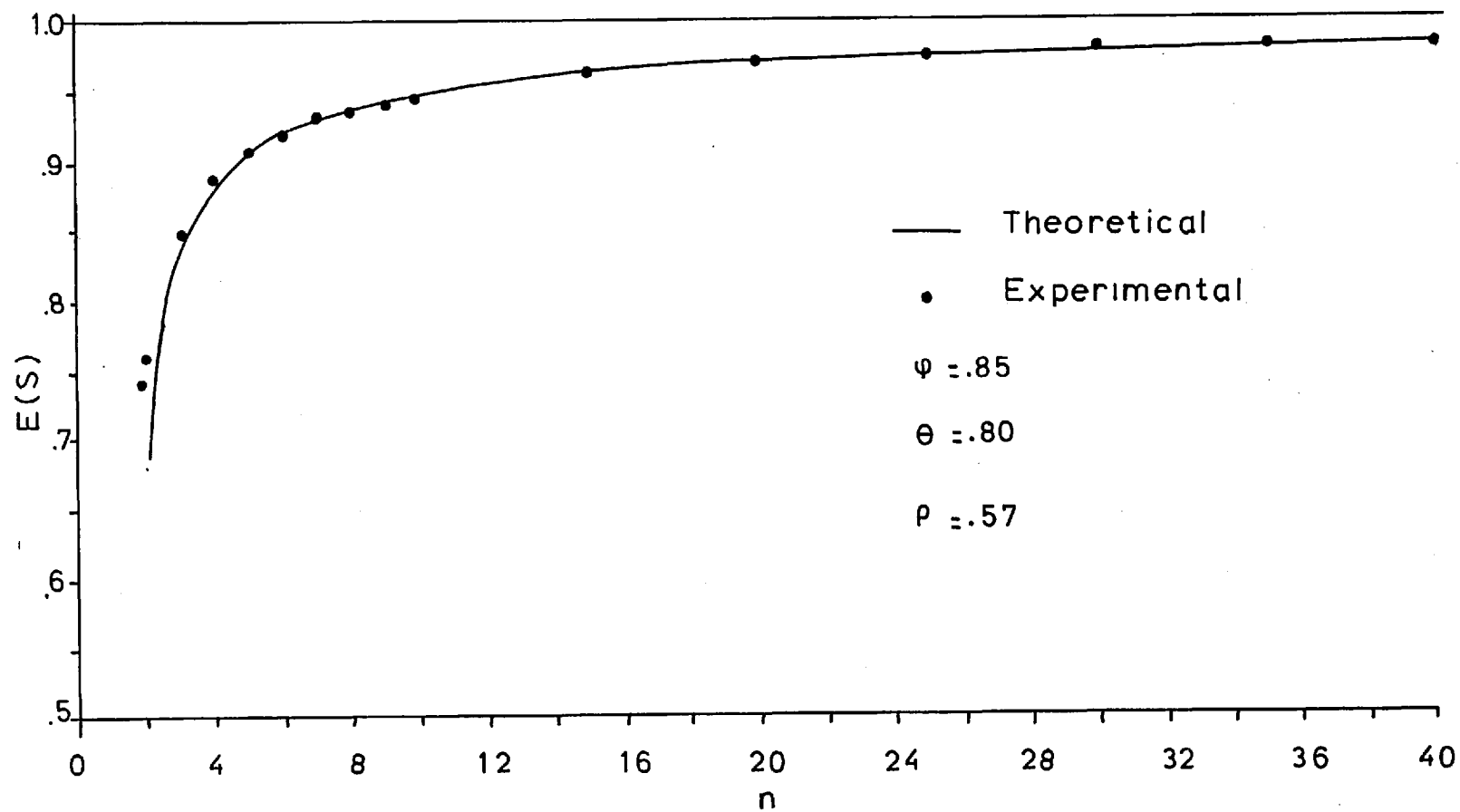


Figure (3.6) - Small sample expectation of S -ARIMA(1,0,1) process--.

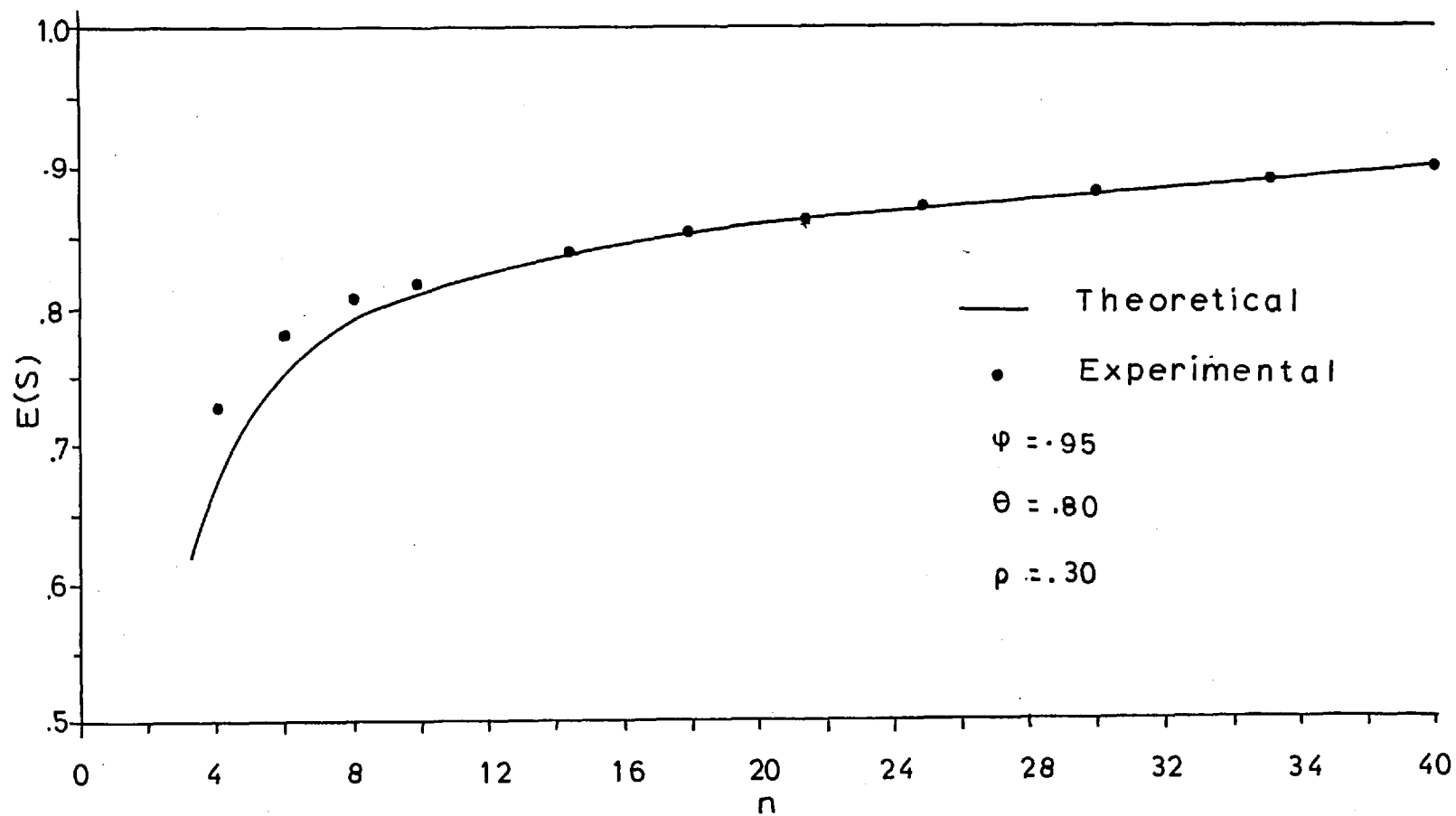


Figure (3.7) - Small sample expectation of S - ARIMA (1,0,1) process - .

Chapter 4

THE WHITE MARKOV PROCESS

4.1 General

The number of stochastic processes that are applicable to hydrological phenomena has been increasing recently, because of the need to preserve the most important descriptive properties of a historic hydrological sequence which constitutes a time series. It is necessary to employ an appropriate stochastic process to generate equally likely synthetic flow sequences that are statistically indistinguishable from the observed sequence, in the sense that, the statistical parameters obtained from both the ensemble of synthetic sequences and the observed sequence tend to be the same in the long-run.

As noted in Chapter 1, a number of models have been developed for modelling annual streamflow. Amongst these, the ARIMA (1,0,1) process is one of the more desirable because of its suitable long-term and short-term properties. In particular, its small sample properties can be derived either by analytical or Monte Carlo techniques.

An intriguing property of this process is that its autocorrelation function is identical to that of a process consisting of the superposition of a lag-one Markov process and NIP which is referred to here as the White Markov process. Consequently, its properties are of interest in modelling annual streamflows, particularly the multisite formulation with which some difficulties were encountered for the ARIMA (1,0,1) process.

In section (4.2) the statistical properties of the WM process are studied and the necessary relationships between the model parameters and the low order moments of the p.d.f are obtained. In this context, it is shown that the ARIMA (1,0,1) process and the WM process are completely identical to each other on the basis of their autocorrelation functions.

Moreover, relationships between the autocorrelation functions of these processes which provide the transition from one process to another are derived. The adequacy of the WM process to model Hurst's law is discussed in section (4.3) and through "box diagrams" a good agreement of the process with Hurst's law is observed. Small sample properties of the standard deviation, variance and serial correlation coefficient are studied analytically in section (4.4). Section (4.5) mentions the application of the WM process to non-Gaussian annual streamflows, where the preservation of coefficient of skewness is important in the generating scheme. In the case of log-normally distributed annual streamflow sequence the application of the WM process to such a situation is shown in section (4.6). The multivariate case of the model is studied in section (4.7), and a comparison with the multisite ARIMA (1,0,1) model is made.

4.2. Description of the White Markov Process.

The appeal of this process as a potential model in synthetic hydrology derives mainly from two functions. Firstly, it is computationally simple and may be expressed as a recurrence relation and secondly the estimates of two most important parameters namely the first order serial correlation, ρ_1 , and the Hurst coefficient, h , can be preserved in generated sequences. Those two properties have also been fulfilled by the ARIMA (1,0,1) process because the WM process is very closely related to the ARIMA (1,0,1) process. For the WM process, although it lies within the Brownian domain of attraction, by convenient choice of its parameters, it is possible to preserve high values of h when ρ_1 is small as will be shown later in this chapter. The WM process can also be considered to be one of the simplest fast fractional Gaussian

process noise approximations which has been proposed by Mandelbrot (1971) to model Hurst's law for hydrologically long sequences. As with other processes belonging to the Brownian domain of attraction, the WM process also yields $h=0.5$ asymptotically for very large sequence lengths.

However, basically the model is a mixture of two very well known stochastic processes, namely the normal independent process, NIP, or white-noise and the lag-one Markov process. The operation performed to get the mixture is to superimpose the white-noise process independently on to the lag-one Markov component ^{which} can be written as

$$X_t - \mu_M = \rho (X_{t-1} - \mu_M) + \sigma_M \sqrt{1 - \rho^2} \epsilon_t \quad (4.1)$$

where X_t and X_{t-1} are the variate values at time points t and $t-1$ respectively, ϵ_t is a r.v. at time t and independent of X_{t-1} , μ_M and σ_M denote the mean and the standard deviation of the process respectively, and finally ρ is the first order autocorrelation coefficient of the Markovian process. The random component ϵ_t of Eq. (4.1) is normally and independently distributed with zero mean and unit variance. The second constituent of the WM process is a process of normally and independently distributed r.v's, η_t with zero mean and unit variance. The independent superimposition of white noise onto Markov process defined by Eq. (4.1) leads to

$$Z_t = X_t + \beta \eta_t \quad (4.2)$$

where β is a model parameter representing the standard deviation of the independently superimposed white noise component.

By substituting Eq. (4.1) into Eq. (4.2) a more explicit description of the new model emerges, that is

$$Z_t = \mu_M + \rho_M (X_{t-1} - \mu_M) + \sigma_M \sqrt{1 - \rho_M} \epsilon_t + \beta \eta_t \quad (4.3)$$

The reason for giving subscripts to ρ_M and σ_M is because of the fact that the overall autocorrelation function and standard deviation of the WM process will be different from that of its Markov component as will be shown in the following section. However, the actual statistical parameters of the process will be a function of σ_M , ρ_M and β only. In order for a process to be applicable in practical situations, its parameters must be related to parameters estimated from a given historical record because of the fact that these historic parameters must be reproduced in the synthetic sequences so that the statistical resemblance between the two will be achieved. In the derivation of the statistical properties of this process Eq. (4.2) will be considered; hence by taking the expectation of both sides,

$$E(Z_t) = E(X_t) + \beta E(\eta_t)$$

and because of $E(\eta_t) = 0$, the result becomes

$$E(Z_t) = E(X_t)$$

which states that the population mean value of the WM process in its form given by Eq.(4.2) is equal to the mean value of its Markov component denoted by μ_M . Hence,

$$\mu = E(Z_t) = \mu_M \quad (4.4)$$

where μ denotes the population mean value of the WM process.

To find the variance first the square of both sides of Eq. (4.2) is taken and the expectation operator is then applied.

$$E(Z_t^2) = E(X_t^2) + 2\beta E(X_t \eta_t) + \beta^2 E(\eta_t^2)$$

where

$$E(X_t^2) = \sigma_M^2 + \mu_M^2$$

and the product term is equal to zero because of the fact that Markov component, X_t , and the WM component, η_t , are superimposed independently.

Hence, the above expression turns out to be

$$\sigma^2 = \sigma_M^2 + \beta^2 \quad (4.5)$$

therefore the variance of the WM process is equal to the sum of the variances of the two main constituents.

To find the autocorrelation function for lag k , in general, the following approach is adopted. The value of the process at time instant t is

$$Z_t = X_t + \beta \eta_t$$

and at the time instant $t+k$ is given as

$$Z_{t+k} = X_{t+k} + \beta \eta_{t+k}$$

By multiplying corresponding sides of these two last expressions and then taking expectations the result becomes

$$E(Z_t \cdot Z_{t+k}) = E(X_t \cdot X_{t+k}) + \beta E(X_t \cdot \eta_{t+k}) + \beta E(X_{t+k} \cdot \eta_t) + \beta^2 E(\eta_t \cdot \eta_{t+k})$$

where

$$E(Z_t \cdot Z_{t+k}) = \sigma^2 \rho_k$$

$$E(X_t \cdot X_{t+k}) = \sigma_M^2 \rho_M^k$$

and because of the independence of η_t from X_t

$$E(X_t \eta_{t+k}) = E(X_{t+k} \eta_t) = E(\eta_t \eta_{t+k}) = 0$$

Therefore,

$$\sigma^2 \rho_k = \sigma_M^2 \rho_M^k \quad (4.6)$$

This equation is the relationship between the autocovariance of the overall process and the autocovariance function of its Markov constituent. As a result the autocorrelation function of the WM process becomes,

$$\rho_k = \frac{\sigma_M^2}{\sigma^2} \cdot \rho_M^k \quad (4.7)$$

Substitution of Eq.(4.5) into Eq.(4.7) leads to

$$\rho_k = \frac{1}{1 + \left(\frac{\beta}{\sigma_M}\right)^2} \rho_M^k \quad (4.8)$$

By defining a new variable α as,

$$\alpha = \frac{1}{1 + \left(\frac{\beta}{\sigma_M}\right)^2} \quad (4.9)$$

which is a function of the ratio of the standard deviation of the white noise and Markov constituents, Eq.(4.8) can be written in a more concise form as

$$\rho_k = \alpha \rho_M^k \quad (4.10)$$

At first glance it is obvious from this last equation that the autocorrelation structure of the WM process is very similar to the ARIMA (1,0,1) process autocorrelation structure given in Eq.(1.21) and Eq.(1.22). In fact, for the first order autocorrelation coefficient Eq.(4.10) yields

$$\rho = \alpha \rho_M \quad (4.11)$$

By re-writing Eq.(4.10) in terms of ρ the following is reached

$$\rho_k = \rho \rho_M^{k-1} \quad (4.12)$$

Then, by comparing Eq.(4.11) and Eq.(4.12) with the ARIMA (1,0,1) autocorrelation structure given by Eq.(1.21) and Eq.(1.22) the following equalities between the parameters of the two processes become apparent, these are

$$\rho_M = \varphi \quad (4.13)$$

and

$$\alpha = \frac{\rho}{\rho_M} \quad (4.14)$$

where ρ is the first order autocorrelation coefficient to be preserved by the two processes and its relation to the ARIMA (1,0,1) process parameters φ and θ is given by Eq.(1.21). Hence, considering Eq.(4.14)

$$\alpha = \frac{(1 - \varphi\theta)(\varphi - \theta)}{\varphi(1 + \theta^2 - 2\varphi\theta)} \quad (4.15)$$

Therefore, Eq.(4.13) and Eq.(4.15) show the full correspondance of the autocorrelation structure of the two processes. As a result all of the advantages claimed for the ARIMA (1,0,1) process for modelling Hurst's law by O'Connell (1974) will also be valid for the WM process. In the case of the WM process the most prominent factor that plays an important role in the long-term behaviour of a sequence generated by the process is the lag-one autocorrelation coefficient ρ_M , of the Markov constituent whereas, for the ARIMA (1,0,1) process this is the parameter φ . On the other hand a logical interpretation of the parameter α of the WM process can be furnished by investigating Eq.(4.8) more closely. By introducing a new variable as

$$\lambda = \frac{\beta^2}{\sigma_M^2} \quad (4.16)$$

where λ always assumes a non-negative value because of the fact that it is the ratio of two variances which are always positive Eq.(4.9)

can be re-written as

$$\alpha = \frac{1}{1 + \lambda} \quad (4.17)$$

Eq.(4.16) reveals the fact that α is a positive value changing between zero and 'one' depending on the value that λ takes. There exist two possible extreme situations. One is that β is actually zero, in which case the WM model is merely equal to the lag-one Markov process. The same situation exists when the ratio β^2 / σ_M^2 is approximately zero, which requires that σ_M is very large compared with β , and as a result of this situation from Eq.(4.16) λ becomes zero and in turn from Eq.(4.17) α is equal to 1. The second extreme case can occur, either when $\sigma_M = 0$ in which case the WM process is the same as the white noise process or if β is very large compared with σ_M so that λ becomes, theoretically, infinite; as a result $\alpha = 0$. Between these two extremes, there exists a mixture resulting in a WM process. A graphical description of these extremes and their mixtures have been given in figure (4.1). Various combinations associated with α or λ value can easily be depicted from such a graph. Thus it becomes apparent that the key parameter is either λ which will be called the mixing ratio or its function, α , defined by Eq.(4.17) and referred to here as the mixing factor. Therefore, knowledge of the value of any one of these two related parameters will represent the relative effect of components on the final process, that is the WM process. When α is closer to zero this is tantamount to saying that the white noise behaviour is more pronounced in the WM process which in turn means that the process is richer in high frequencies and as a result it can be concluded that the transient effect in Hurst's law will be shorter. On the contrary, if α is closer to one, that is the Markov behaviour is predominant in the WM process, the low frequencies will be abundant in the overall process, which is quantitatively equal to

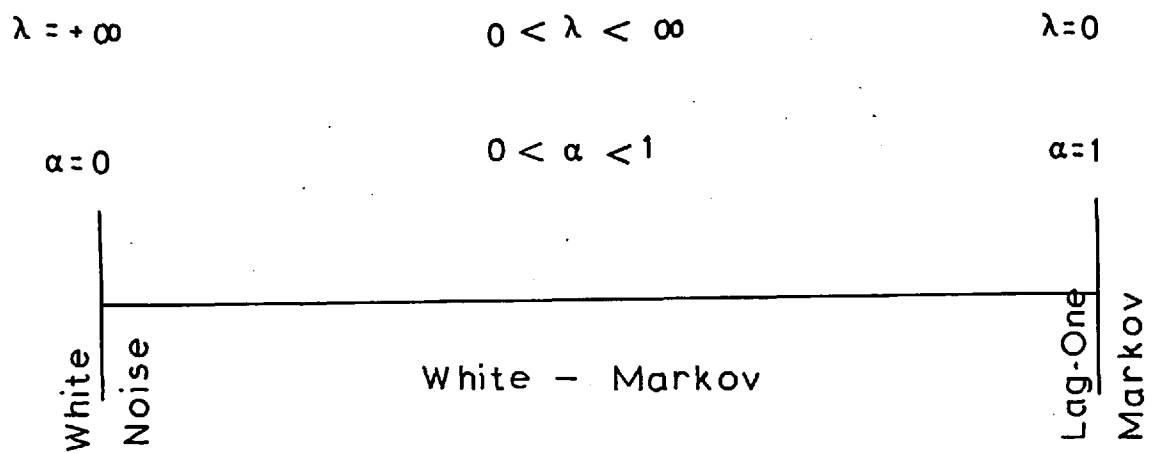


Figure (4.1) - Various WM processes

saying that the transient effect in Hurst's law increases. Consequently, by increasing the mixing factor, α , only, and keeping ρ_M fixed, it becomes possible to obtain a more persistent WM process and conversely for a fixed α value, by increasing ρ_M , the WM process can again be made more persistent, or a variable combination of α and ρ_M can be combined to obtain a more persistent WM process.

However, in choosing the parameters α and ρ_M in order to preserve a fixed overall ρ value there appears to be a restriction imposed on α and ρ_M . The relevant relationship among these three parameters is given by Eq.(4.11) and for various constant ρ values the plotting of ρ_M versus α has been given in figure(4.2) which indicates the relevant ranges of α and ρ_M in order to obtain a fixed ρ value. Each one of the curves shown in Figure(4.2) is a rectangular hyperbola. For the sake of argument only, if ρ were found to be equal to 0.2 then, it is clear from this figure that the restrictions on α and ρ_M are $0.2 < \alpha, \rho_M < 1.0$ and in general the parameters restriction is given by

$$\rho < \alpha, \rho_M < 1.0 \quad (4.18)$$

α and ρ_M are chosen in such a way that their product will be equal to ρ as given in Eq.(4.11)

4.3. Preservation of Hurst's Law

The WM process which theoretically lies within the Brownian domain of attraction, is characterized by an asymptotic h value equal to 0.5. However, the finite memory considerations of the same process will yield estimates of h that are greater than 0.5 over long transients and on the basis of this fact the process can be claimed to be adequate to model Hurst's law, as in the case of the ARIMA(1,0,1) process proposed by O'Connell (1971).

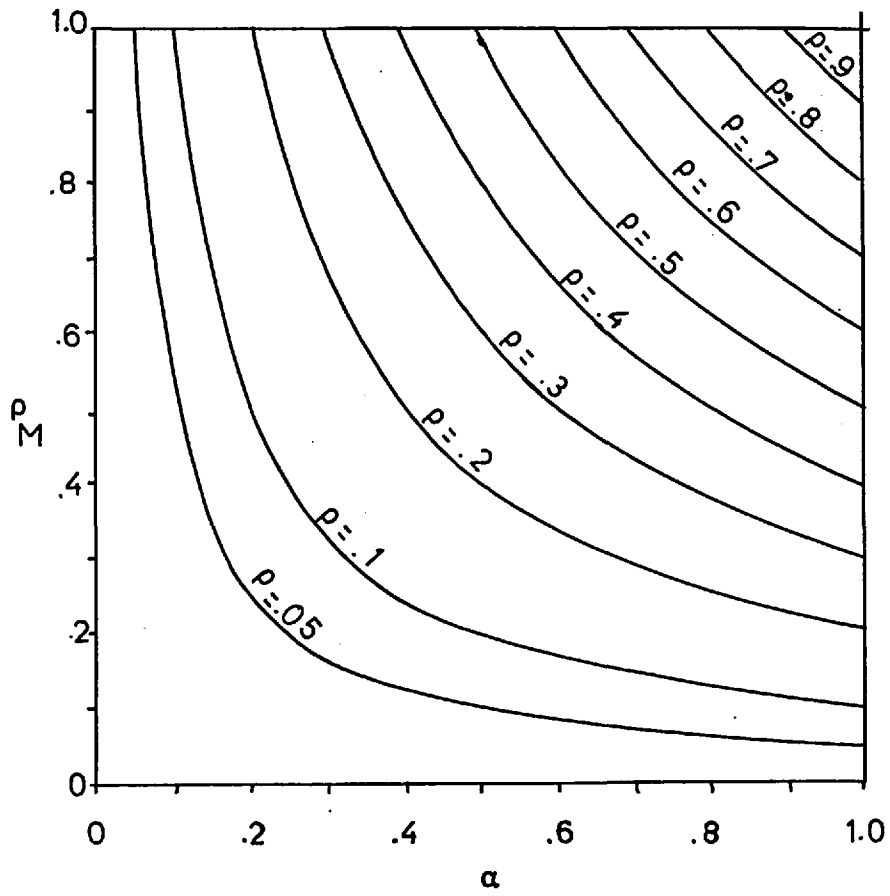


Figure (4.2) - Parameter restriction

In order to illustrate the ability of the WM process to model Hurst's law for values n of comparable with the longest geophysical records available a simulation approach first proposed by Mandelbrot and Wallis (1969b) has been adopted here. A set of α and ρ_M values of interest have been employed in generating 9000 synthetic observations for constructing a 'pox diagram' (Mandelbrot and Wallis, 1969b). The fitting of a straight line to the pox diagram has been achieved by the least squares procedure used by Wallis and Matalas (1970) and hence the slope of the pox diagram which provides an estimate, H , of the population Hurst coefficient, h , has been calculated. About a hundred pox diagrams for various combinations of α and ρ_M have been plotted by computer and only five of them have been presented in figures (4.3) - (4.7). The visual inspection of these five graphs confirms an overall agreement of the WM process with Hurst's law.

While it must be realized that h equals to 0.5 ultimately for the WM process, but as figures (4.3) - (4.7) reveal the estimate H is greater than 0.5 over a finite initial transient of length n_m . When n_m is exceeded then the population $h = 0.5$ will be attained. From the definition of the autocorrelation function of the WM process given by Eq.(4.12) it is obvious that ρ_M controls the 'effective memory' of the process which is defined as the lag- k , at which ρ_k becomes effectively zero. It can be said at this point that ρ_M plays the same role on the autocorrelation structure of the WM process as ϕ for the ARIMA(1,0,1) process. As ρ_M decreases, so does n_m , and as a result n_m is intimately related to the effective memory. Another conclusion obtained for the ARIMA(1,0,1) process by O'Connell (1974) states that for small values of H , n_m , in general tends to exceed the effective memory; however, for larger values of H , n_m approaches the 'effective memory'. This conclusion is valid also for the WM process; a similar result has also been obtained for the type 1 approximation to dfgn by Mandelbrot and Wallis (1969b).

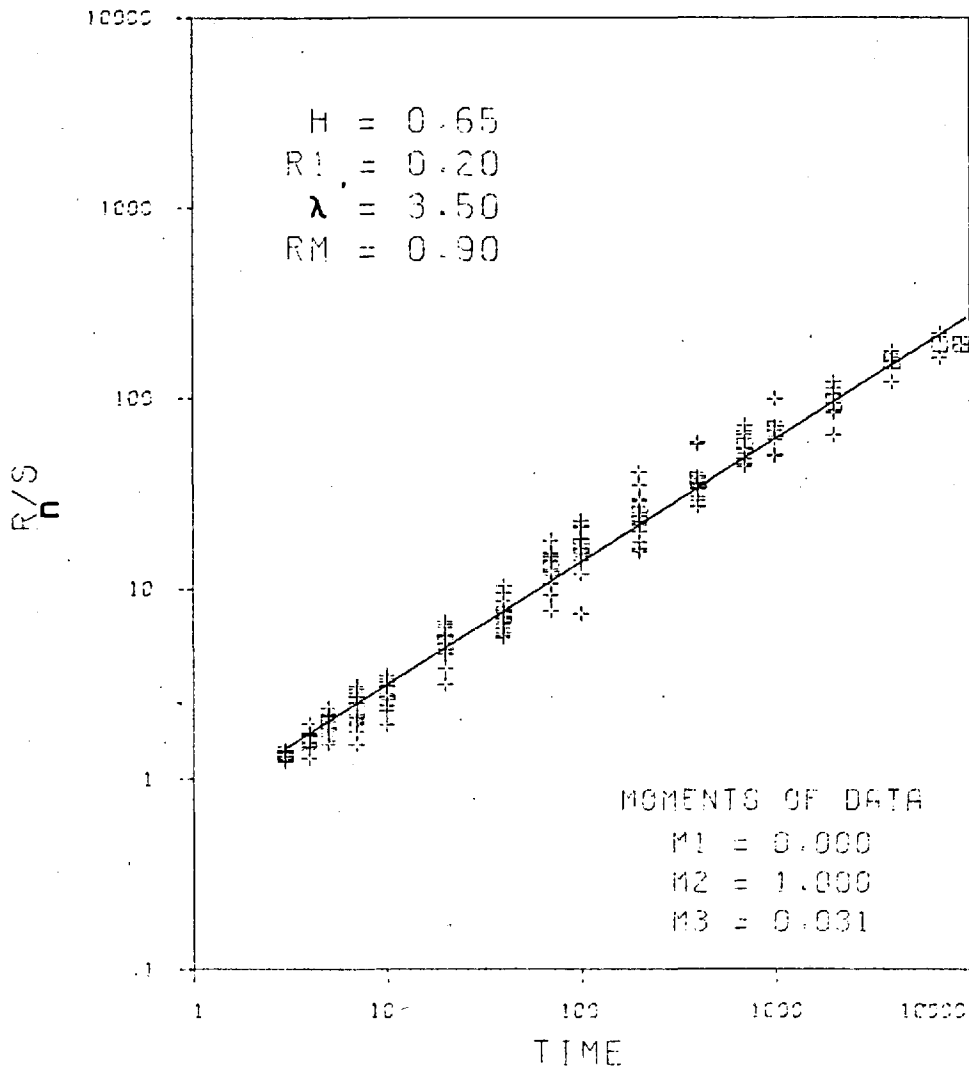


Figure (4.3) - Pox diagram of $\ln(R_n/S)$ versus $\ln(n)$ (time) for a series of 9000 values from a WM process with $\alpha = .222$. The mean values of R_n/S have been plotted as little squares. Note that p_1 is denoted as R_1 .

The selected values of subsample lengths were 3,4,5,7,10,20, 40,70,100,200,400,700,1000,2000,4000,7000,9000. For every value of $n_s \geq 500$, the subsample starting time was made equal to 1,100,200,.....,1400. For every value of $n_s \geq 500$, the subsample starting time was made equal to 1000,2000,...8000 or $(9000 - n_s + 1)$ whichever was the smaller (as used by Mandelbrot and Wallis 1969b).

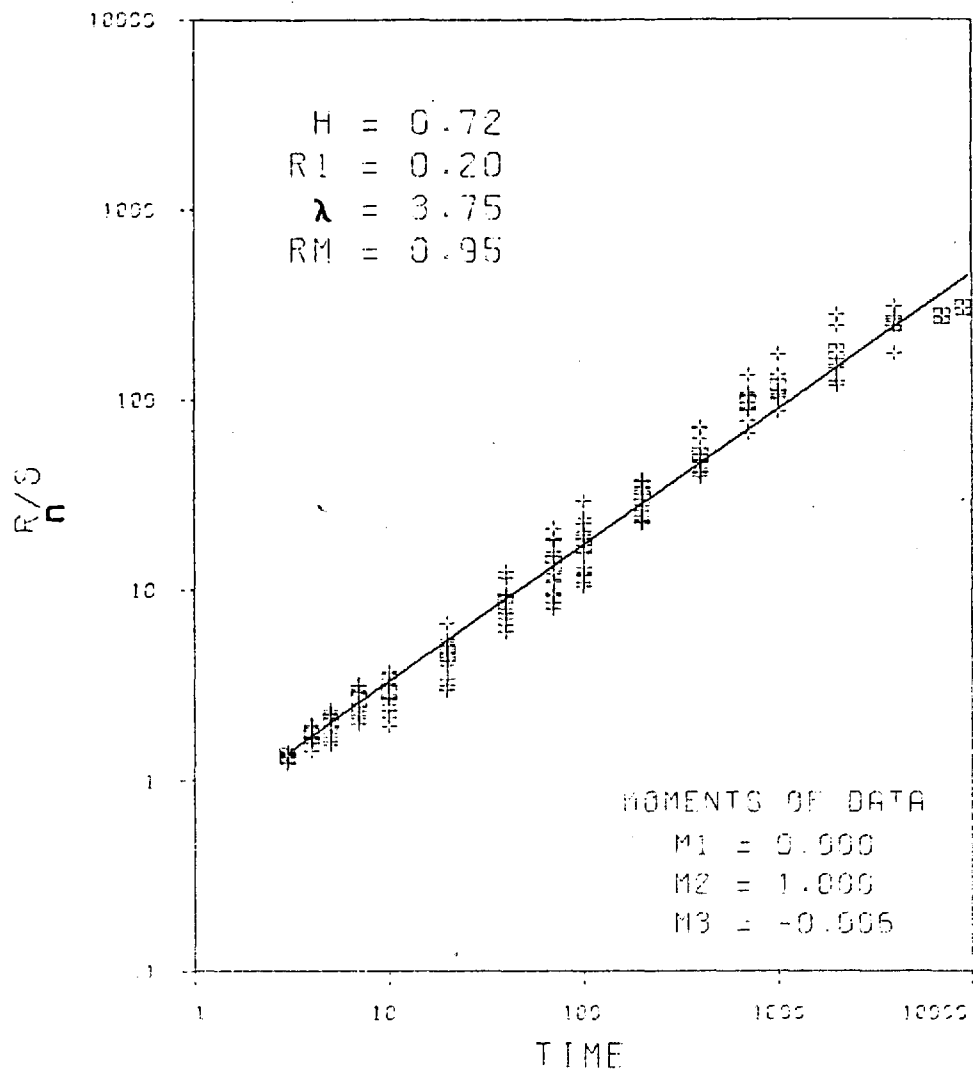


Figure (4.4) - Pox diagram of $\ln(R/S)$ versus $\ln(n)$ (time) for a series of 9000 values from a WM process with $\alpha = .211$. Details are as for figure (4.3).

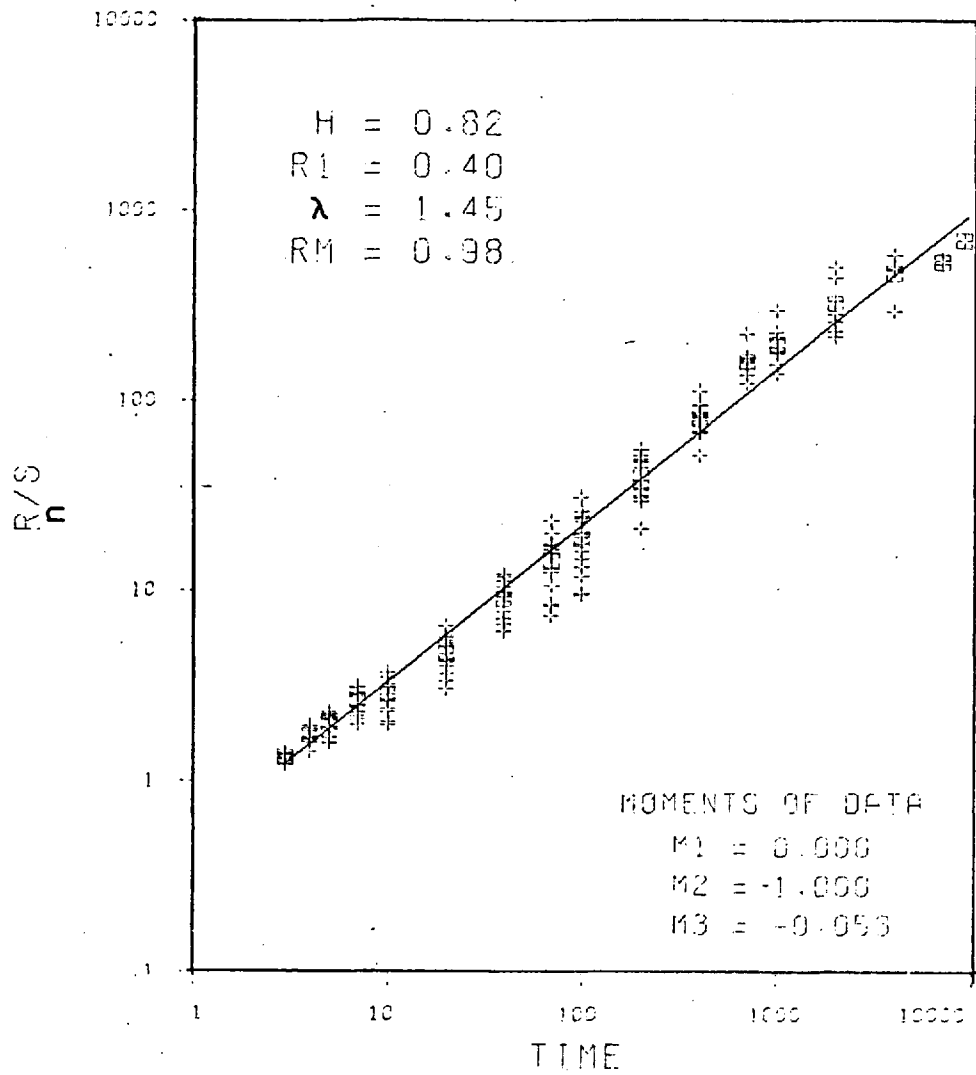


Figure (4.5) - Pox diagram of $\ln(R/S)$ versus $\ln(n)$ (time) for a series of 9000 values from a WM process with $\alpha = .310$. Details are as for figure (4.3).

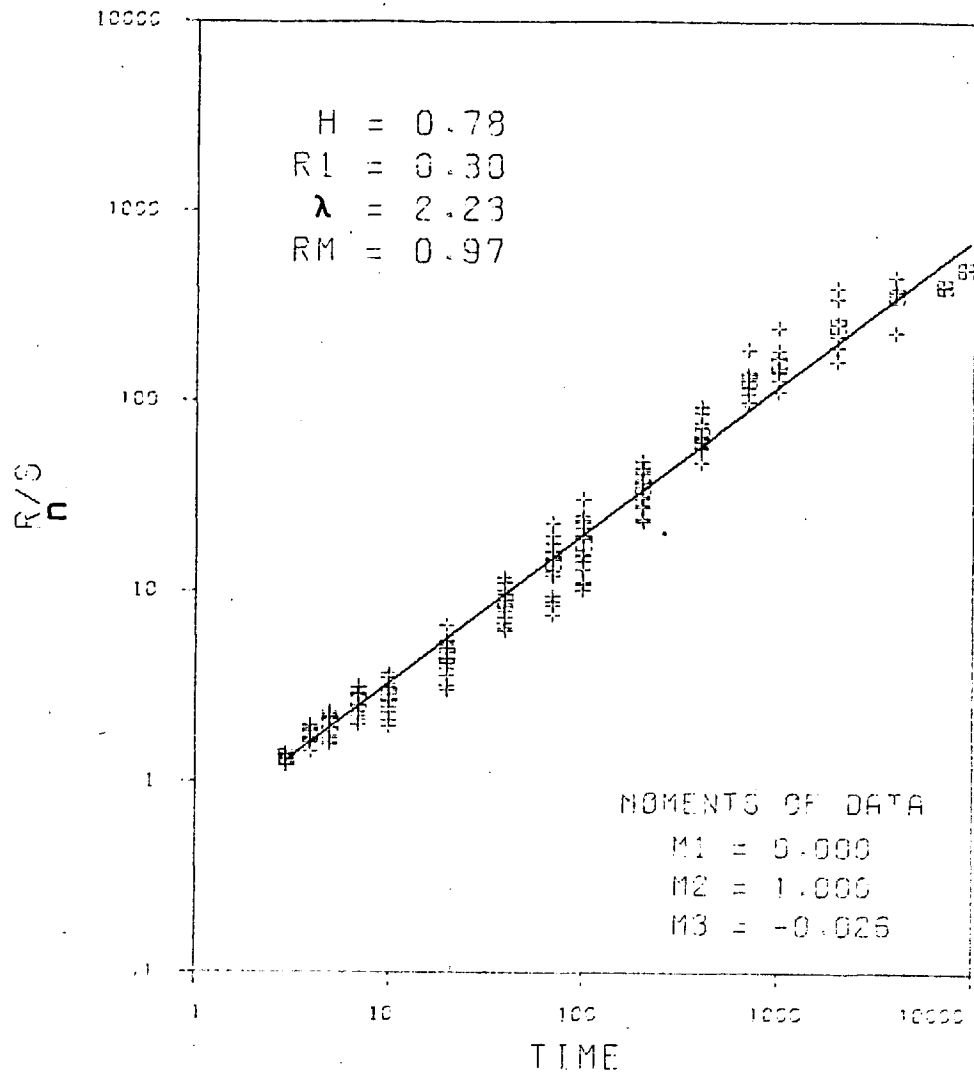


Figure (4.6) - Pox diagram of $\ln(R/S)$ versus $\ln(n)$ (time) for a series of 9000 values from a WM process with $\alpha = .408$. Details are as for figure (4.3).

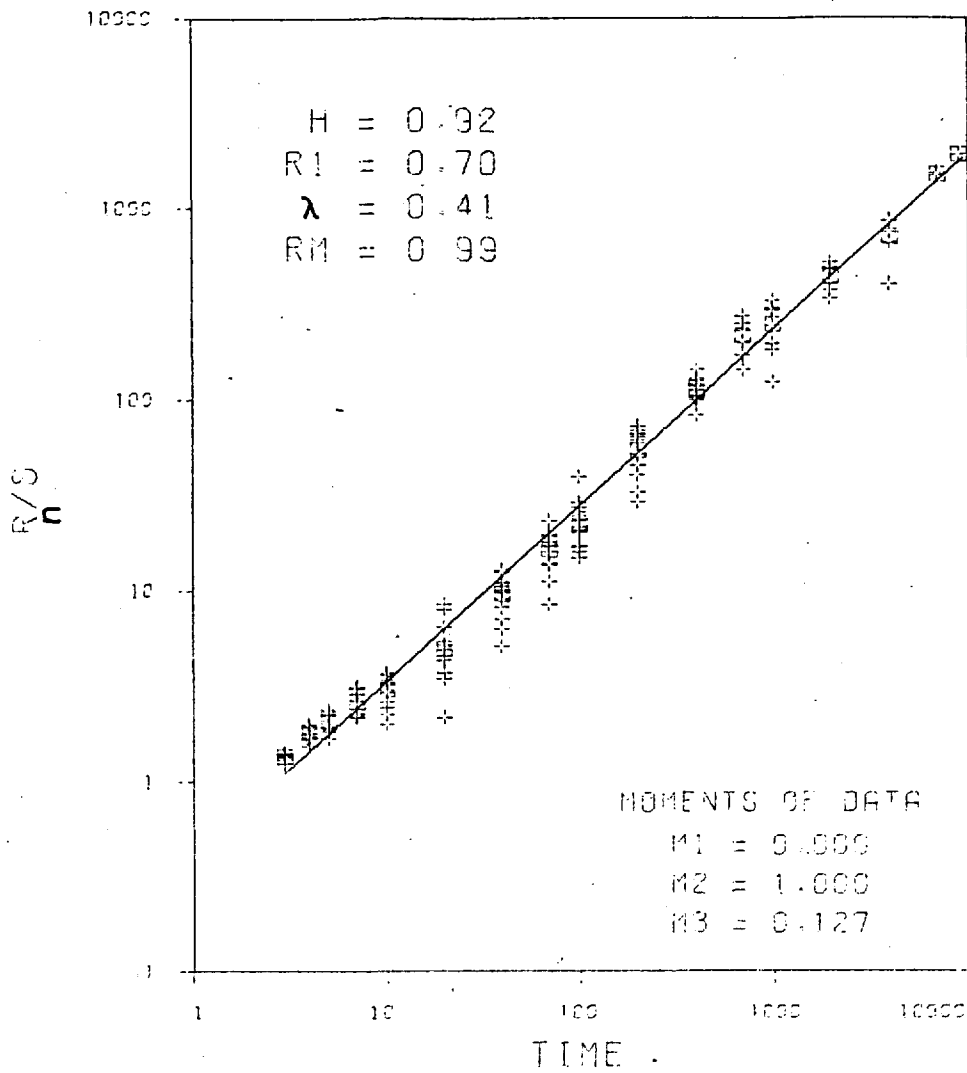


Figure (4.7) - Pox diagram of $\ln(R/S)$ versus $\ln(n)$ (time) for a series of 9000 values from a WM process with $\alpha = .708$. Details are as for figure (4.3).

The "pox diagrams" are governed solely by the autocorrelation function i.e. for a given correlogram and n value there corresponds one pox diagram and in turn one estimate H only, whatever the underlying generating process is. Correspondingly all of the results obtained for the ARIMA(1,0,1) process by O'Connell (1974) can be used for the WM process provided that the necessary conversions of φ and θ in terms of α and ρ_M are considered as given by Eq.(4.13) and Eq.(4.15). Therefore, an estimate, $\bar{H}$, of the population h value for $n < n_m$ in the case of WM process can be obtained from the simulation work carried out by O'Connell (1974) which gives table (4.1). In this table $\bar{H}$ has been derived as the mean of ten values of H , each defined as the slope of a pox diagram for $n_m \leq 1000$. In the design of water resources, the economic time horizon is usually 50 - 100 years, in which case it should be sufficient to use approximations to dfGn with memories of the order of a few hundred years for generating synthetic sequences. For this purpose, like the ARIMA(1,0,1) process, the WM process will also be sufficient. A good agreement with Hurst's law is achieved with values of ρ_M in the range 0.80 - 0.95 as can be seen from the pox diagrams. For any particular value of ρ_M , the value of n_m decreases as α decreases, hence the approximation to dfGn becomes poorer.

4.4. Small Sample Properties of Estimates of σ^2 , σ , ρ and h

For the purpose of generating synthetic sequences by using an approximate model to simulate long-run properties of a historic sequence of length n , the model parameters, σ^2 (or σ), ρ and h have all to be estimated from this given finite historic record. However, the corresponding estimates S^2 (or S), $\hat{\rho}$ and K or H are known to be biased and consequently any historic record estimates of σ^2 , σ , ρ and h require a bias correction based on the assumption of a particular generating mechanism. Some of these bias correction procedures have been already examined in chapter 3.

ρ_M	.99	.98	.97	.95	.90	.85	.80
ρ	.111	.051	.026	.000	-.038	-.076	-.094
α	.112	.052	.027	.000	-.042	-.089	-.117
$\bar{H}$	.737	.676	.634	.576	.490	.439	.403
ρ	.350	.205	.139	.073	.000	-.042	-.076
α	.352	.210	.144	.077	.000	-.049	-.095
$\bar{H}$	.839	.781	.737	.673	.576	.515	.472
ρ	.562	.384	.287	.179	.061	.000	-.044
α	.570	.391	.295	.189	.067	.000	-.055
$\bar{H}$	.892	.840	.800	.735	.639	.576	.529
ρ	.706	.540	.433	.300	.140	.057	.000
α	.711	.550	.446	.315	.155	.067	.000
$\bar{H}$	.920	.875	.839	.782	.687	.623	.576
ρ	.797	.659	.558	.418	.229	.126	.055
α	.806	.670	.572	.438	.253	.149	.069
$\bar{H}$	.937	.897	.864	.812	.723	.661	.614
ρ	.857	.745	.657	.523	.322	.203	.119
α	.869	.760	.678	.550	.358	.238	.149
$\bar{H}$	.947	.911	.881	.833	.750	.690	.645
ρ	.895	.807	.731	.612	.411	.282	.188
α	.905	.822	.756	.645	.468	.332	.235
$\bar{H}$	.953	.920	.890	.848	.770	.714	.670
ρ	.921	.851	.789	.684	.493	.360	.260
α	.936	.870	.810	.720	.549	.424	.325
$\bar{H}$	.956	.926	.900	.858	.785	.732	.691

Table (4.1) - Values of $\bar{H}$ and ρ for a selection of values of ρ_M and α . Each value of $\bar{H}$ is the mean of ten values of H , each derived as the slope of a least squares line fitted to the mean values of R_n/S for values of $n < 1000$. The total sample size in all cases was 9000.

The variance σ^2 was estimated according to Eq.(3.7) and the expectation of this estimate, S^2 , for the WM process is obtained as

$$E(S^2) = \sigma^2 \left\{ 1 - \frac{2\alpha\rho_M}{n(n-1)} \left[\frac{n(1-\rho_M) - (1-\rho_M^n)}{(1-\rho_M^2)} \right] \right\} \quad (4.19)$$

which reduces to the lag-one Markov case when $\alpha = 1$ which is given in Eq.(3.22). When $\rho_M > 0$ then the bias correction factor

$$f(n; \rho_M, \alpha) = 1 - \frac{2\alpha\rho_M}{n(n-1)} \left[\frac{n(1-\rho_M) - (1-\rho_M^n)}{(1-\rho_M^2)} \right] \quad (4.20)$$

is positive and the magnitude of bias becomes

$$B(n; \rho_M, \alpha) = \frac{2\alpha\rho_M}{n(n-1)} \left[\frac{n(1-\rho_M) - (1-\rho_M^n)}{(1-\rho_M^2)} \right] \quad (4.21)$$

which confirms the fact that the bias effect decreases as n increases. The unbiased estimate of the population variance for a given length n can be achieved by substituting Eq.(4.20) into Eq.(3.25) which yields

$$S^{*2} = \frac{S^2}{f(n; \rho_M, \alpha)} \quad (4.22)$$

By taking the expectation of both sides of Eq.(4.22)

$$E(S^{*2}) = \sigma^2 \quad (4.23)$$

so that S^{*2} will be unbiased provided that the underlying generating process is known to be the WM process.

The population standard deviation, σ , was estimated as the square

root of S^2 given in Eq.(3.7) and its approximate expectation for small samples has been given in general by Eq.(3.45) as

$$E(S) = \sigma \sqrt{F} \left[1 - \frac{1}{4(n^*-1)F^2} \right]$$

where F is equivalent to the bias correction factor for σ^2 given in Eq.(4.20) and n^* can be obtained from the general expressions of Eq.(3.42) and Eq.(3.43) for the WM process as,

$$n^* = n \left\{ \frac{(n-1)(1 - \rho_M^2) + \alpha^2 \rho_M^2 [1 - \rho_M^{2(n-1)}]}{(n-1) + \alpha^2 \rho_M^2 [1 - \rho_M^{2(n-1)}]} \right\} \quad (4.24)$$

or approximately

$$n^* = n \frac{(1 - \rho_M^2)}{(1 - \alpha^2 \rho_M^2 n)} \quad (4.25)$$

The substitution of Eq.(4.24) or Eq.(4.25) into Eq.(3.45) yields the analytical form of the expected standard deviation for the WM process. The bias correction factor for σ becomes

$$g(n; \rho_M, \alpha) = \sqrt{F} \left[1 - \frac{1}{4(n^*-1)F^2} \right] \quad (4.26)$$

or the amount of bias associated with the small sample estimate of σ can be given as

$$B(n; \rho_M, \alpha) = \sigma(1 - \sqrt{F}) \left[1 - \frac{1}{4(n^*-1)F^2} \right] \quad (4.27)$$

The unbiased estimate of the standard deviation is obtained as

$$S^* = \frac{S}{g(n; \rho_M, \alpha)} \quad (4.28)$$

which by taking the expectation of both sides becomes

$$E(S^*) = \sigma \quad (4.29)$$

and hence confirming the unbiased nature of S^* .

In addition, the small sample expectation of the lag k serial correlation coefficient, when the open-series definition given in Eq.(3.57) adopted becomes

$$E(\hat{\rho}_k) = \rho_k - \frac{\alpha}{(n-k)} \left\{ \frac{(1+\rho_M)^2 - \rho_M^k [1+2\alpha\rho_M - \rho_M^2(1-2\alpha)]}{(1-\rho_M^2)} + 2k\alpha\rho_M^k \right\} \quad (4.30)$$

which for ($\alpha = 1$; $\rho = \rho_M$) reduces to the result for lag-one Markov process case given in Eq.(3.65). For $k = 1$ Eq.(4.30) reduces to

$$E(\hat{\rho}) = \rho - \frac{\alpha}{(n-1)} \left\{ \frac{(1+\rho_M)^2 - \rho_M [1+2\rho - \rho_M^2(1-2\alpha)]}{(1-\rho_M^2)} + 2\rho \right\} \quad (4.31)$$

When ($\alpha = 1$; $\rho = \rho_M$) is substituted into Eq.(4.31) than the lag-one Markov process case given in Eq.(3.66) is obtained.

If the circular definition of the serial correlation as well as the Box and Jenkins definition are considered then the small sample expectation of $\hat{\rho}_k$ becomes

$$E(\hat{\rho}_k) = \rho_k - \frac{1}{n} \left[k\alpha\rho_M^k - \alpha\rho_M^k - k\alpha\rho_M^{n-k} + 2\alpha^2 k\rho_M^k + \frac{2\alpha\rho_M(1-\alpha\rho_M^k)}{(1-\rho_M)} + 1 \right] \quad (4.32)$$

which, for $k = 1$, reduces to

$$E(\hat{\rho}) = \rho - \frac{1}{n} \left[\frac{2\rho(\alpha+1) - 4\rho^2}{(1-\rho_M)} \right] \quad (4.33)$$

A similar bias correction procedure to that obtained for the ARIMA(1,0,1) process can be prepared by numerical solution of either Eq.(4.31) or Eq.(4.33). Even the tables for the bias correction of $\hat{\rho}$ in the case of the ARIMA(1,0,1) process can directly be applied to the WM process provided that the conversions between (ρ, φ) and (α, ρ_M) are considered.

Both H and K defined in chapter 2 are known to be biased estimates of the population Hurst coefficient, h . So far in the hydrologic literature there exists no analytical expression for small sample bias corrections for estimates of h . Therefore, the only way of assessing biases in estimates of h may be quantified through Monte Carlo simulation only if h itself is known. Some hydrological processes such as approximations to the dfGn do not yield themselves to simulation experiments because of a prohibitive investment in computer time.

However, the bias correction procedure for K and H in samples of size n can be achieved by defining $E(K)_n$ or $E(H)_n$ provided that the underlying generating process is known. Because of the lacking analytical expressions for $E(K)_n$ and $E(H)_n$ they can be defined by Monte Carlo simulation technique. The simplicity of the WM process makes it feasible to employ Monte Carlo techniques on the computer. The estimate, H or K, of the Hurst's coefficient, h , can be preserved in the synthetic sequences by matching the observed value of h from a historic sequence of size n with $E(K)_n$ or $E(H)_n$ hence reproducing low frequency properties of the observed sequence. In addition, to this the WM process also allows the high frequency properties of the observed sequence to be modelled by preserving $\hat{\rho}$.

In defining the small sample properties of estimates of h for a process

a choice must be made between H or K. In general, as mentioned in Chapter 2, H exhibits smaller bias but larger variance than K. Although the deficiencies that K suffers have been pointed out by Mandelbrot and Wallis (1969e) it is quicker to compute than H. On the other hand the choice of a set of sub-samples to be used in the box diagram from which the estimate H is found is rather subjective and there is no definite universal rule which serves this purpose. As a result K was adopted in the simulation experiments to define $E(K)_n$ for the WM process which was the case for the ARIMA(1,0,1) process (O'Connell, 1974).

However, the simulation experiments for defining $E(K)_n$ have not been carried out for the WM process, but because of the equivalence of this process to the ARIMA(1,0,1) process, the computer experiments have been converted to WM process case through using Eq.(4.13) and Eq.(4.15). In this way table(4.2)-(4.7) have been prepared so that α and ρ_M can be selected to preserve a specified $E(K)_n$ value. It is evident from these tables that for sample sizes 25, 50 and 100 the range of K values which can be modelled for all the selected values of α and ρ_M is approximately 0.65 - 0.87 which covers most of the observed values of K and ρ that are encountered in streamflow sequence. The use of these tables will be illustrated in a later section.

4.5. The Non-Gaussian Case

It frequently occurs that streamflow values, in histogram form, do not conform with the normal distribution. When such a situation exists, the coefficient of skewness, γ used in WM process, which is a measure of the deviation of a distribution function from the normal one must also be preserved in synthetic sequences. This is tantamount to saying that the synthetic flow sequence will resemble the historic one with respect to $\hat{\mu}$, $\hat{\sigma}$, $\hat{\rho}$, $\hat{h}$ and in addition to these $\hat{\gamma}$ which is the estimate of the skewness coefficient.

$\rho_M = 0.96$							
		$E(K)_n$			$E(\hat{\rho})_n$		
α	$\rho \backslash n$	25	50	100	25	50	100
.060	.058	.654	.661	.657	-.005	.005	.025
.152	.146	.683	.679	.698	-.005	.038	.090
.260	.250	.681	.704	.724	.000	.083	.132
.374	.357	.693	.737	.767	.069	.160	.227
.475	.457	.718	.752	.781	.099	.196	.264
.568	.545	.735	.776	.811	.144	.259	.366
.646	.620	.756	.796	.827	.219	.310	.444
.710	.682	.763	.818	.843	.243	.382	.504
.765	.734	.791	.823	.855	.300	.424	.567
.808	.776	.803	.825	.861	.353	.475	.621
.850	.818	.801	.852	.866	.377	.565	.647

Table (4.2) - Values of $E(K)_n$ and $E(\hat{\rho})_n$ for selected values of ρ_M and α .

$\rho_M = 0.92$							
		$E(K)_n$			$E(\hat{\rho})_n$		
α	$\rho \backslash n$	25	50	100	25	50	100
.053	.049	.657	.664	.654	.001	.012	.028
.134	.114	.687	.680	.686	.005	.046	.079
.205	.188	.686	.705	.709	.072	.093	.123
.292	.268	.699	.735	.745	.082	.160	.208
.379	.343	.725	.751	.756	.116	.199	.240
.462	.425	.740	.782	.783	.162	.269	.332
.540	.496	.772	.783	.800	.218	.309	.390
.610	.560	.773	.796	.803	.273	.364	.437
.630	.616	.774	.820	.825	.285	.432	.517

Table (4.3) - Values of $E(K)_n$ and $E(\hat{\rho})_n$ for selected values of ρ_M and α .

$\rho_M = 0.88$							
		$E(K)_n$			$E(\hat{\rho})_n$		
α	$\rho \backslash n$	25	50	100	25	50	100
.052	.046	.659	.665	.651	.005	.041	.026
.130	.102	.689	.678	.676	.012	.050	.078
.188	.165	.688	.703	.696	.080	.097	.117
.265	.233	.703	.728	.728	.091	.158	.196
.343	.302	.729	.745	.737	.126	.198	.225
.420	.370	.742	.772	.764	.176	.262	.309
.495	.435	.771	.772	.778	.223	.302	.363
.562	.495	.771	.788	.782	.278	.356	.407
.625	.550	.774	.808	.802	.291	.416	.481

Table (4.4) - Values of $E(K)_n$ and $E(\hat{\rho})_n$ for selected values of ρ_M and α .

$\rho_M = 0.84$							
		$E(K)_n$			$E(\hat{\rho})_n$		
α	$\rho \backslash n$	25	50	100	25	50	100
.052	.044	.660	.663	.647	.007	.016	.029
.114	.096	.690	.676	.668	.016	.056	.077
.183	.154	.688	.699	.687	.036	.099	.113
.255	.214	.705	.720	.714	.038	.156	.189
.330	.277	.725	.746	.733	.142	.203	.237
.388	.325	.753	.749	.747	.180	.245	.290
.487	.409	.753	.766	.753	.240	.298	.330
.550	.454	.761	.784	.777	.254	.355	.406

Table (4.5) - Values of $E(K)_n$ and $E(\hat{\rho})_n$ for selected values of ρ_M and α .

$\rho_M = 0.80$							
		$E(K)_n$			$E(\hat{\rho})_n$		
α	$\rho \backslash n$	25	50	100	25	50	100
.069	.055	.666	.657	.649	-.006	.022	.041
.139	.119	.685	.694	.678	.056	.079	.098
.235	.188	.708	.698	.708	.114	.103	.166
.325	.260	.736	.727	.719	.139	.192	.227
.413	.331	.737	.750	.741	.175	.247	.289
.500	.400	.756	.764	.746	.263	.315	.343

Table (4.6) - Values of $E(K)_n$ and $E(\hat{\rho})_n$ for selected values of ρ_M and α .

$\rho_M = 0.75$							
		$E(K)_n$			$E(\hat{\rho})$		
α	$\rho \backslash n$	25	50	100	25	50	100
.072	.054	.665	.655	.645	-.004	.023	.041
.153	.115	.684	.689	.671	.059	.079	.095
.240	.179	.707	.691	.698	.117	.102	.161
.329	.246	.733	.719	.707	.141	.189	.219
.419	.313	.739	.740	.728	.177	.240	.277
.500	.377	.751	.754	.732	.264	.307	.329

Table (4.7) - Values of $E(K)_n$ and $E(\hat{\rho})_n$ for selected values of ρ_M and α .

The modification needed in the model structure to preserve $\hat{\gamma}$, as well, have already been presented for the lag-one Markov process, the type 1 approximation to the dfGn and the ARIMA(1,0,1) process by Matalas (1967), Matalas and Wallis (1971) and O'Connell (1974) respectively. It is the third order moment that makes a distribution function skewed if that moment's value is other than zero. Therefore, by considering Eq(4.2) the following result will be valid

$$E(Z_t) = 0 \quad (4.34)$$

$$E(Z^2) = \sigma^2 = \sigma_M^2 + \beta^2 \quad (4.35)$$

and

$$E(Z_t^3) = E(X_t^3) + 3\beta E(X_t^2 \eta_t) + 3\beta^2 (X_t \eta_t^2) + \beta^3 E(\eta_t^3) \quad (4.36)$$

If the normality of the lag-one Markov component, X_t , is assumed then recalling the property of independent superimposition Eq.(4.36) reduces to,

$$E(Z_t^3) = \beta^3 E(\eta_t^3)$$

Or, more specifically

$$\gamma_Z = (1 - \alpha)^{3/2} \cdot \gamma_\eta \quad (4.37)$$

which means that the skewness, γ_Z , of streamflow sequence, Z_t , can entirely be accounted only by the skewness coefficient, γ_η , of the white-noise component, η_t , of WM process provided that the lag-one Markovian component of the process is normally distributed. Hence, if the process Z_t is required to have a skewness γ_Z then this may be achieved through replacing the term η_t in Eq.(4.2) by an independent r.v. ε_t which may be defined by the Wilson-

Hilferty transformation as follows,

$$\varepsilon_t = \frac{2}{\gamma_\varepsilon} \left[1 - \frac{\gamma_\varepsilon \eta_t}{6} - \frac{\gamma_\varepsilon^2}{36} \right]^3 - \frac{2}{\gamma_\varepsilon} \quad (4.38)$$

where η_t is NIP(0,1) and the r.v. ε_t will have zero mean, unit variance and skewness γ_ε . MacMahon and Miller (1971) have observed that for values of γ_ε greater than about 0.3 Eq.(4.38) is inadequate. A modified transformation has been proposed by Kirby(1972) which overcomes the problem. The flows generated by using the above procedure will have an approximate gamma distribution

After having abstracted the estimates of μ , σ , γ , ρ and h from a historic sequence of length n two types of resemblance, either type A and type B, as mentioned in chapter 3 can be maintained between synthetic and historic sequences as quoted by O'Connell (1974) for the ARIMA(1,0,1) process.

a) Type A Resemblance

In this case the estimates $\hat{\mu}$, $\hat{\sigma}$, $\hat{\rho}$, $\hat{\gamma}$ and h derived from an available historic sequence of length n are assumed to play the role of the unknown population counterparts; then synthetic sequences of infinite length are generated and the estimates of the same parameters are abstracted this time from the synthetic sequences of infinite length. If the two sets of estimates are equal to each other then resemblance of type A is said to be maintained. Unfortunately, all of the approximations to dfGn that fall within the Brownian domain of attraction give a population value of $h = 0.5$, therefore resemblance of type A cannot be maintained in terms of h for the WM process. This fact is true of any finite memory approximation to dfGn. However, resemblance in terms of an estimate of h can be maintained within finite synthetic sequences. The following 6 - step procedure is to be followed

to maintain a type A resemblance

1) find the estimates $\hat{\mu}$, $\hat{\sigma}$, $\hat{\rho}$, $\hat{\gamma}$ and K from a given historic sequence of length n.

2) By considering n and assuming that

$$E(K)_n \approx K \quad (4.39)$$

and

$$\rho \approx \hat{\rho}$$

find the corresponding model parameters α and ρ_M from Tables (4.2) - (4.7). Interpolation may be necessary here.

3) Find the estimate of λ from Eq.(4.17) as

$$\hat{\lambda} = \frac{1}{\alpha} - 1 \quad (4.40)$$

and subsequently the estimates of α_M and β can be calculated from the simultaneous solution of Eq.(4.5) and Eq.(4.16) which gives

$$\hat{\alpha}_M = \frac{\hat{\sigma}_Z}{\sqrt{1+\hat{\gamma}}} \quad (4.41)$$

and

$$\hat{\beta} = \frac{\hat{\gamma}\hat{\sigma}}{\sqrt{1+\hat{\gamma}}} \quad (4.42)$$

4) The estimated skewness $\hat{\gamma}_Z$ is then used to define an estimate of γ_ϵ using Eq.(4.37) which yields

$$\hat{\gamma}_\epsilon = \frac{\hat{\gamma}_Z}{(1-\alpha)^{3/2}} \quad (4.43)$$

The term ϵ_t can then be generated using Eq.(4.38).

5) The quantities $\hat{\mu}_M = \hat{\mu}$, $\hat{\sigma}_M$, $\hat{\rho}_M$ and $\hat{\beta}$ are then incorporated into Eq.(4.1) and Eq.(4.2),

6) whence a sequence of lag-one Markov component values, X_t , of a desired length is generated by using Eq.(4.1). This sequence is then independently superimposed on to the white-noise component ϵ_t , generated in step 4 , according to the following relation

$$Z_t = X_t + \beta \epsilon_t \quad (4.44)$$

b) Type B resemblance

In this case resemblance is maintained in terms of estimates of parameters obtained from both historic and synthetic sequences of the same length n . The small sample properties of the coefficient of skewness are unknown. Therefore it will not be included in maintaining this type of resemblance. The necessary steps in the procedure are as follows.

1) Find the estimates $\hat{\mu}$, $\hat{\sigma}$, $\hat{\rho}$ and K from a historic sequence of length n .

2) From tables (4.2) - (4.7) identify values of α and ρ_M such that

$$E(K)_n \approx K$$

and

$$E(\hat{\rho})_n \approx \hat{\rho}$$

Again some interpolation will be involved.

3) Eq.(4.19) and Eq.(3.45) can be re-written as

$$E(S^2)_n = \sigma^2 f(n; \rho, \alpha) \quad (4.45)$$

and

$$E(S)_n = \sigma g(n; \rho_M, \alpha) \quad (4.46)$$

where the analytical forms of $f(n, \rho_M, \alpha)$ and $g(n, \rho_M, \alpha)$ are given by Eq.(4.20) and Eq.(4.26) respectively. After having obtained α and ρ_M under (2) an unbiased estimate of σ_z^2 or σ_z may be defined as

$$\sigma_z^{*2} = \frac{\hat{\sigma}^2}{f(n; \rho_M, \alpha)} \quad (4.47)$$

and

$$\sigma_z^* = \frac{\hat{\sigma}}{g(n; \rho_M, \alpha)} \quad (4.48)$$

respectively.

(4) The estimates of $\hat{\lambda}$, σ_M and $\hat{\beta}$ are found by making use of Eq.(4.40), Eq.(4.41) and Eq.(4.42) respectively.

(5) The quantities $\hat{\mu}_M = \mu$, $\hat{\sigma}_M$, $\hat{\rho}$ and $\hat{\beta}$ are then incorporated into Eq.(4.3) whence synthetic flows may be generated. On some occasions the estimates ρ and K may not be compatible with the values of $E(\hat{\rho})_n$ and $E(K)_n$ in tables (4.2) - (4.7) thus preventing the required resemblance from being maintained. In such a case either one of the parameters can be preserved in synthetic sequences and thus an approximate resemblance can be maintained.

4.6. The Log-Normal WM Process

If historic flows conform to a log-normal distribution then these flows can be transformed into a normal space through using

$$Y_t = \log(X_t - a) \quad (4.49)$$

where α is a lower bound of r.v. X_t . The new r.v. Y_t can be assumed to be a WM process. The relationships between the parameters of Y_t and X_t are given by Aitchison and Brown (1957) as,

$$\mu_X = \alpha + \exp\left(\frac{\sigma_Y^2}{2} + \mu_Y\right) \quad (4.50)$$

$$\sigma_X^2 = \exp\left[2(\sigma_Y^2 + \mu_Y)\right] - \exp(\sigma_Y^2 + 2\mu_Y) \quad (4.51)$$

$$Y_X = \frac{\exp(3\sigma_Y^2) - 3\exp(\sigma_Y^2) + 2}{\left[\exp(\sigma_Y^2) - 1\right]^{3/2}} \quad (4.52)$$

The relationship between the first order autocorrelation coefficient, ρ_X and ρ_Y of variates X_t and Y_t respectively has been given by Matalas (1967) for the lag-one Markov process as

$$\rho_X = \frac{\exp(\rho_Y \sigma_Y^2) - 1}{\exp(\sigma_Y^2) - 1} \quad (4.53)$$

This result has been shown to be process free by O'Connell (1974). Once estimates μ_X , σ_X^2 and Y_X are obtained from a given historic sequence then the simultaneous solution of Eq.(4.50), Eq.(4.51) and Eq.(4.52) gives the estimates of μ_Y , σ_Y^2 and α . If the r.v. Y_t is assumed to be generated by a WM process with parameters within the parameter space covered by tables (4.2)-(4.7) then long-term persistence will also be preserved in X-space. The model parameters α and ρ_M can be obtained by using either the type A or type B procedure for maintaining resemblance between historic and synthetic sequences.

By replacing the estimate of ρ_X into Eq.(4.53) an estimate of ρ_Y can be obtained. Although resemblance of type B can be maintained in Y-space, there is no guarantee that such a resemblance will be maintained in X-space because of the non-linear transformations relating X_t and Y_t .

4.7. The Multivariate Case

Up to this point only the information that is yielded by a record at a single site has been taken into consideration. However, a model preserving the spatial correlations between sites in addition to the temporal correlations at each site is frequently required. When the events at each site within a catchment area are entirely independent from each other then the single site model could be applied independently at each site. On the other hand, if a catchment area is considered then the process that takes place on it is generally sampled at various distinct sites and time instances. Therefore, one usually expects that all the records at these distinct sites are interdependent which is tantamount to saying that the records are spatially correlated apart from the correlation in the structure of each record which is referred to as the temporal correlation. The spatial dependence can be specified in general by the cross-correlation between pairs of records at distinct sites; therefore in particular the lag-zero cross correlation corresponds to the linear dependence between the records, when all the observations in the two records are considered concurrently. If all the observations of one record are shifted by one time unit with respect to the observations of the other record then the lag-one cross-correlation results. In the same way higher order lag cross-correlation can be conceived.

The formulation of a stochastic model can be devised in such a way that it takes into account both temporal and spatial correlations; in this form it is frequently referred to as a multivariate generating process.

Consequently, the purpose of a multivariate version of a stochastic model is to ensure the statistical resemblance between the historic and synthetic records at various sites is maintained, on the basis of the means, variances, serial correlations of each record and in addition to these the cross-correlations among the records.

A detailed review of the multivariate models developed and used in synthetic hydrology so far, and then the necessary algebraic calculations for the multivariate ARIMA(1,0,1) process are given by O'Connell (1974). The multivariate Markov process was first devised by Matalas (1967) and Matalas and Wallis(1971) specify a multivariate dfGn model.

Although one tends to expect the expressions for the multivariate WM case to be rather similar to those for the multivariate Markov process , they are indeed very different from each other as will be appreciated from the following equations. They also differ from the equations for the multivariate ARIMA(1,0,1) model. This latter aspect is of particular interest as some difficulties were encountered in developing the multivariate ARIMA(1,0,1) model (O'Connell, 1974).

The general form of the multivariate case of the WM process can be written as

$$\underline{Z}_t = \underline{A}\underline{X}_{t-1} + \underline{B}\underline{\epsilon}_t + \underline{C}\underline{\eta}_t \quad (4.54)$$

where $\underline{Z}_t$ is an $(mx1)$ column matrix of synthetic flows, the i -th element of which is equal to $z_t^{(i)} - \mu_z^{(i)}$, $\underline{A}$ is an (mxm) square matrix of coefficients, $\underline{X}_{t-1}$ is an $(mx1)$ column matrix of the lag-one Markov component, $\underline{\epsilon}_t$ and $\underline{\eta}_t$ are both $(mx1)$ column matrices of independent r.v's which are also independent from each other, $\underline{B}$ and $\underline{C}$ are (mxm) matrices of coefficients and finally m denotes the number of sites considered. It is the main objective of this section to evaluate the elements of the coefficient matrices $\underline{A}$, $\underline{B}$ and $\underline{C}$ in

terms of the means, variances, serial correlation coefficients of individual records and low order cross correlation coefficients among them, such that the required resemblance will be maintained between historic and synthetic sequences in terms of A , B and C. Although it is mathematically feasible to construct a multivariate model that preserves all the lag correlations it is practically impossible to estimate higher order lag correlations reliably from a given set of historic data of finite length. It is because of this difficulty that hydrologists are interested in the development of multivariate models which necessitates the use of low order moments only.

In order to find the elements of the matrices A, B and C the following algebra is necessary. First of all both sides of Eq.(4.54) are postmultiplied by the transpose, $\underline{Z}_t^T$, of the matrix Z and then by taking expectation

$$E(\underline{Z}_t \underline{Z}_t^T) = \underline{A} E(\underline{X}_{t-1} \underline{Z}_t^T) + \underline{B} E(\underline{\epsilon}_t \underline{Z}_t^T) + \underline{C} E(\underline{\eta}_t \underline{Z}_t^T) \quad (4.55)$$

where from Eq.(4.45), $\underline{Z}_t^T$ is obtained as

$$\underline{Z}_t^T = \underline{X}_{t-1}^T \underline{A}^T + \underline{\epsilon}_t^T \underline{B}^T + \underline{\eta}_t^T \underline{C}^T \quad (4.56)$$

By substituting Eq.(4.56) into the r.h.s. of the Eq.(4.55) and after making the necessary simplifications the following equation emerges

$$E(\underline{Z}_t \underline{Z}_t^T) = \underline{A} E(\underline{X}_{t-1} \underline{X}_{t-1}^T) \underline{A}^T + \underline{B} \underline{B}^T + \underline{C} \underline{C}^T \quad (4.57)$$

or, by denoting

$$\underline{M}_O^Z = E(\underline{Z}_t \underline{Z}_t^T) \quad (4.58)$$

$$\underline{M}_O^X = E(\underline{X}_t \underline{X}_t^T) \quad (4.59)$$

where $\underline{M}_0^Z$ and $\underline{M}_0^X$ are the lag-zero covariance matrices of the random variates $\underline{Z}_t$ and $\underline{X}_t$ respectively. In the case of m sites both of the matrices are (mxm) square matrices whose explicit forms can be given as follows

$$\underline{M}_0^Z = \begin{bmatrix} E(Z_t^{(1)} Z_t^{(1)}) & E(Z_t^{(1)} Z_t^{(2)}) & \dots & E(Z_t^{(1)} Z_t^{(m)}) \\ E(Z_t^{(2)} Z_t^{(1)}) & E(Z_t^{(2)} Z_t^{(2)}) & \dots & E(Z_t^{(2)} Z_t^{(m)}) \\ \vdots & \vdots & \ddots & \vdots \\ E(Z_t^{(m)} Z_t^{(1)}) & E(Z_t^{(m)} Z_t^{(2)}) & \dots & E(Z_t^{(m)} Z_t^{(m)}) \end{bmatrix} \quad (4.60)$$

where the diagonal elements are the lag-zero covariances of individual record, and the off-diagonal elements are the lag-zero cross covariances of any two records. In order to have a more compact form of Eq.(4.60) the variables at each site are assumed to be standardized as

$$\underline{Z}_t^{(i)} = \frac{Z_t^{(i)} - \mu^{(i)}}{\sigma^{(i)}} \quad (4.61)$$

where $\mu^{(i)}$ and $\sigma^{(i)}$ are the mean and standard deviation of sequence at site i and $Z_t^{(i)}$ is the non-standardized form of the variable. By the virtue of this standardizing procedure Eq.(4.60) becomes

$$\underline{M}_0^Z = \begin{bmatrix} 1 & \rho_{12}^Z & \rho_{13}^Z & \dots & \rho_{1m}^Z \\ \rho_{21}^Z & 1 & \rho_{23}^Z & \dots & \rho_{2m}^Z \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{m1}^Z & \rho_{m2}^Z & \rho_{m3}^Z & \dots & 1 \end{bmatrix} \quad (4.62)$$

where the diagonal elements are all equal to unity because of the fact that they are the variances of the standardized variables at individual sites. The off-diagonal elements are the lag-zero cross correlations between the records. Once the set of historic records have been provided each of the elements of $\underline{M}_0^Z$ can be estimated and consequently the matrix is a known quantity. As can be seen from Eq.(4.62), $\underline{M}_0^Z$ is a symmetric matrix. In Eq.(4.59) the matrix $\underline{M}_0^X$ has got the same meaning as $\underline{M}_0^Z$ but pertains to the Markov component and can be explicitly written as

$$\underline{M}_0^X = \begin{bmatrix} 1 & \rho_{12}^X & \rho_{13}^X & \dots & \rho_{1m}^X \\ \rho_{21}^X & 1 & \rho_{23}^X & \dots & \rho_{2m}^X \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{m1}^X & \rho_{m2}^X & \rho_{m3}^X & \dots & 1 \end{bmatrix} \quad (4.63)$$

where the off-diagonal elements correspond to the lag-zero correlation coefficients of the lag-one Markov component of WM process. These elements cannot be estimated from a given set of data, therefore $\underline{M}_0^X$ is an unknown matrix. Finally, Eq.(4.57) can be re-written as,

$$\underline{M}_0^Z = \underline{A} \underline{M}_0^X \underline{A}^T + \underline{B} \underline{B}^T + \underline{C} \underline{C}^T \quad (4.64)$$

where $\underline{A}$, $\underline{B}$, $\underline{C}$ and $\underline{M}_0^X$ are unknown terms, hence in order to find a unique solution for them another further three equations are required.

A second equation can be obtained by post-multiplying Eq.(4.54) by

the transpose, $\underline{Z}_{t+1}^T$, of $\underline{Z}_{t+1}$ and taking expectations of both sides which leads to

$$E(\underline{Z}_t \underline{Z}_{t+1}^T) = \underline{A} E(\underline{X}_{t-1} \underline{X}_t^T) \underline{A}^T + \underline{B} E(\underline{\varepsilon}_t \underline{X}_t^T) \underline{A}^T \quad (4.65)$$

By incorporating the following equations

$$\underline{M}_1^Z = E(\underline{Z}_t \underline{Z}_{t+1}^T) \quad (4.66)$$

$$\underline{M}_1^X = E(\underline{X}_{t-1} \underline{X}_t^T) \quad (4.67)$$

and

$$\underline{B}^T = E(\underline{\varepsilon}_t \underline{X}_t^T) \quad (4.68)$$

Eq.(4.68) becomes,

$$\underline{M}_1^Z = \underline{A} \underline{M}_1^X \underline{A}^T + \underline{B} \underline{B}^T \underline{A}^T \quad (4.69)$$

where $\underline{M}_1^Z$, has the lag-one serial correlation coefficient of the records as its main-diagonal elements and the lag-one cross correlations between the records as the off-diagonal elements. $\underline{M}_1^X$ is the lag-one cross correlation matrix of the Markov component.

Each one of the elements of matrix $\underline{M}_1^Z$ can be estimated from a set of given historic records and therefore it is a known quantity. The explicit form of the matrix $\underline{M}_1^Z$ can be given as

$$\underline{M}_1^Z = \begin{bmatrix} Z_{\rho(1)11} & Z_{\rho(1)12} & Z_{\rho(1)13} & \dots & Z_{\rho(1)1m} \\ Z_{\rho(1)21} & Z_{\rho(1)22} & Z_{\rho(1)23} & \dots & Z_{\rho(1)2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_{\rho(1)m1} & Z_{\rho(1)m2} & Z_{\rho(1)m3} & \dots & Z_{\rho(1)mm} \end{bmatrix} \quad (4.70)$$

where sub-subscripts outside the brackets denote the lag value. In Eq.(4.69) $\underline{M}_1^X$ has just the same structure as the $\underline{M}_1^Z$ matrix, but relates to the Markov component of the model itself; as a result it is an unknown quantity. The major properties of matrices $\underline{M}_1^Z$ and $\underline{M}_1^X$ are that they are non-symmetric because of the fact that in general

$$Z_{ij}^{(1)} \neq Z_{ji}^{(1)} \quad \text{for } i \neq j \quad (4.71)$$

To obtain the further two equations required Eq.(4.54) is postmultiplied by the transposes $\underline{Z}_{t+2}^m$ and $\underline{Z}_{t+3}^T$ of $\underline{Z}_{t+2}$ and $\underline{Z}_{t+3}$ respectively and then the expectations are taken which leads to

$$\underline{M}_2^Z = \underline{A} \underline{M}_2^X \underline{A}^T + \underline{B} \underline{B} \underline{A}^T \quad (4.72)$$

and

$$\underline{M}_3^Z = \underline{A} \underline{M}_3^X \underline{A}^T + \underline{B} \underline{B} \underline{A}^T \quad (4.73)$$

In both of these two last equations the matrices $\underline{M}_2^Z$ and $\underline{M}_3^Z$ have the same

interpretations as the previous $\underline{M}_0^Z$ and $\underline{M}_1^Z$ matrices, the only difference being with respect to their lags.

The four basic equations Eq.(4.64), Eq.(4.72), Eq.(4.69) and Eq.(4.73) are the equations necessary to solve for the four unknowns , $\underline{A}$, $\underline{B}$, $\underline{C}$ and $\underline{M}_0^X$, but among the terms corresponding to the Markov component, of these four equations, there exist the following relationships, in general,

$$\underline{M}_1^X = \underline{A}^1 \underline{M}_0^X \quad (4.74)$$

and in particular

$$\underline{M}_1^X = \underline{A} \underline{M}_0^X \quad (4.75)$$

$$\underline{M}_2^X = \underline{A}^2 \underline{M}_0^X \quad (4.76)$$

$$\underline{M}_3^X = \underline{A}^3 \underline{M}_0^X \quad (4.77)$$

By substituting these last three equations into the basic equations a new system of basic equations is obtained as,

$$\underline{M}_0^Z = \underline{A} \underline{M}_0^X \underline{A}^T + \underline{B} \underline{B}^T + \underline{C} \underline{C}^T \quad (4.78)$$

$$\underline{M}_1^Z = \underline{A} \underline{M}_0^X \underline{A}^T + \underline{B} \underline{B}^T \underline{A}^T \quad (4.79)$$

$$\underline{M}_2^Z = \underline{A} \underline{M}_0^X \underline{A}^T + \underline{B} \underline{B}^T \underline{A}^T \quad (4.80)$$

$$\underline{M}_3^Z = \underline{A} \underline{M}_0^X \underline{A}^T + \underline{B} \underline{B}^T \underline{A}^T \quad (4.81)$$

It is obvious that the last three basic equations do not include the matrix $\underline{C}$ and it is because of this property that the solution can be split up into two parts. In the first part, the three unknowns $\underline{A}$, $\underline{B}$ and $\underline{M}_0^X$ can first be evaluated by using the last three basic equations, whereas in the second part, which follows on completion of the first part, the only remaining unknown $\underline{C}$ can be solved from the first basic equation that is Eq.(4.78). Another property is that the second terms on the r.h.s. of the last three basic equations are the same and can be eliminated by taking the difference of any two equations. Hence, subtraction of Eq.(4.80) from Eq.(4.81) gives

$$\underline{M}_2^Z - \underline{M}_3^Z = \underline{A} \underline{M}_0^X \underline{A}^T - \underline{A} \underline{M}_0^X \underline{A}^T \quad (4.82)$$

and of Eq.(4.80) from Eq.(4.79) gives

$$\underline{M}_1^Z - \underline{M}_2^Z = \underline{A} \underline{M}_0^X \underline{A}^T - \underline{A} \underline{M}_0^X \underline{A}^T \quad (4.83)$$

By premultiplying Eq.(4.83) by the matrix $\underline{A}$ and then subtracting it from Eq.(4.82) a solution for $\underline{A}$ is obtained as

$$\underline{A} = (\underline{M}_2^Z - \underline{M}_3^Z) (\underline{M}_1^Z - \underline{M}_2^Z)^{-1} \quad (4.84)$$

where the r.h.s. is known from the given historic data.

Let a new matrix, $\underline{D}$, be denoted as,

$$\underline{D} = \underline{A} \underline{M}_0^X \underline{A}^T \quad (4.85)$$

which is symmetric and when it is substituted in Eq.(4.83) then the result becomes

$$\underline{M}_1^Z - \underline{M}_2^Z = \underline{A} \underline{D} + \underline{A}^2 \underline{D} = (\underline{A} + \underline{A}^2) \underline{D}$$

or

$$\underline{D} = (\underline{I} + \underline{A})^{-1} \underline{A}^{-1} (\underline{M}_1^Z - \underline{M}_2^Z) \quad (4.86)$$

where $\underline{I}$ is a unit matrix. Hence the calculation of $\underline{D}$ by Eq.(4.86) in terms of known quantities makes it possible to solve for the matrix $\underline{M}_0^X$ from Eq.(4.85) as

$$\underline{M}_0^X = \underline{A}^{-1} \underline{D} (\underline{A}^T)^{-1} \quad (4.87)$$

which is a symmetric matrix. Eq.(4.87) can be re-written more explicitly by substituting the equivalent of $\underline{D}$ which leads to

$$\underline{M}_0^X = \underline{A}^{-1} (\underline{I} - \underline{A}^{-1} \underline{A}^{-1} (\underline{M}_1^Z - \underline{M}_2^Z)) (\underline{A}^T)^{-1} \quad (4.88)$$

Eq.(4.72) gives

$$\underline{B} \underline{B}^T = (\underline{M}_2^Z - \underline{A}^3 \underline{M}_0^X \underline{A}^T) (\underline{A}^T)^{-1}$$

or

$$\underline{B} \underline{B}^T = \underline{M}_2^Z (\underline{A}^T)^{-1} - \underline{A}^3 \underline{M}_0^X \underline{A}^T (\underline{A}^T)^{-1}$$

where $\underline{A}^T (\underline{A}^T)^{-1} = \underline{I}$ and hence

$$\underline{B} \underline{B}^T = \underline{M}_2^Z (\underline{A}^T)^{-1} - \underline{A}^3 \underline{M}_0^X \quad (4.89)$$

$\underline{B}\underline{B}^T$ in Eq.(4.89) is a symmetric matrix confirming that the r.h.s. of the same equation is also symmetric. Hence by defining a symmetric matrix, $\underline{S}_1$, as

$$\underline{B}\underline{B}^T = \underline{S}_1 \quad (4.90)$$

therefore a unique solution of matrix $\underline{B}$ can be obtained only by assuming that $\underline{B}$ is a lower triangular matrix. This gives the same number of non-zero coefficients $m(m + 1)/2$ as required to specify $\underline{S}_1$. The solution is obtained by expanding Eq.(4.90) into $m(m + 1)/2$ equations that define the matrix equation. Subsequently the elements of $\underline{B}$ are found recursively.

Finally, the solution of matrix $\underline{C}$ is obtained from Eq.(4.64) as

$$\underline{C}\underline{C}^T = \underline{M}_0^Z - \underline{A}\underline{M}_0^X \underline{A}^T - \underline{B}\underline{B}^T \quad (4.91)$$

where again $\underline{C}\underline{C}^T$ is a symmetric matrix and its solution is also achieved by lower triangularization.

As a result there appears to be the same difficulty in the solution of the multivariate equations of the WM process as was the case for the ARIMA(1,0,1) process; because of the restriction as to the assumption of lower order triangularization of matrices $\underline{B}$ and $\underline{C}$. The total number of non-zero elements of unknown matrices is

$$\frac{m(m+1)}{2} + \frac{m(m+1)}{2} + \frac{m(m+1)}{2} + m^2 \quad (4.92)$$

whereas the total number of elements of known matrices is given as

$$\frac{m(m+1)}{2} + m^2 + m^2 + m^2 \quad (4.93)$$

As a result there are not enough elements to define the elements of matrices $\underline{A}$, $\underline{B}$, $\underline{C}$ and $\underline{M}_0^X$.

4.7.1. Multivariate Log-Normal WM Process.

In the case where the variates, $X_i(t)$, at each site conforms to a three parameter log-normal distribution then $Z_i(t) = \ln \left[\frac{X_i(t) - a_i}{\sigma_i} \right]$ will follow a multivariate normal distribution, where a_i denotes the lower bound of variable $X_i(t)$ at site i . A multivariate normal WM process, discussed in the previous section can then be applied to $Z_i(t)$ which can be written as

$$\underline{z}(t) = \underline{A}' \underline{x}(t-1) + \underline{B}' \underline{\varepsilon}(t) + \underline{C}' \underline{\eta}(t) \quad (4.94)$$

where $\underline{z}(t)$ and $\underline{x}(t-1)$ are vectors of standardized normal variates with elements

$$z_i(t) = \frac{Z_i(t) - \mu'_i}{\sigma'_i} \quad (4.95)$$

$$x_i(t-1) = \frac{X_i(t-1) - \mu'_i}{\sigma'_i} \quad (4.96)$$

where $x_i(t-1)$ is the lag-one Markov component of the overall WM process. The matrices $\underline{A}'$, $\underline{B}'$ and $\underline{C}'$ must be defined from the correlation matrices $\underline{M}_0^{z'}$, $\underline{M}_1^{z'}$, $\underline{M}_2^{z'}$ and $\underline{M}_3^{z'}$ of the process $Z(t)$. Moreover, by considering the logarithmic transformation the elements of the matrices $\underline{M}_0^{z'}$, $\underline{M}_1^{z'}$, $\underline{M}_2^{z'}$ and $\underline{M}_3^{z'}$, $\rho_{ij}^{z'}(0)$, $\rho_{ij}^{z'}(1)$, $\rho_{ij}^{z'}(2)$ and $\rho_{ij}^{z'}(3)$ respectively, must be defined from the corresponding counterparts of matrices $\underline{M}_0^z$, $\underline{M}_1^z$, $\underline{M}_2^z$ and $\underline{M}_3^z$ through the following equations given by O'Connell (1974)

$$\rho_{ij}^{z'}(0) = \frac{[\exp(\sigma_i' \sigma_j' \rho_{ij}^{z'}(0)) - 1]}{[\exp(\sigma_i'^2) - 1]^{1/2} [\exp(\sigma_j'^2) - 1]^{1/2}} \quad (4.97)$$

$$\rho_{ij}^z(1) = \frac{[\exp(\dot{\sigma}_i \cdot \dot{\sigma}_j \cdot \rho_{ij}^z(1)) - 1]}{[\exp(\dot{\sigma}_i^2) - 1]^{1/2} [\exp(\dot{\sigma}_j^2) - 1]^{1/2}} \quad (4.98)$$

$$\rho_{ij}^z(2) = \frac{[\exp(\dot{\sigma}_i \cdot \dot{\sigma}_j \cdot \rho_{ij}^z(2)) - 1]}{[\exp(\dot{\sigma}_i^2) - 1]^{1/2} [\exp(\dot{\sigma}_j^2) - 1]^{1/2}} \quad (4.99)$$

and

$$\rho_{ij}^z(3) = \frac{[\exp(\dot{\sigma}_i \cdot \dot{\sigma}_j \cdot \rho_{ij}^z(3)) - 1]}{[\exp(\dot{\sigma}_i^2) - 1]^{1/2} [\exp(\dot{\sigma}_j^2) - 1]^{1/2}} \quad (4.100)$$

The estimates of the elements of $\underline{M}_0^z$, $\underline{M}_1^z$, $\underline{M}_2^z$ and $\underline{M}_3^z$ may be inserted into equations (4.97) - (4.100) which in turn yields estimates of elements of $\underline{M}_0^{z'}$, $\underline{M}_1^{z'}$, $\underline{M}_2^{z'}$ and $\underline{M}_3^{z'}$. In addition the estimates of $\dot{\mu}_i$, $\dot{\sigma}_i$, and $\dot{\sigma}_i^2$ can be obtained by employing equations (4.50) - (4.52). After having obtained all of the elements of matrices $\underline{M}_0^{z'}$, $\underline{M}_1^{z'}$, $\underline{M}_2^{z'}$ and $\underline{M}_3^{z'}$ then the matrices $\underline{A}'$, $\underline{B}'$, $\underline{C}'$ and $\underline{M}_0^{x'}$ can be solved for as discussed in the previous section.

4.8. Summary

A model referred to as the white Markov process has been devised to simulate annual streamflow sequences and its long-term behaviour has been examined using computer simulation experiments which revealed that the WM process can be used to approximate the behaviour of long geophysical records. On the other hand, the preservation of the lag-one autocorrelation coefficient shows the fact that the short-term properties of the process are modelled in addition to the long-term properties.

The small sample properties of the model have been analytically derived and they have been used in formulating the process to maintain the required statistical resemblance between historic and synthetic sequences. Finally,

the necessary equations for the application of the WM process to a multisite situation have been derived. However, the solutions obtained may prove slightly restrictive.

Chapter 5

BAYESIAN APPROACH TO AUTOREGRESSIVE PROCESSES

5.1. General

In the design and operation of water resource systems, hydrologists have been trying to estimate the parameters of the appropriate model for their purpose, by using classical statistical methods very frequently. However, recently a new mode of estimation has become popular that is the Bayesian estimation of parameters which allows the hydrologist, apart from the historic evidence, to incorporate his own experience about the hydrological event concerned. In addition the Bayesian approach has an advantage for small samples whereas for large samples both the classical and Bayesian estimates yield the same result. It is hoped that the Bayesian estimates will be more realistic than their classical counterparts because of the fact that the latter has only one information source, that is the historic performance of the hydrological event, whereas, the former has in addition to this historic data, the experience enjoyed by the hydrologist as well.

A statistical inference about any one of the parameters of a hydrological model can be achieved in two ways; either by the application of the sampling theory or by use of Bayes theorem. The methods employed in Chapter 3 in connection with the small sample properties of various parameters fall within the sampling theory framework. However, the main purpose of this chapter will be the derivation of the necessary expressions which make it possible to apply the Bayes theorem in general and in particular to the ARIMA(1,0,1) process. The general formulae derived in this chapter are all readily reducible to the Markov process case, for which some results have been given by Lenton et. al. (1973).

An extensive exploitation of the Bayesian approach to parameter estimation appears in the econometrics field, where a detailed study has been

given by Zellner (1971). In his treatise a very comprehensive consideration of various aspects of Bayesian inference has been given, such as univariate and multivariate regression analysis and particularly first order Gaussian autoregressive processes.

In hydrology, an account of Bayesian parameter estimation and inference have been given by Lenton et. al. (1973) for the lag-one autocorrelation coefficient, ρ , only on the basis of the Markov process and world-wide prior information about ρ . Another work that appeared in the hydrological literature is that of Conover (1971) who also applied the Bayesian approach to estimating ρ , on the basis of hypothetical prior information about ρ stating that ρ is normally distributed which is not a realistic assumption in practice.

However, a topic related to the Bayesian approach, that is Bayesian decision theory has been more extensively applied to hydrological design systems. To this end Tschannerl (1970) studied various efficient uses of data sources within a decision framework, when he applied Bayesian decision making to the population mean value of Markovian annual streamflows. By assuming the population mean as a r.v. he assigned a normal pdf to it and then digitized this pdf in order to find the probabilities associated with a set of possible mean values and subsequently through the use of an appropriate loss function, a mean value minimizing this loss function is decided on.

Decision theory could also be used for the choice of the underlying distribution function as was advocated by Fiering and Jackson (1971) who considered two distribution functions only and generated synthetic streamflows based on each distribution so that an optimal design value associated with each distribution was obtained through an economic objective function. Moreover, the consequences of accepting the design corresponding to the

first distribution, assuming that the second were the true distribution is evaluated in terms of economic loss. Conversely the economic loss involved in assuming the second distribution when the first was, in fact the true one must be evaluated. The optimal design will then consequently suggest the appropriate distribution for the flows.

In section (5.2) a general review of Bayes theorem is given. The marginal prior pdf of φ which is one of the model parameters of the ARIMA(1,0,1) process is derived in section (5.3). The likelihood function which is representative of information contained in the historic data is dealt with in section (5.4) and a general expression is derived for it on the basis of an assumption that the underlying generating process is ARIMA(1,0,1). The combination of the prior pdf and the likelihood function so as to give an improved information level named the posterior pdf is achieved in section (5.5) through Bayes theorem. Moreover, in the subsections of section (5.5) the posterior pdf of the parameter φ is obtained for various situations, where first μ and σ are both considered to be random variables, then one of them is allowed to be a r.v. whereas the other one is taken as a constant and finally both μ and σ are thought of as being constants. A very brief review of Bayesian estimates is given in section (5.6).

5.2. Bayes Theorem

Bayes theorem which is also referred to in the probability literature, as the principle of inverse probability, is a means whereby historic information i.e. recorded data and prior information about the event before the data is recorded, can be combined together to give more detailed information called the posterior information.

After the identification of the model that gives rise to the recorded historic observations, the data are then regarded as being generated by

this model whose parameters are assumed to be unknown. In the Bayesian approach the unknown population parameters are all assumed to be r.v's, and as a result of this assumption the parameters will have individual and in addition joint pdf's. Therefore, the above mentioned prior and posterior informations will be expressed formally in terms of pdf's. The prior pdf is denoted by $P(\omega)$, where ω is a set of values that the parameter of the model can assume and the historic information that is the observed data is represented by $l(D|\omega)$ which means that given a set of ω values, in addition to the model itself, then, there exists only one pdf reflecting the random nature of the historic sequence, i.e. it is a kind of conditional pdf. As a result the joint pdf of the data with the parameters may be expressed in two different ways as follows

$$P(D, \omega) = P(\omega) \cdot l(D|\omega) \quad (5.1)$$

and

$$P(D, \omega) = P(D) \cdot l(\omega|D) \quad (5.2)$$

These two equations yield the same result and therefore by setting the r.h.s's equal to each other

$$P(D) \cdot l(\omega|D) = P(\omega) \cdot l(D|\omega)$$

and by dividing both sides by $P(D)$

$$l(\omega|D) = \frac{P(\omega) \cdot l(D|\omega)}{P(D)} \quad (5.3)$$

The denominator of this last equation is a fixed value for a given set of

historic observations and moreover, its main role is to give a standardized pdf in the sense that the area under the function $[(\omega | D)]$ is equal to unity. Hence, by denoting $l(\omega | D)$ by $P(\omega | D)$ Bayes theorem can be re-written as

$$P(\omega | D) \propto P(\omega) l(D | \omega) \quad (5.4)$$

or, in an equality form as

$$P(\omega | D) = N \cdot P(\omega) \cdot l(D | \omega) \quad (5.5)$$

where N is a standardizing factor and the remaining terms in Eq.(5.5) have the following meanings

$P(\omega | D)$: Posterior pdf of the parameters

$P(\omega)$: Prior pdf of the parameters

$l(D | \omega)$: Likelihood function

Eq.(5.4) is the general form of the posterior pdf. When the lag-one Markov process is adopted to represent the hydrological event then Eq.(5.4) becomes in terms of the parameters of this model

$$P(\mu, \sigma, \rho | D) \propto P(\mu, \sigma, \rho) \cdot l(D | \mu, \sigma, \rho) \quad (5.6)$$

where $P(\mu, \sigma, \rho | D)$ is the joint posterior pdf of the parameters of the lag-one Markov process, $P(\mu, \sigma, \rho)$ is the joint prior pdf in general, of the same parameters and lastly $l(D | \mu, \sigma, \rho)$ is the likelihood function. For the ARIMA(1,0,1) process which is the main concern of this chapter the posterior pdf is given by

$$P(\mu \sigma \phi \theta | D) \propto P(\mu \sigma \phi \theta) \cdot l(D | \mu \sigma \phi \theta) \quad (5.7)$$

where the terms have similar interpretations to the terms in Eq.(5.6)

Some properties of these terms must be derived , before a detailed study of them can be given for a particular process. First of all the prior pdf, $P(\omega)$, is independent of the historic data and the model to be employed in the generating scheme, this is because of the fact that the sources of information for the prior pdf can be derived from any one of the following areas

i) The experience gained by the hydrologist about the phenomenon from previous similar studies

ii) Consideration of all the recorded data all over the world, the extraction of an estimate of the parameter from each set of data and then frequency analysis of the estimates obtained, to which subsequently a theoretical pdf curve can be fitted ; see Lenton et. al. (1973)

iii) Mere theoretical considerations, such as a normality assumption for the prior pdf of the parameter in order to render a complicated derivation into a tractable form ; see Conover (1970)

A concrete example of a prior pdf is provided as follows. Assuming that attention is confined to the historic sequence of annual streamflows at a fixed site in a given catchment area, the Bayesian estimate of the mean value of this sequence is to be calculated. From the streamflow measurements made at various points of the main river and its tributaries, a set of sequences each attached to one site in the drainage basin is obtained. Obviously, when the mean of the streamflow sequence at each site is found they will differ from each other to some extent. However, there are two assumptions that can be made at this stage; one is to assume that all of the means will eventually converge to the same population value as the length of sequences increase, in which case it is irrelevant to think of a pdf of this unique population value. Therefore, the prior pdf of population mean is a constant as a result of which from Eq.(5.4) it can be said that virtually

the posterior pdf will be proportional to the likelihood function only, from which the estimate of the mean can be extracted in the classical way that is by finding the value which maximizes the likelihood function.

However, on the contrary, the assumption may be that each of the mean values, at different sites in the same catchment area, will converge ultimately as the sample size increases to different population values which consequently illustrates the idea of the prior pdf of the population mean over that catchment area. The way to obtain this prior pdf is to fit a suitable theoretical pdf to the frequency distribution of these distinct population means. In a similar way the same argument can be pursued in order to find the pdf of any other parameter, such as the standard deviation, σ , the lag-one serial correlation coefficient, ρ , etc. In a work presented by Tschannerl (1970) the prior pdf of the mean of a streamflow sequence has been assumed to be a normal pdf without any practical justification.

Another constituent of Bayes theorem is the likelihood function which has already been mentioned briefly in chapter 3. All of the information included in an observed sequence is quantitatively represented by the likelihood function. The combination of the prior pdf and the likelihood function through Eq.(5.4) yield the posterior pdf. As it is evident from Eq.(5.4) when the prior pdf of a parameter is assumed to be a mere constant, then the posterior pdf becomes equal to the likelihood function, and as a result Bayes theorem is no longer valid but instead the sampling theory comes into play.

5.3. The Prior pdf of φ .

There are various sources of independent information about φ that give some insight into its prior pdf, $P(\varphi)$. First of all, although theoretically φ varies between -1 and +1 in hydrology for the annual streamflows, φ takes values mostly between 0 and +1. In addition to this,

as noted by O'Connell (1974) in relation to the agreement with Hurst's law obtained with the ARIMA(1,0,1) process, ϕ might be expected to lie in the range 0.80 - 0.95. Therefore, the expected value of the prior pdf could be assumed to be equal to any appropriate value between 0.80 and 0.95. Another very useful source of information would be a frequency diagram of which could be obtained from 140 annual streamflow sequences given by Yevjevich (1963). A theoretical curve fitted to this histogram would yield a prior pdf of ϕ , but this is not the approach adopted in this chapter.

However, by virtue of the relationship among ϕ , θ and ρ given as

$$\rho = \frac{(1 - \phi\theta)(\phi - \theta)}{1 + \theta^2 - 2\phi\theta} \quad (5.8)$$

the prior pdf of ϕ can be obtained from the prior pdf of ρ through Eq.(5.8) under the assumption that θ is a fixed value. Lenton et. al. (1973) gave the prior pdf of ρ on the basis of the world wide data presented by Yevjevich (1963), as the beta distribution on -1 to +1, which is expressed as

$$f(\rho) = (1 + \rho)^{r-1} (1 - \rho)^{t-r-1} \quad (5.9)$$

The first two moments of this distribution are

$$m = \frac{2r}{t} - 1 \quad (5.10)$$

$$\sigma^2 = \frac{4r(t-r)}{t^2(t+1)} \quad (5.11)$$

where the parameters r and t can be estimated from the data supplied by

Yevjevich (1963) by means of the method of moments. The values of r and t found by Lenton et. al. (1973) from Yevjevich data are

$$r = 15.499$$

and

$$t = 26.387$$

As a result, in this particular case the Beta distribution becomes

$$f(\rho) \propto (1 - \rho)^{k_1} (1 + \rho)^{k_2} \quad (5.12)$$

where $k_1 = t - r - 1 = 9.888$ and $k_2 = r - 1 = 14.499$. By substituting Eq.(5.8) in Eq.(5.12) the joint prior pdf of φ and θ is obtained as

$$f(\varphi, \theta) = N \left[1 - \frac{(1 - \varphi\theta)(\varphi - \theta)}{1 + \theta^2 - 2\varphi\theta} \right]^{k_1} \cdot \left[1 - \frac{(1 - \varphi\theta)(\varphi - \theta)}{1 + \theta^2 - 2\varphi\theta} \right]^{k_2} \quad (5.13)$$

Because of the difficulty encountered in dealing with the numerical calculation of a bivariate pdf given in Eq.(5.13) a simplification assumption will be made by regarding one of the variables as a constant quantity so as to examine the performance of the other variable. However, because of the fact that φ apparently plays a more important role than θ in the long-term properties of the ARIMA(1,0,1) process, all of the following work assumes θ to be a fixed to some preselected value between -1 and +1 thus leaving the Bayesian properties of φ only for examination. Alternatively, interest could be centered on the Bayesian behaviour of the parameter θ by fixing φ to a preselected value, and then by using the analytical expressions of the subsequent section, finding the posterior pdf of θ only. Although it is not a very tedious work to undertake, the latter approach is not adopted in this thesis, the only difference between the two being numerical integration

with respect to either φ or Θ as necessary.

However, a third alternative, which entails the combined Bayesian behaviour of φ and Θ can be considered, that is, to allow both φ and Θ to vary, in which case the numerical integration necessary to derive the contours of the joint posterior pdf of φ and Θ become enormous, although not impossible to tackle. This approach has not been adopted here.

5.4. The Likelihood Function of the ARIMA(1,0,1) Process.

After the prior pdf the second component of Bayes theorem leading to the posterior pdf is the likelihood function, $l(D|\omega)$. The explicit form of the ARIMA(1,0,1) process has been provided by O'Connell (1974) as

$$X_t = \mu(1 - \varphi) + \varphi X_{t-1} + \sigma(\eta_t - \Theta \eta_{t-1}) \quad (5.14)$$

where μ and σ are the mean and standard deviation of the process respectively, whereas φ and Θ are parameters related to the autocorrelation structure given by Eq.(1.21) and Eq.(1.22) and finally η_t and η_{t-1} are normally and independently distributed r.v's at time instances t and $t - 1$, with zero mean and variance given by

$$\sigma^2 = \frac{1 - \varphi^2}{(1 + \Theta^2 - 2\varphi\Theta)} \quad (5.15)$$

The first step to be taken in search for the likelihood function of the ARIMA(1,0,1) process, is to find the expression of the likelihood function of the random component, η_t , which can be written without any difficulty because the $\eta_0, \eta_1, \eta_2, \dots, \eta_n$ are normally and independently distributed. One important point to consider is that, for generating n observations by using Eq.(5.14), $(n + 1)$ values of η_t 's are required, including the initial r.v. η_0 . Therefore, the general form of the

likelihood function becomes because of independence

$$l(\eta_1, \eta_2, \eta_3, \dots, \eta_n | \eta_0) = f(\eta_0) \prod_{i=1}^n f(\eta_i) \quad (5.16)$$

where the $f(\eta_i)$ have normal pdf's given as

$$f(\eta_i) = \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp\left(-\frac{\eta_i^2}{2\sigma_\eta^2}\right) \quad (5.17)$$

As a result by substituting Eq.(5.17) in Eq.(5.16),

$$l(\eta_1, \eta_2, \eta_3, \dots, \eta_n | \eta_0) = \frac{f(\eta_0)}{n\sqrt{2\pi\sigma_\eta^2}} \exp\left(-\frac{1}{2\sigma_\eta^2} \sum_{i=1}^n \eta_i^2\right) \quad (5.18)$$

Various assumptions concerning the pdf of η_0 , denoted by $f(\eta_0)$, will be considered in a later section, but it is useful at this stage to think of Eq.(5.18) as being conditional on η_0 .

Now, it can be deduced from Eq.(5.14) that, when a transition from the variables η_t to the actual observations, X_t , is performed the η_i 's in Eq.(5.18) will all be a function of both the parameters of the model namely φ and θ and the observations themselves. Such a transition will be achieved first by working out the Jacobian of the transformation of η_t to X_t which can be calculated as

$$J = \frac{\delta(\eta_1, \eta_2, \dots, \eta_n)}{\delta(x_1, x_2, \dots, x_n)} = \begin{vmatrix} 1/\sigma & 0 & 0 & \dots & 0 \\ \varphi/\sigma & 1/\sigma & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1/\sigma \end{vmatrix} \quad (5.19)$$

The Jacobian is an $(n \times n)$ lower triangular matrix whose determinant is equal to the product of the elements of the main diagonal only. Therefore,

$$|J| = \frac{1}{\sigma^n} \quad (5.20)$$

The likelihood of the ARIMA(1,0,1) process, thus turns out to be conditional on two initial variables, namely, X_0 and η_0 ; in general

$$l(x_1, x_2, \dots, x_n | \eta_0, x_0) \propto \frac{f(\eta_0)f(x_0)}{\sigma^n \sigma_\eta^n} \exp \left[-\frac{1}{2\sigma_\eta^2} \sum_{i=1}^n \eta_i^2(x_1, x_2, \dots, x_n) \right] \quad (5.21)$$

where $f(X_0)$ is the pdf of the initial observation X_0 and $\eta_i^2(x_1, x_2, \dots, x_n)$ is the new form of the r.v's in terms of the given observations. In writing Eq.(5.21) an implicit assumption made is that η_0 and X_0 are independent from each other. Otherwise the joint pdf of η_0 and X_0 should appear in Eq.(5.21).

However, in order to bring Eq.(5.21) into a more revealing form, the terms in the exponent must be explicitly written in terms of the given observations. This is achieved after some considerable algebra. First of all, to facilitate the algebra, instead of the actual observations their standardized counterparts will be assumed. Hence, with the standardized observations Eq.(5.14) can simple be re-written as,

$$y_t = \phi y_{t-1} + \eta_t - \theta \eta_{t-1} \quad (5.22)$$

where

$$y_t = \frac{X_t - \mu}{\sigma} \quad (5.23)$$

is the standardized observation. With these standardized values the η_i 's are now a function of $y_0, y_1, y_2, \dots, y_n$. The next step is to express the individual η_i in terms of standardized observations and y_0, η_0, φ and θ . From Eq.(5.22) given η_0, y_0, φ and θ the r.v's, η_i , may be expressed as

$$\eta_1 = y_1 - \varphi y_0 + \theta \eta_0$$

$$\eta_2 = y_2 + (\theta - \varphi)y_1 - \theta(\varphi y_0 - \theta \eta_0)$$

$$\eta_3 = y_3 + (\theta - \varphi)(y_2 + \theta y_1) - \theta^2(\varphi y_0 - \theta \eta_0) \quad (5.24)$$

⋮

$$\eta_n = y_n + (\theta - \varphi)(y_{n-1} + \theta y_{n-2} + \theta^2 y_{n-3} + \dots + \theta^{n-2} y_1) - \theta^{n-1}(\varphi y_0 - \theta \eta_0)$$

By defining a new variable as

$$A = \varphi y_0 - \theta \eta_0 \quad (5.25)$$

the substitution of Eq.(5.25) into Eq.(5.24) gives

$$\eta_1 = y_1 - A$$

$$\eta_2 = y_2 + (\theta - \varphi)y_1 - \theta A$$

$$\eta_3 = y_3 + (\theta - \varphi)(y_2 + \theta y_1) - \theta^2 A \quad (5.26)$$

$$\eta_n = y_n + (\theta - \varphi)(y_{n-1} + \theta y_{n-2} + \theta y_{n-3} + \dots + \theta^{n-2} y_1) - \theta^{n-1} A$$

Now, by taking the square of each equation in the above system of equations given by Eq.(5.26) and then adding them all together to form the summation in Eq.(5.21) the likelihood function of the ARIMA(1,0,1) process becomes

$$l(x_1, x_2, \dots, x_n | \eta_0, x_0, \mu, \sigma, \varphi, \theta) \propto \frac{f(\eta_0) f(x_0)}{\sigma^n \sigma_\eta^n} \exp \left\{ -\frac{1}{2 \sigma_\eta^2 \sigma^2} \left[\sum_{i=1}^n (x_i - \mu)^2 + 2(\theta - \varphi) \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^{n-j} (x_i - \mu)(x_{i+j} - \mu) \right. \right. \\ \left. \left. + (\theta - \varphi)^2 \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{j-i} (x_i - \mu) \right]^2 - 2A \sum_{i=1}^n \theta^{i-1} (x_i - \mu) \right. \right. \\ \left. \left. - 2A(\theta - \varphi) \sum_{i=2}^n \theta^{i-1} \sum_{j=2}^i \theta^{i-j} (x_{j-1} - \mu) + A^2 \frac{1 - \theta^{2n}}{1 - \theta^2} \right] \right\} \quad (5.27)$$

where the standardized variables have been converted back into the original variables. Before proceeding further the form of the pdf's of both η_0 and x_0 must be specified. In fact, both η_0 and x_0 are normally distributed

and furthermore X_0 will be assumed to have the same mean and standard deviation as the observations X_t and in the same way the mean and standard deviation of the initial r.v. η_0 will be the same as r.v. η_i . With these assumptions Eq.(5.27) becomes

$$\begin{aligned}
 l(x_1, x_2, \dots, x_n | \eta_0, x_0, \mu, \sigma, \varphi, \theta) \propto & \frac{1}{\sigma^{n+1} \sigma_\eta^{n+1}} \exp \left\{ -\frac{(x_0 - \mu)^2}{2\sigma^2} \right. \\
 & - \frac{\eta_0^2}{2\sigma_\eta^2} - \frac{1}{2\sigma_\eta^2 \sigma^2} \left[\sum_{i=1}^n (x_i - \mu)^2 + 2(\theta - \varphi) \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^{n-j} (x_i - \mu)(x_{i+j} - \mu) \right. \\
 & + (\theta - \varphi)^2 \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{j-i} (x_i - \mu) \right]^2 - 2A \sum_{i=1}^n \theta^{i-1} (x_i - \mu) \\
 & \left. \left. - 2A(\theta - \varphi) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} (x_{i-1} - \mu) + A^2 \frac{1 - \theta^{2n}}{1 - \theta^2} \right] \right\} \quad (5.28)
 \end{aligned}$$

Therefore, the general expression of the posterior pdf defined by Eq.(5.7) becomes

$$\begin{aligned}
 P(\mu, \sigma, \varphi, \theta | D, \eta_0, x_0) \propto & \frac{P(\mu, \sigma, \varphi, \theta)}{\sigma^{n+1} \sigma_\eta^{n+1}} \exp \left\{ -\frac{(x_0 - \mu)^2}{2\sigma^2} - \frac{\eta_0^2}{2\sigma_\eta^2} \right. \\
 & - \frac{1}{2\sigma_\eta^2 \sigma^2} \left[\sum_{i=1}^n (x_i - \mu)^2 + 2(\theta - \varphi) \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^{n-j} (x_i - \mu)(x_{i+j} - \mu) \right. \\
 & \left. + (\theta - \varphi)^2 \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{j-i} (x_i - \mu) \right]^2 - 2A \sum_{i=1}^n \theta^{i-1} (x_i - \mu) \right.
 \end{aligned}$$

$$\left. -2A(\theta - \varphi) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} (x_{i-1} - \mu) + A^2 \frac{1 - \theta^{2n}}{1 - \theta^2} \right] \quad (5.29)$$

which can be written in a more compact form as shown in appendix (5.2)

$$P(\mu, \sigma, \varphi \mid \eta_0, x_0, \theta, D) \propto \frac{P(\varphi \mid \theta)}{\sigma^{n+2} \sigma_{\eta}^{n+1}} \exp(-P)$$

where the huge expression of P has been displayed in appendix (5.1).

5.5. The Posterior pdf of φ .

By definition, the posterior pdf of a parameter is proportional to the product of the prior pdf of the same parameter and the likelihood function of the generating process. In the case of more than one parameter the prior and posterior distributions will be the joint pdf's of the parameters concerned. For the ARIMA(1,0,1) process the joint posterior pdf of the parameters μ , σ , φ , and θ cannot be found until the form of the joint prior pdf of the parameters is specified. It is almost always impossible to propose a single joint prior pdf and because of this, ways of simplifying the joint prior pdf concerned are sought.

The simplest form would be to assume all the parameters to be independently distributed from each other prior to obtaining the data, but this would be a very unrealistic assumption, especially for the ARIMA(1,0,1) process where the independence of θ and φ is disproved by Eq.(5.8). Theoretically, the mean and the standard deviation are independent from each other for the normal processes as was shown in appendix (2.3). In practice, the assumption of independence between the mean and the standard deviation cannot be justified because of the fact that rivers with large means tend to

have large variances and as quoted by Lenton et. al. (1973), the coefficient of variation of all rivers tend to be concentrated around the value 0.2-0.3. Therefore, when the inverse of the coefficient of variation is adopted, which is given as $v = \mu / \sigma$, which does not vary very much in the case of streamflow, the independence of v and σ can be deduced (Lenton et. al. 1973). It is reasonable to assume that ϕ and θ are independent from μ and σ and hence, from v because of the fact that, all of the knowledge about ϕ and θ do not give any information about μ and σ and vice versa.

By taking into account all the above points the joint pdf of the ARIMA(1,0,1) process parameters can be re-written as product of various pdf's:

$$P(v, \sigma, \phi, \theta) = P(v) \cdot P(\sigma) \cdot P(\phi, \theta) \quad (5.30)$$

where $P(v)$ and $P(\sigma)$ are the prior pdf's of the coefficient of variation and the standard deviation respectively. The joint prior pdf, $P(\phi, \theta)$ of ϕ and θ on the r.h.s. of Eq.(5.30) can be thought of as a conditional pdf as explained in Section (5.3). Therefore,

$$P(v \sigma \phi \theta) = P(v) P(\sigma) P(\phi | \theta) \quad (5.31)$$

is the final form of the prior pdf where $P(\phi | \theta)$ is the conditional prior pdf of ϕ given θ , which can be found numerically from Eq.(5.13) and then incorporated in Bayes theorem as

$$P(v, \sigma, \phi | \eta_0, x_0, \theta, D) \propto P(v, \sigma, \phi, \theta) I(x_1, x_2, \dots, x_n | \eta_0, x_0, v, \sigma, \phi, \theta) \quad (5.32)$$

where the l.h.s. denotes the joint posterior pdf of the parameters v , σ and ϕ and the data itself. For annual hydrological streamflow sequences the distribution of v is assumed to be a constant. As to the prior pdf of σ

there exists no hydrological work but it has been borrowed here from the probability theory as suggested by Jeffreys (1961) the logarithm of σ is assumed to be uniformly distributed over the range $(0, +\infty)$, indicating a complete ignorance about the behaviour of σ . Hence,

$$P(\sigma) \propto \frac{1}{\sigma} \quad (5.33)$$

In the general expression of the posterior pdf, all of the terms that include μ have to be converted into v before the integration is executed. As a result, by substituting Eq.(5.33) and Eq.(5.28) into Eq.(5.32) the general form of the joint posterior pdf of μ , σ and φ is obtained as

$$\begin{aligned} P(\mu, \sigma, \varphi | \eta_0, x_0, \theta, D) \propto & \frac{P(\varphi | \theta)}{\sigma^{n+2} \sigma_{\eta}^{n+1}} \exp \left\{ -\frac{(x_0 - \mu)^2}{2\sigma^2} - \frac{\eta_0^2}{2\sigma^2} \right. \\ & - \frac{1}{2\sigma_{\eta}^2 \sigma^2} \left[\sum_{i=1}^n (x_i - \mu)^2 + 2(\theta - \varphi) \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^{n-j} (x_i - \mu)(x_{i+j} - \mu) \right. \\ & + (\theta - \varphi)^2 \sum_{j=1}^{n-1} \sum_{i=1}^j \theta^{j-i} (x_i - \mu)^2 - 2A \sum_{i=1}^n \theta^{i-1} (x_i - \mu) \\ & \left. \left. - 2A(\theta - \varphi) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} (x_{i-1} - \mu) + A^2 \frac{1 - \theta^{2n}}{1 - \theta^2} \right] \right\} \quad (5.34) \end{aligned}$$

In this form Eq.(5.34) seems to be difficult to integrate with respect to μ or v . To reduce it into a more manageable form it is necessary to introduce some constants by expanding each of the terms in the brackets of the exponent one by one as shown in appendix (5.2), resulting in

$$P(v, \sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(v)P(\varphi|\theta)}{\sigma^{n+2} \sigma_\eta^{n+1}} \exp \left\{ -\frac{C_1}{2\sigma_\eta^2 \sigma^2} \left[(\sigma v - E_1)^2 + D_1 \right] \right\} \quad (5.35)$$

where

$$E_1 = -\frac{B_1}{2C_1} \quad (5.36)$$

$$D = \frac{A_1}{C_1} - \left(\frac{B_1}{2C_1} \right)^2 \quad (5.37)$$

where A_1 , B_1 and C_1 are given in appendix (5.2). The joint prior pdf of the parameters μ , σ and ρ of the lag-one Markov process can be obtained by substituting $\eta_0 = 0$, $\theta = 0$ and $\varphi = \rho$ into Eq.(5.35).

5.5.1. Process with Three Unknown Parameters (μ , σ , φ)

In the following evaluations all of the numerical constants are absorbed into the proportionality sign. In the most general form the parameters μ , σ , φ and θ for the ARIMA(1,0,1) process will all be assumed as r.v's with prior pdf's attached to them. In the previous sections θ has already been assumed to be constant, hence, in general, the joint posterior pdf of the remaining three parameters is given by Eq.(5.35) where v is employed instead of μ .

However, to find the individual posterior pdf of any one of the parameters Eq.(5.35) has to be integrated with respect to the other two

parameters, hence eliminating their effects. With the assumption that v has a uniform prior pdf over the range $(-\infty, +\infty)$, then Eq.(5.35) becomes,

$$P(\sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+2} \sigma_\eta^{n+1}} \exp\left(-\frac{C_1 D_1}{2 \sigma^2 \sigma_\eta^2}\right) \int_{-\infty}^{+\infty} \exp\left[-\frac{C_1}{2 \sigma^2 \sigma_\eta^2} (\sigma v - E_1)^2\right] dv \quad (5.38)$$

where $P(v)$ has been absorbed into the proportionality sign. The integral in Eq.(5.38) has been evaluated in appendix (5.2) by Eq.(A5.25) giving

$$P(\sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{\sqrt{C_1} \sigma^{n+2} \sigma_\eta^{n+1}} \exp\left(-\frac{C_1 D_1}{2 \sigma^2 \sigma_\eta^2}\right) \quad (5.39)$$

In order to derive the posterior pdf of φ only, Eq.(5.35) has to be integrated with respect to σ , as given in appendix (5.2) by Eq.(A5.31)

$$P(\varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{\sqrt{C_1}} F_1^{-(n+1)/2} \quad (5.40)$$

where

$$F_1 = A_1 - \frac{B_1^2}{4 C_1} \quad (5.41)$$

In the derivation of Eq.(5.40) the initial r.v. η_0 was assumed to be equal to zero, without any prior pdf. However, on the contrary if η_0 is assumed to be normally and independently distributed with zero mean and variance equal to σ_η^2 , then, a further constant, t , should be added to A_1 , which is given by

$$t = \eta_0^2 + \theta \eta_0 \sum_{i=1}^n \theta^{i-1} (x_i - \mu) + \theta \eta_0 (\theta - \varphi) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} (x_{i-1} - \mu) \\ + \theta^2 \eta_0^2 \frac{1 - \theta^{2n}}{1 - \theta^2} + 2 \varphi \theta \eta_0 x_{0-1} \frac{1 - \theta^{2n}}{1 - \theta^2} \quad (5.42)$$

and the term that is to be added to B_1 is

$$u = 2\theta\varphi\eta_0 \frac{1 - \theta^{2n}}{1 - \theta^2} \quad (5.43)$$

To check the correctness of the posterior pdf given by Eq.(5.40), it suffices to insert $\theta = 0$ in the formula. Hence, it reduces to the lag-one Markov process case for which the equivalent expressions have been given by Thonber (1967).

To get the posterior pdf Eq.(5.40) can be readily evaluated numerically, for any desired set of parameters μ , σ , φ , θ , N , η_0 and X_0 . Figures (5.1) - (5.4) show the graphical representation of the prior pdf of Eq.(5.13), the likelihood function of Eq.(5.28) and the posterior pdf of Eq.(5.40) under the assumption that the initial r.v. η_0 is equal to zero. A first glance at figures (5.1) - (5.4) gives the expected impression that although the prior pdf which represents all the available information before the data is obtained has a greater variance, the combination of this prior knowledge with the likelihood function, representative of information provided by data, results in a posterior pdf which has relatively smaller variance indicating that there has been some reduction in the uncertainty compared with the prior pdf. In order to see the effect of sample size various posterior pdf's with different sample sizes have been plotted in figures (5.5) - (5.8). As it is evident from these figures as the sample size increases the variance of the posterior pdf decreases thus ensuring a better Bayesian estimate of the parameter.

One very useful property of the Bayesian estimation procedure is that it is possible to revise the estimate using additional data, in which case the previously found posterior pdf will assume the role of the new prior pdf before the additional data become available. The whole scheme is then repeated to arrive at a new posterior pdf which it is hoped will have a

smaller variance than the previous posterior pdf, indicating a reduction in the amount of uncertainty.

5.5.2. Process with Two Unknown Parameters (σ, φ)

One of the approaches taken to the Bayesian estimation procedure by Thornber (1967) is to assume one of two of the population parameters to be known exactly. The application of this approach to the ARIMA(1,0,1) process with two unknown parameters is effected by assuming either μ or σ to be known and then studying the combined effect of two remaining parameters (σ, φ) or (μ, φ) on posterior pdf of φ . In this section the population mean has been assumed to be known leaving σ as a r.v. with a prior pdf defined by Eq.(5.33).

Therefore, all of the product terms are a function of μ only, and can be eliminated from Eq.(5.29) by assuming them to be constants and absorbed in the proportionality sign. The posterior pdf of φ is obtained after integrating Eq.(5.36) with respect to σ only, which leads to as shown in appendix (5.3)

$$P(\varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{B^{(n+1)/2}} \quad (5.44)$$

The numerical solution of Eq.(5.44) and the likelihood function for a given μ value have been plotted with the prior pdf in figures (5.9) - (5.12). In addition to these, the effect of sample size has been shown in figures (5.13) - (5.16). Again the same conclusions reached in the 3-parameter case remain valid, that is the variance of the posterior pdf is smaller than the variance of the prior pdf and moreover the increase in the sample size shows itself as a decrease in the variance.

5.5.3. Process with One Unknown Parameter (φ)

By assuming, both the mean and the standard deviation of the ARIMA(1,0,1) process to be known the posterior pdf of φ can simply be thought of being equal to Eq.(A5.22) in appendix (5.2) without further modifications and is given as

$$P(\mu, \sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+2} \sigma_{\eta}^{n+1}} \exp \left\{ -\frac{C_1}{2\sigma^2 \sigma_{\eta}^2} \left[(\mu - E)^2 + D_1 \right] \right\} \quad (5.45)$$

This equation has been evaluated numerically leading to the plottings of figures (5.17) - (5.24) which show a very strong effect of the likelihood function on the posterior pdf.

In order to be able to compare all of the three approaches, the corresponding posterior pdf's, with the same set of parameters, have been plotted on the same diagram as shown in figures (5.25) - (5.28). One of the prominent features of these figures is that as the number of known parameters increases the information content, which is measured by the inverse of the variance as first proposed by Matalas and Langbein (1962) will be increasing, this is because of the fact that each one of the known parameters reduces the amount of uncertainty resulting in the final product, which is the posterior pdf in this special case. Another property that can easily be deduced from these figures is that the amount of uncertainty reduced by the assumption that σ is known is comparatively smaller than the case when both σ and μ are known. As a result the assumption of known μ has far more greater reduction in uncertainty than known σ . These properties can be quantitatively accounted for by comparing the means and the standard deviations of various posterior probability distribution as given in table (5.1). As a conclusion it can be said that the estimate of the mean is rather important to the estimation of φ , the prominent control parameter of the long-term persistence of the ARIMA(1,0,1) process.

In order to be able to compare, quantitatively, the three types of posterior pdf's their means and standard deviations have been calculated and presented on table (5.1). It is evident from this table that type-I and type-II posterior pdf parameters are very close to each other which suggests that they are characteristically extremely similar to each other. This property confirms an already made assertion, that is, the effect of standard deviation on φ is very small indeed. On the other hand, the type-III posterior pdf is distinctly different from the other two posterior pdf's, hence, confirming the importance of the mean in estimating the parameter φ .

5.6. Bayes Estimates

The two prerequisites in the Bayesian estimation of any parameter are the posterior pdf of the parameter concerned and the loss function, $l(\varphi, \hat{\varphi})$, which can theoretically assume any form of function provided that it is non-negative for all values of φ_i and $\hat{\varphi}$ and has a unique zero value when $\varphi = \hat{\varphi}$ only. The loss function is a r.v. for finite lengths of observed sequences and consequently the loss function will have its own statistical properties such as the mean, the variance etc. In the Bayesian framework, the value of $\hat{\varphi}^*$ that minimizes the mathematical expectation of the loss function will be referred to as the Bayes estimation of the parameter

$$\varphi = \min_{\hat{\varphi}} E[l(\varphi, \hat{\varphi})] = \min_{\hat{\varphi}} \int_R l(\varphi, \hat{\varphi}) \cdot P(\varphi | \eta_0, x_0, \theta, D) d\varphi \quad (5.46)$$

where R is the region over which φ can assume values, that is, for the ARIMA(1,0,1) process $-1 \leq R \leq +1$.

It is a very difficult task to find a practical loss function that conforms with reality; because of this, so far in the hydrological literature, only theoretical loss functions have been used. One very frequently used loss

function is the quadratic loss function which is expressed as

$$l(\varphi, \hat{\varphi}) = (\varphi - \hat{\varphi})^2 \quad (5.46)$$

It has been shown by Zellner (1971) that when a quadratic loss function is considered then the Bayes estimate of the parameter is equal to the mean value of the posterior pdf. The posterior distribution of φ in all the above cases is conditional on Θ . However, what is desired is the overall optimal choice of φ and Θ i.e. corresponding to minimum expected loss. Assuming a loss function is available for φ the losses associated with various values of φ can be worked out for a particular value of Θ using the posterior pdf in Eq.(5.40). For the evaluation mentioned above the following table can be very usefull

$\varphi \backslash \Theta$	φ_1	φ_2	φ_3	φ_4	.	.	.	φ_n
Θ_1	l_{11}	l_{12}	l_{13}	l_{14}	.	.	.	l_{1n}
Θ_2	l_{21}	l_{22}	l_{23}	l_{24}	.	.	.	l_{2n}
Θ_3	l_{31}	l_{32}	l_{33}	l_{34}	.	.	.	l_{3n}
.	.	.	.	.	.	.	.	.
Θ_m	l_{m1}	l_{m2}	l_{m3}	l_{m4}	.	.	.	l_{mn}

where l_{ij} are the losses associated with φ_i and Θ_j . If l_{34} is the minimum loss then the Bayesian estimates of φ and Θ turn out to be φ_4 and Θ_3 respectively.

5.7. Summary

A general approach leading to the marginal probability distribution functions of various parameters of the ARIMA(1,0,1) model have been presented based on some simplifying assumptions. In particular the posterior pdf of φ which is the controlling parameter of long-term behaviour of an ARIMA(1,0,1) process has been derived. All of the analytical expressions obtained in this chapter can simply be reduced to the lag-one Markov process case, provided that $\varphi = \rho$ and $\Theta = 0$ are substituted in the derived equations.

Fig. (5.1) - Type-I

-.-.- Maximum Lik.
 - - - Prior pdf
 ——— Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 10$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

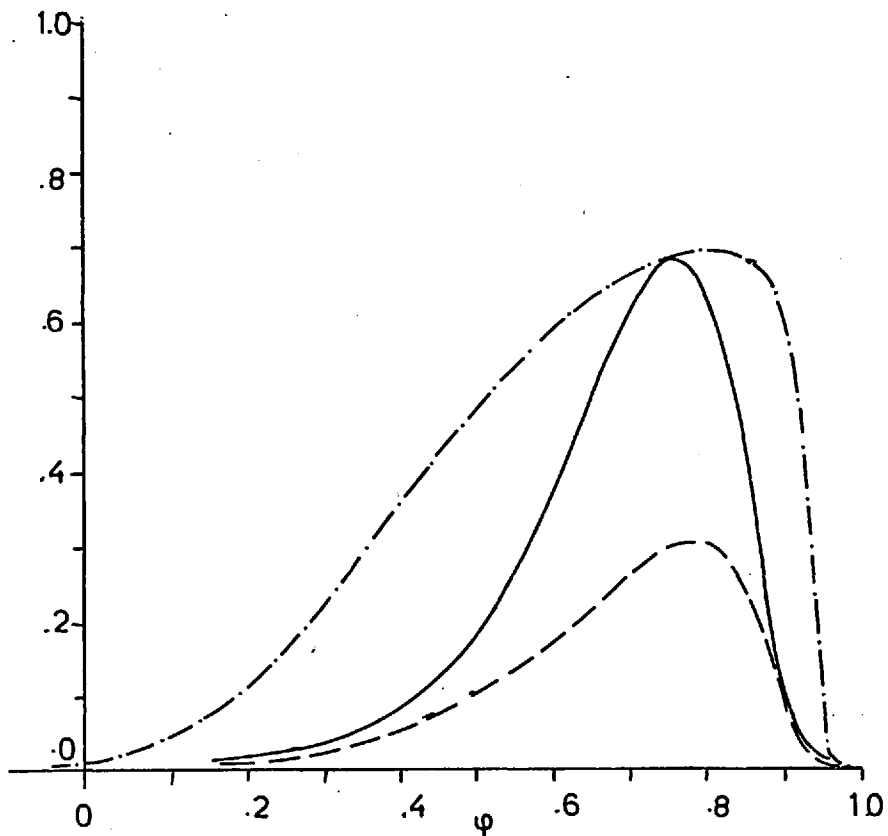


Fig. (5.2) - Type-I

-.-.- Maximum Lik.
 - - - Prior pdf
 ——— Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 20$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

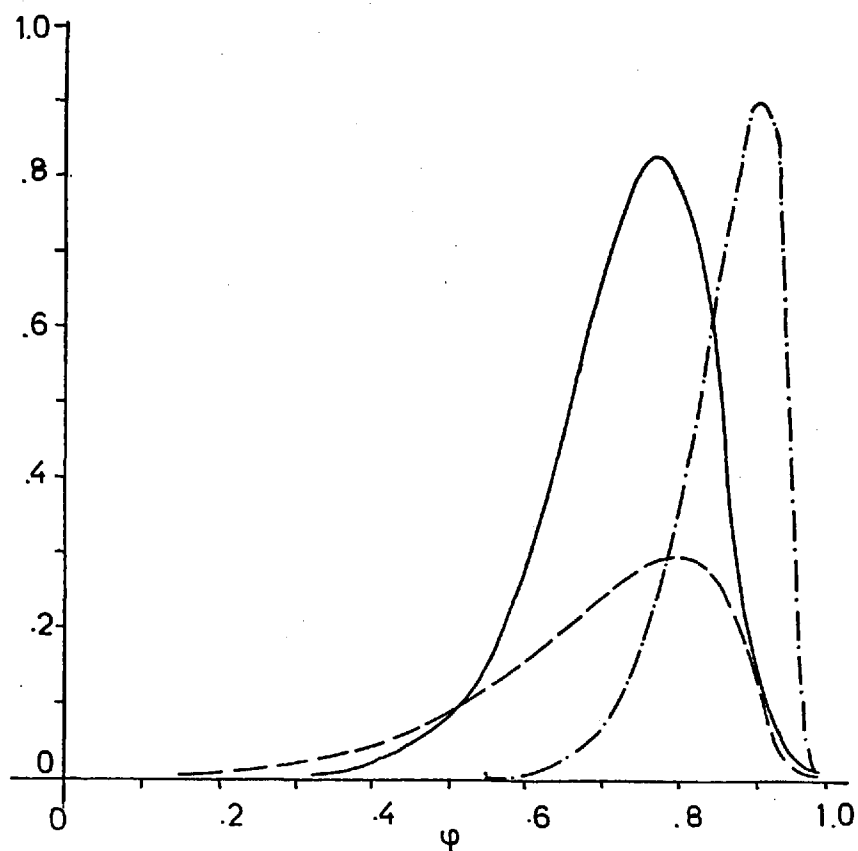


Fig. (5.3) - Type-I

..... Maximum Lik.
 ----- Prior pdf
 _____ Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 30$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

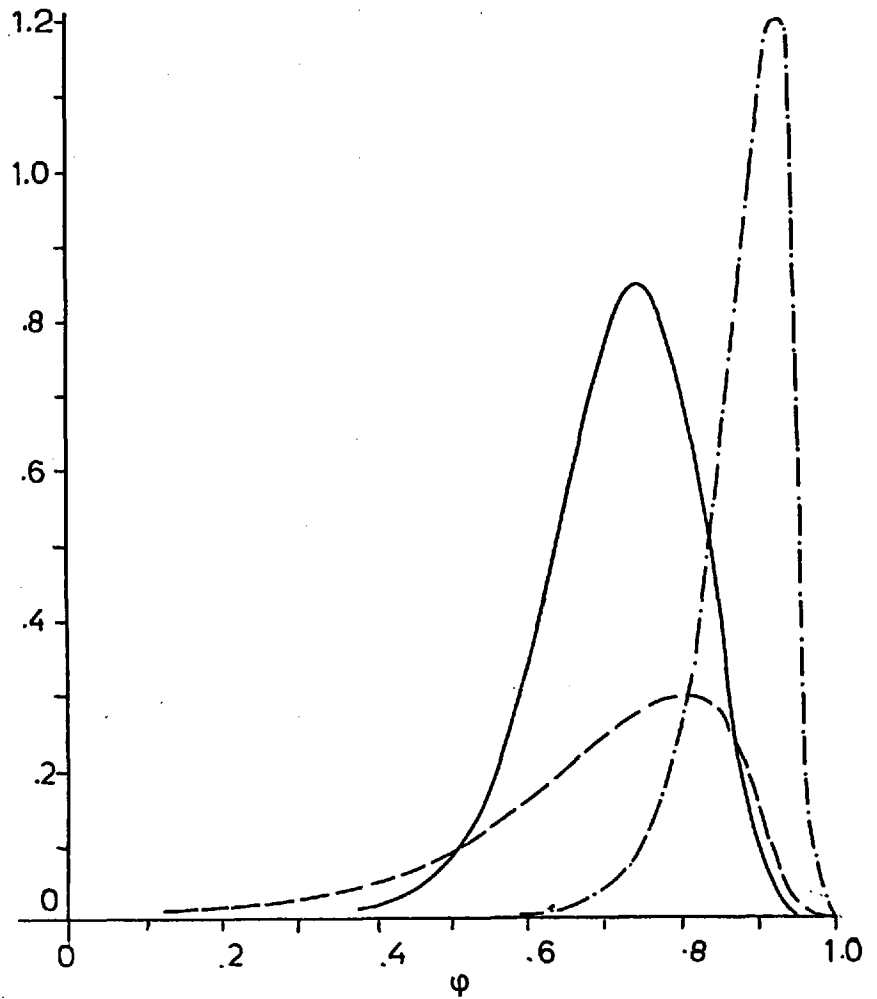


Fig. (5.4) - Type-I

..... Maximum Lik.
 ----- Prior pdf
 _____ Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 40$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

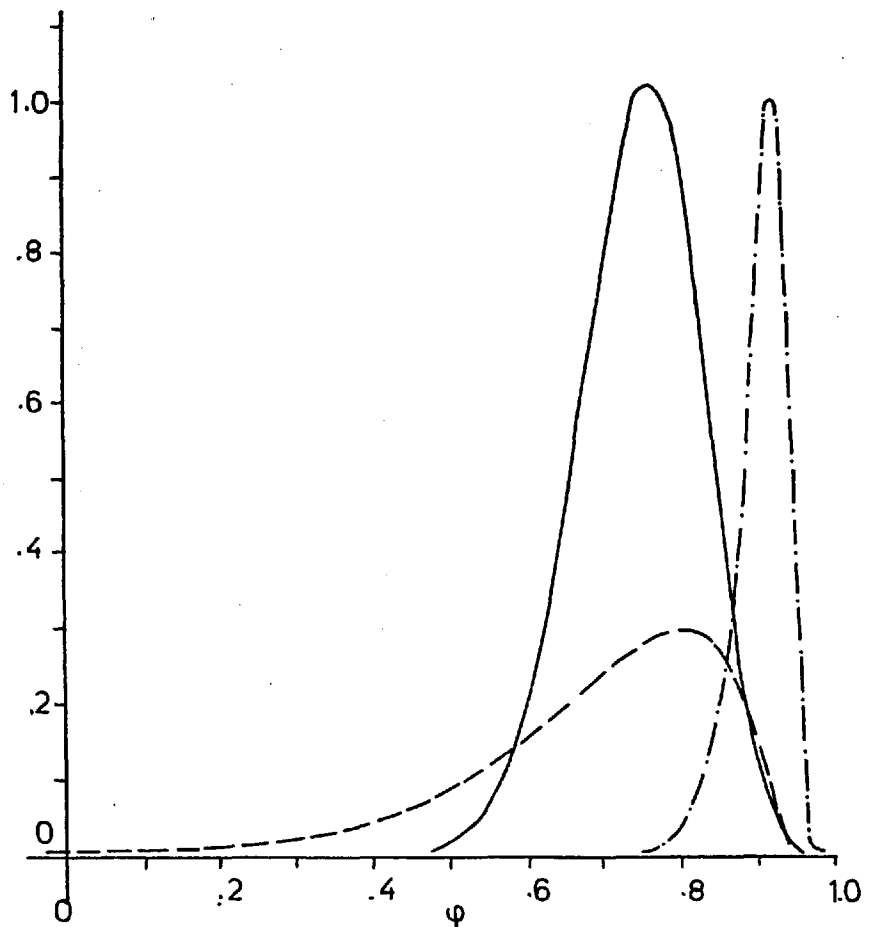


Fig. (5.5) - Type-I

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$\eta_0 = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$

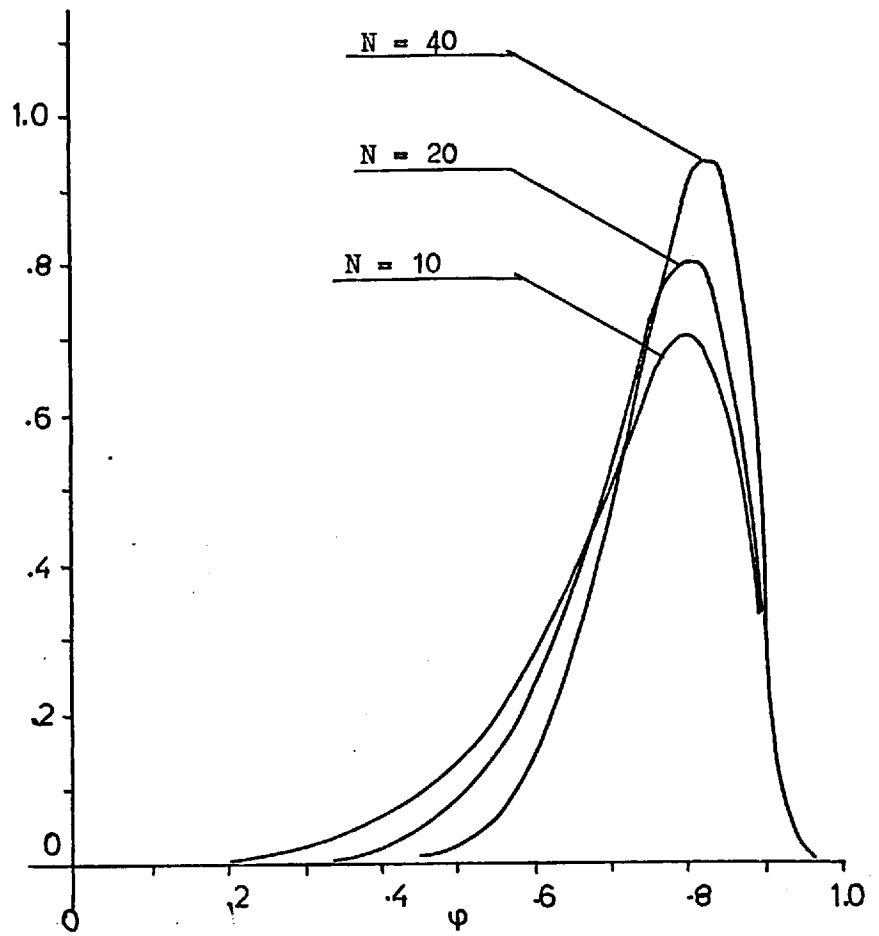


Fig. (5.6) - Type-I

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.80$$

$$\theta = 0.60$$

$$\eta_0 = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$

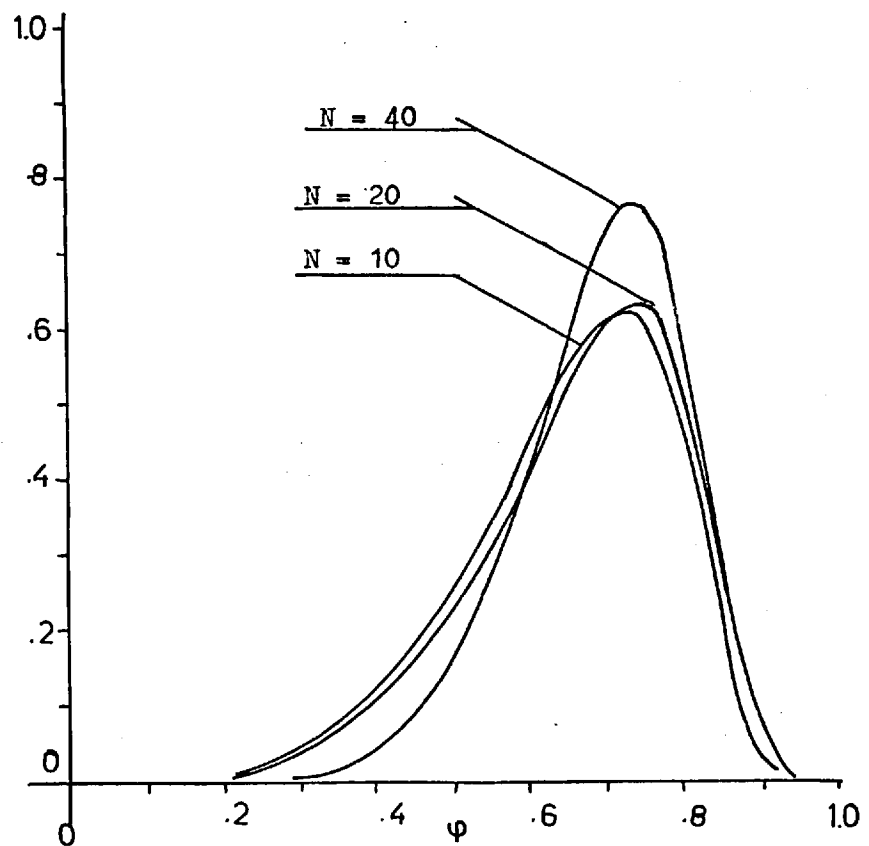


Fig. (5.7) - Type-I

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.70$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

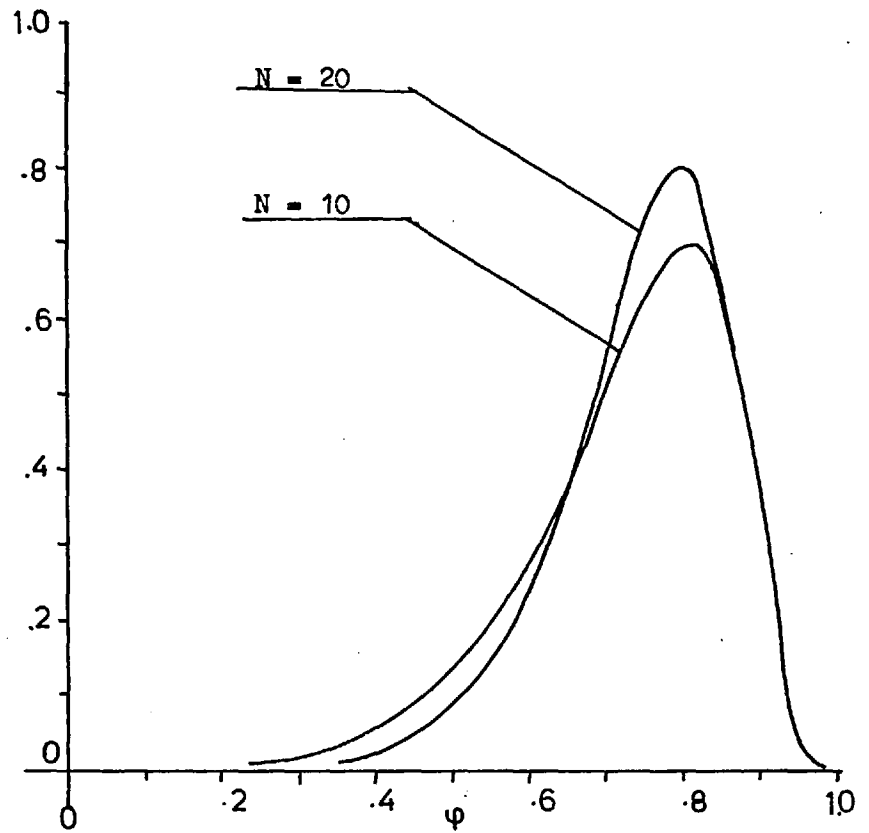


Fig. (5.8) - Type-I

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.50$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

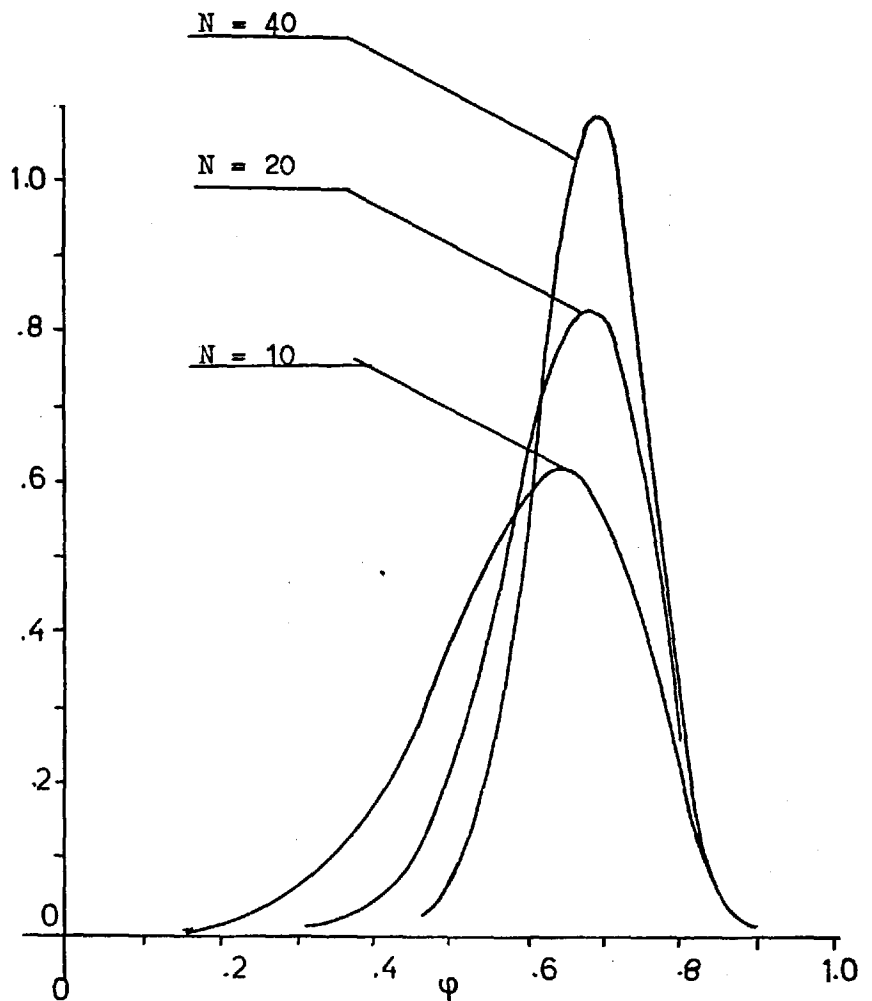


Fig. (5.9) - Type-II

..... Maximum Lik.
 ----- Prior pdf
 ————— Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 10$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

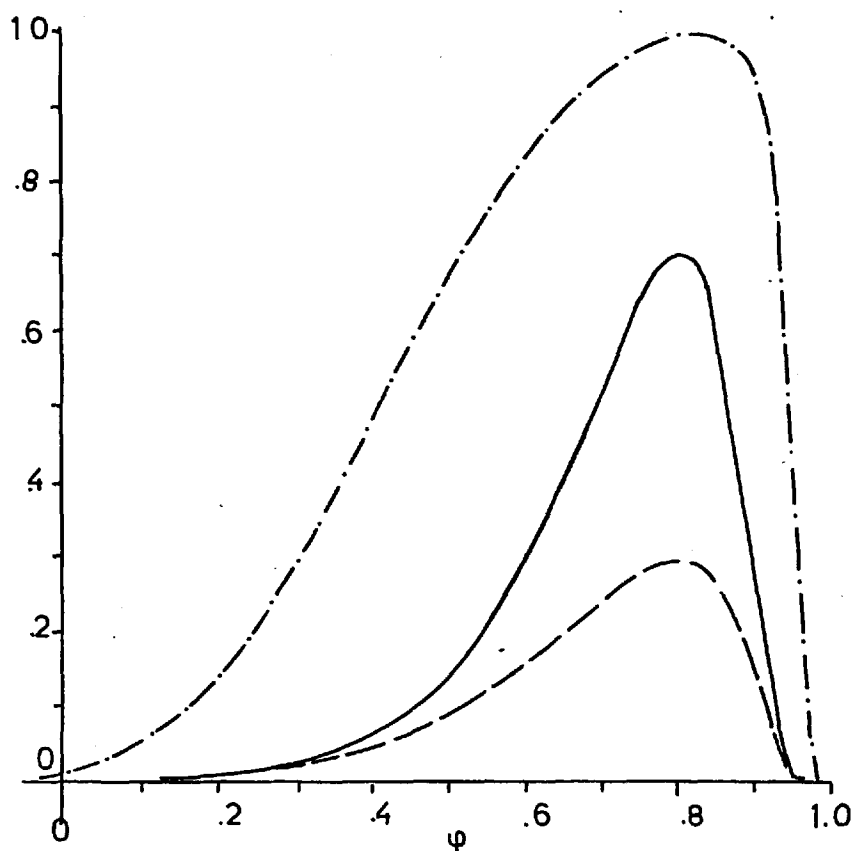


Fig. (5.10) - Type-II

..... Maximum Lik.
 ----- Prior pdf
 ————— Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 20$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

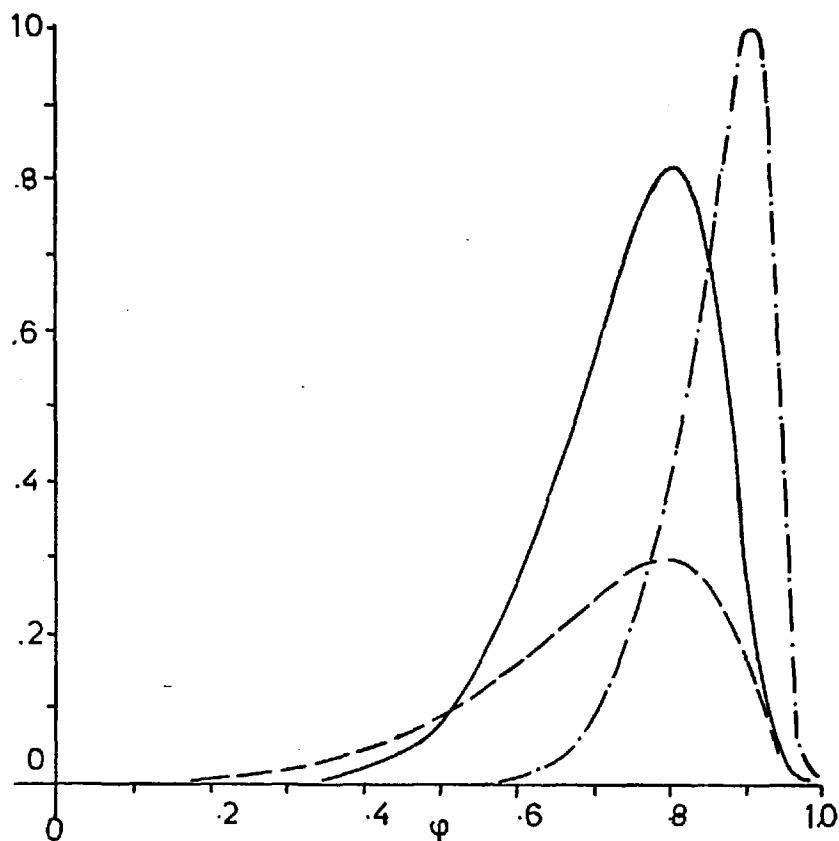


Fig. (5.11) - Type-II

..... Maximum Lik.
 ----- Prior pdf
 _____ Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 30$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

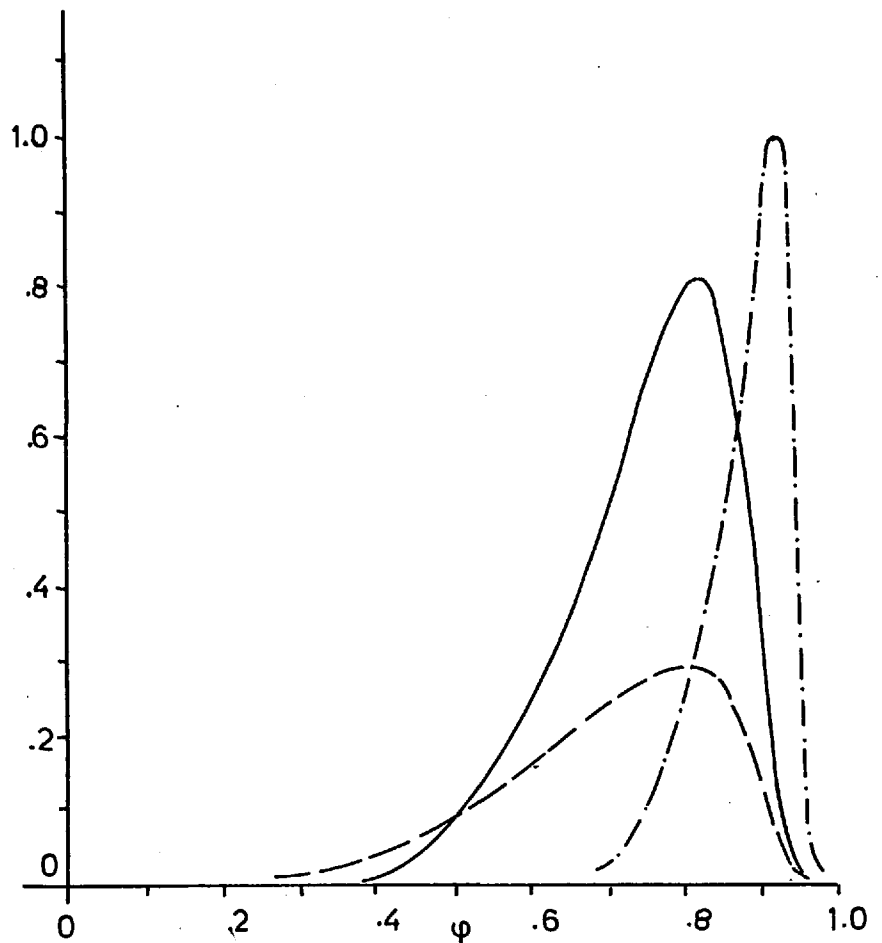


Fig. (5.12) - Type-II

..... Makimum Lik.
 ----- Prior pdf
 _____ Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.0$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 40$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

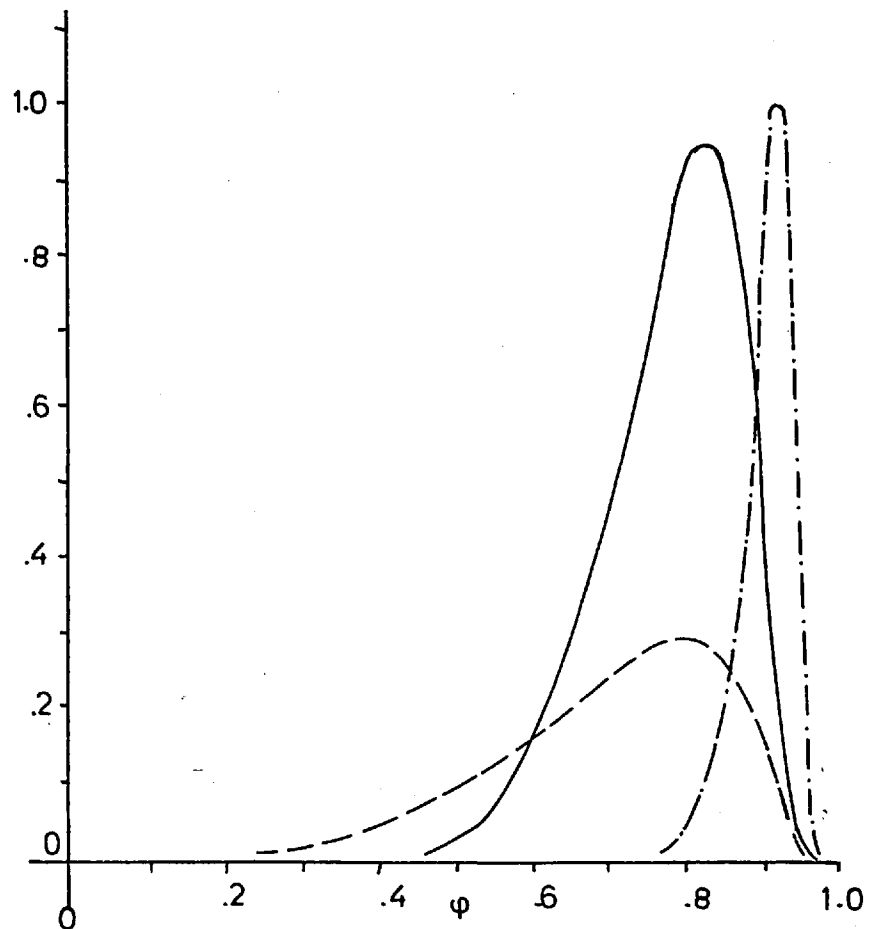


Fig. (5.13) - Type-II

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$\eta_0 = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$

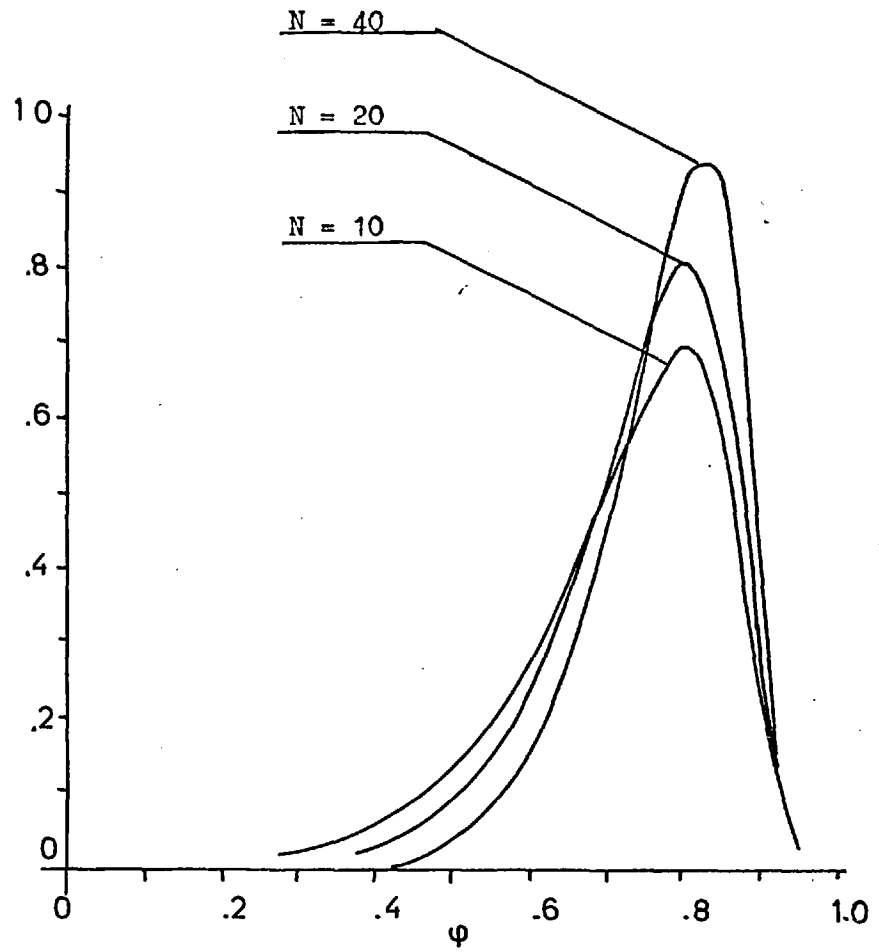


Fig. (5.14) - Type-II

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.80$$

$$\theta = 0.60$$

$$\eta_0 = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$

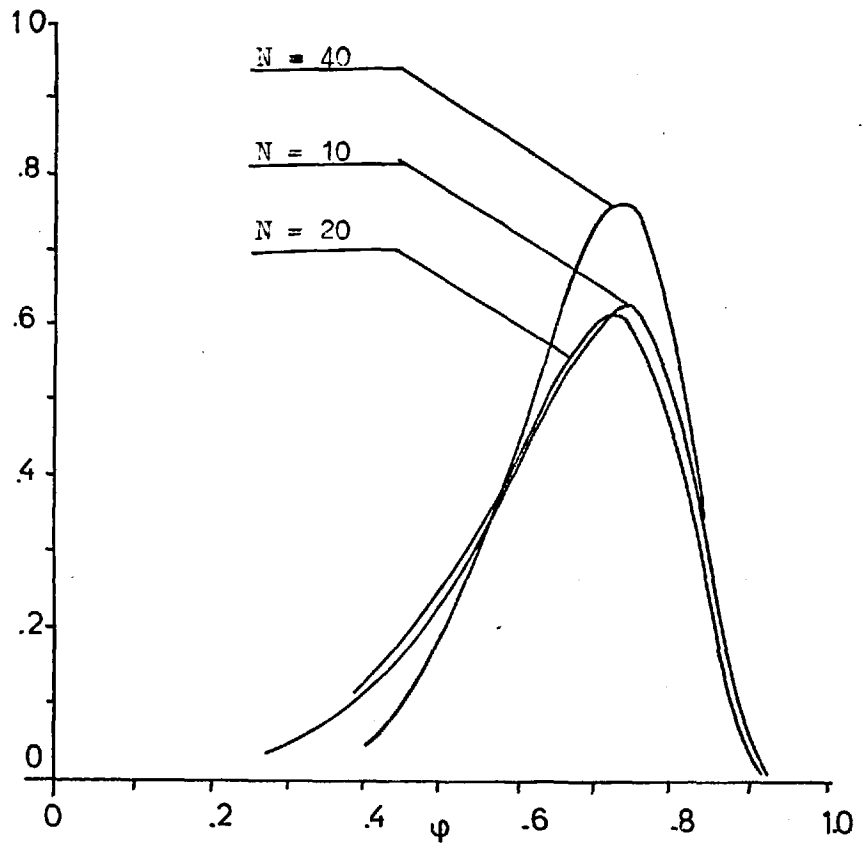


Fig. (5.15) - Type-II

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.70$$

$$\eta_0 = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$

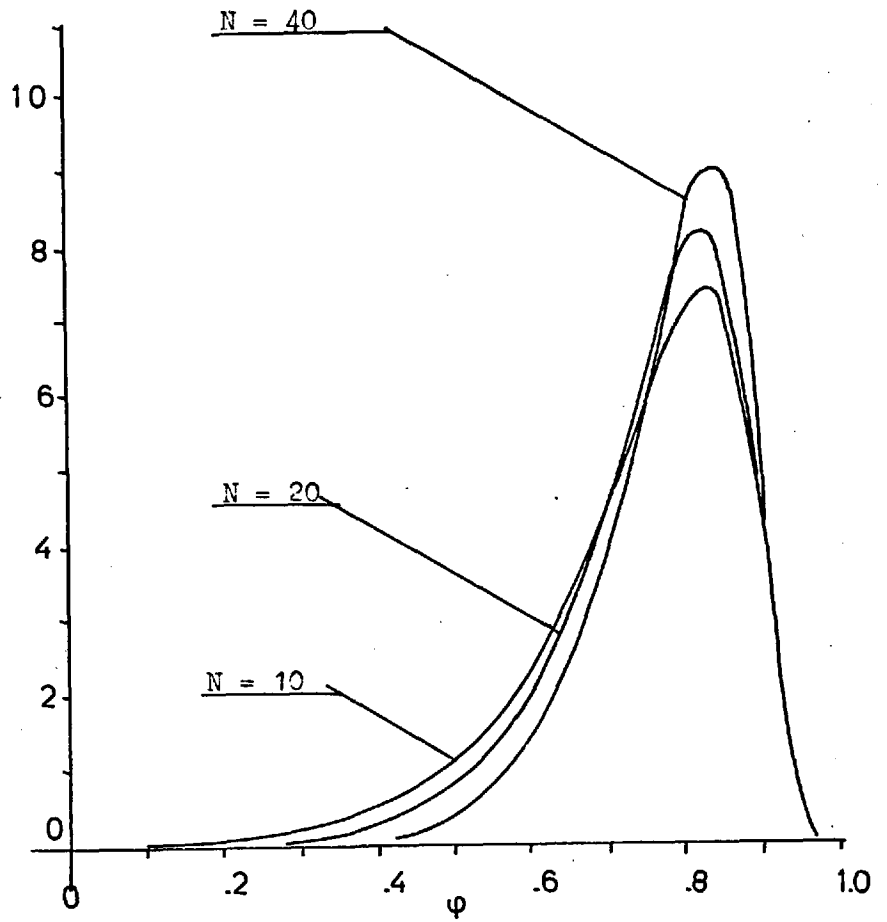


Fig. (5.16) - Type-II

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.50$$

$$\eta_0 = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$

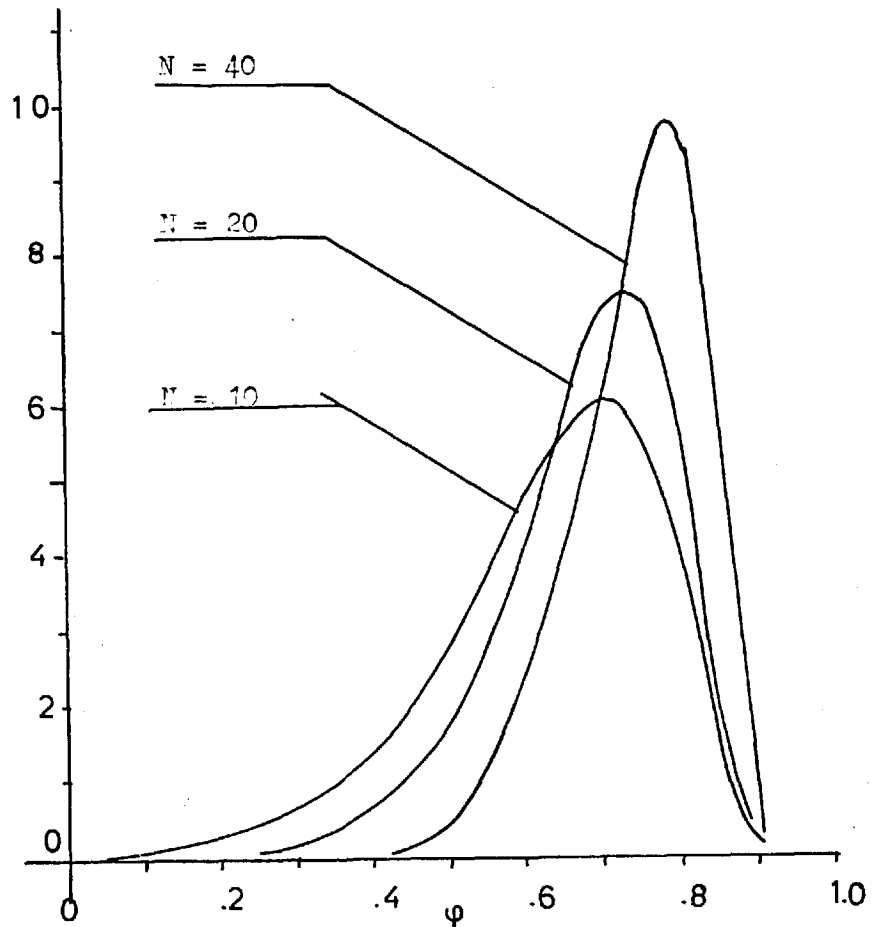


Fig. (5.17) - Type-III

..... Maximum Lik.
 ----- Prior pdf
 _____ Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 10$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

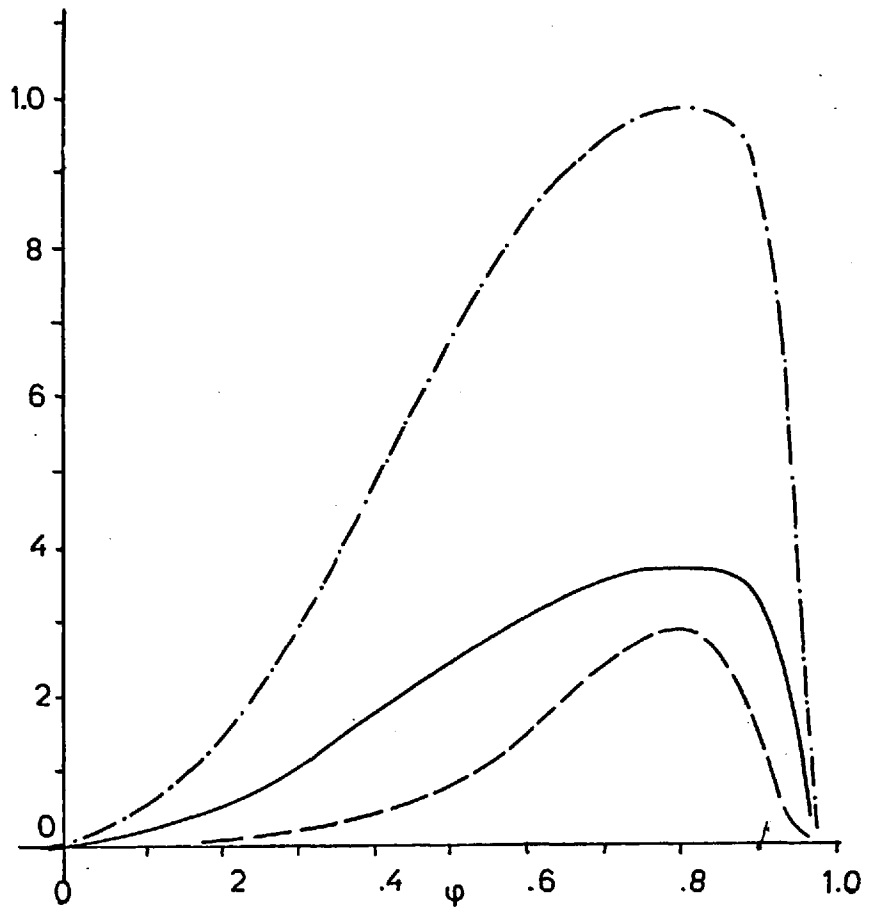


Fig. (5.18) - Type-III

..... Maximum Lik.
 ----- Prior pdf
 _____ Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 20$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

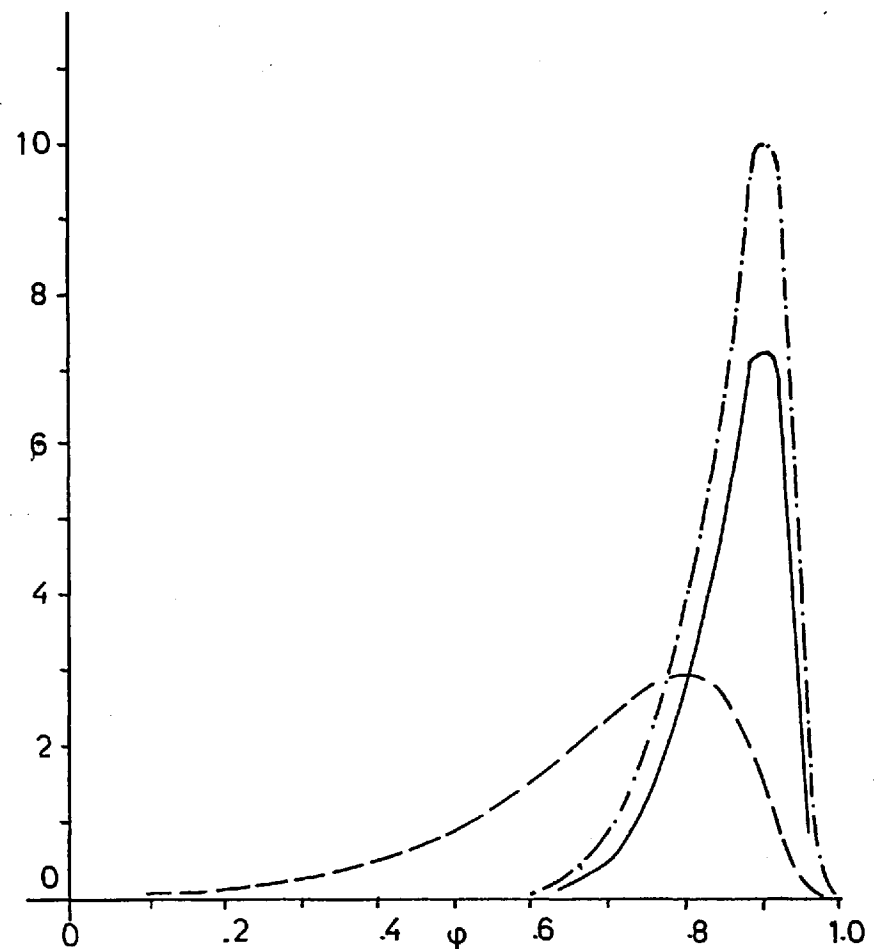


Fig. (5.19) - Type-III

..... Maximum Lik.
 ----- Prior pdf
 ————— Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 30$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

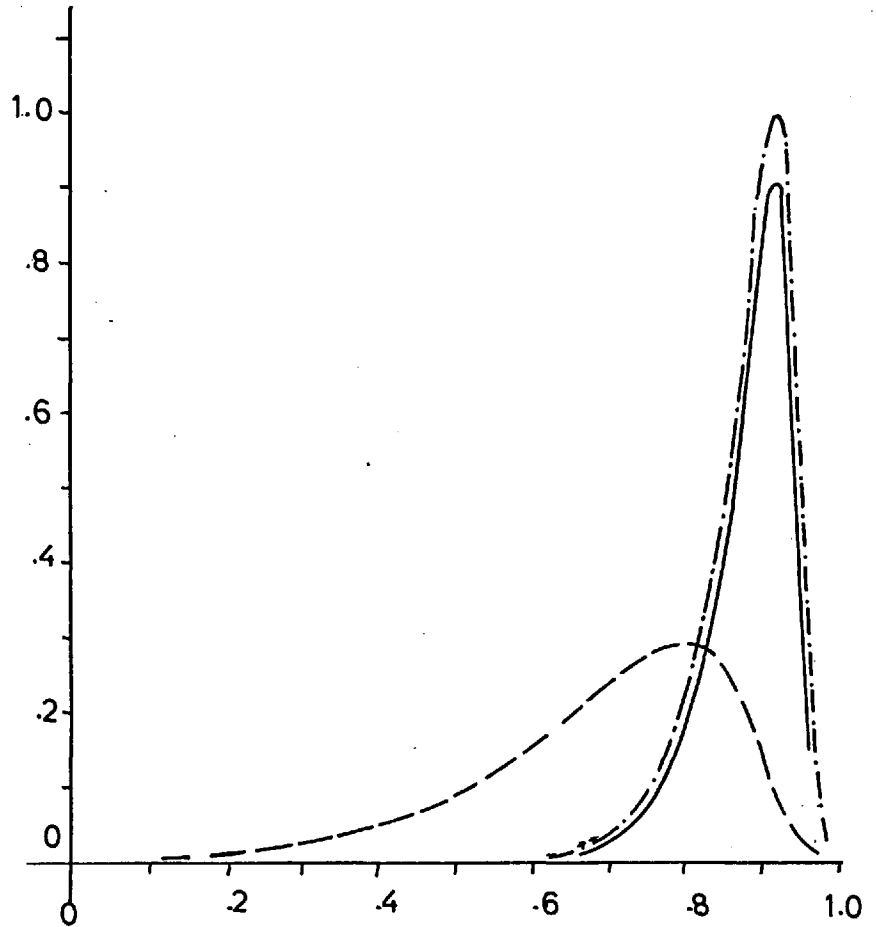


Fig. (5.20) - Type-III

..... Maximum Lik.
 ----- Prior pdf
 ————— Posterior pdf

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 40$$

$$\eta_0 = 0.0$$

$$x_0 \rightarrow N(\mu, \sigma^2)$$

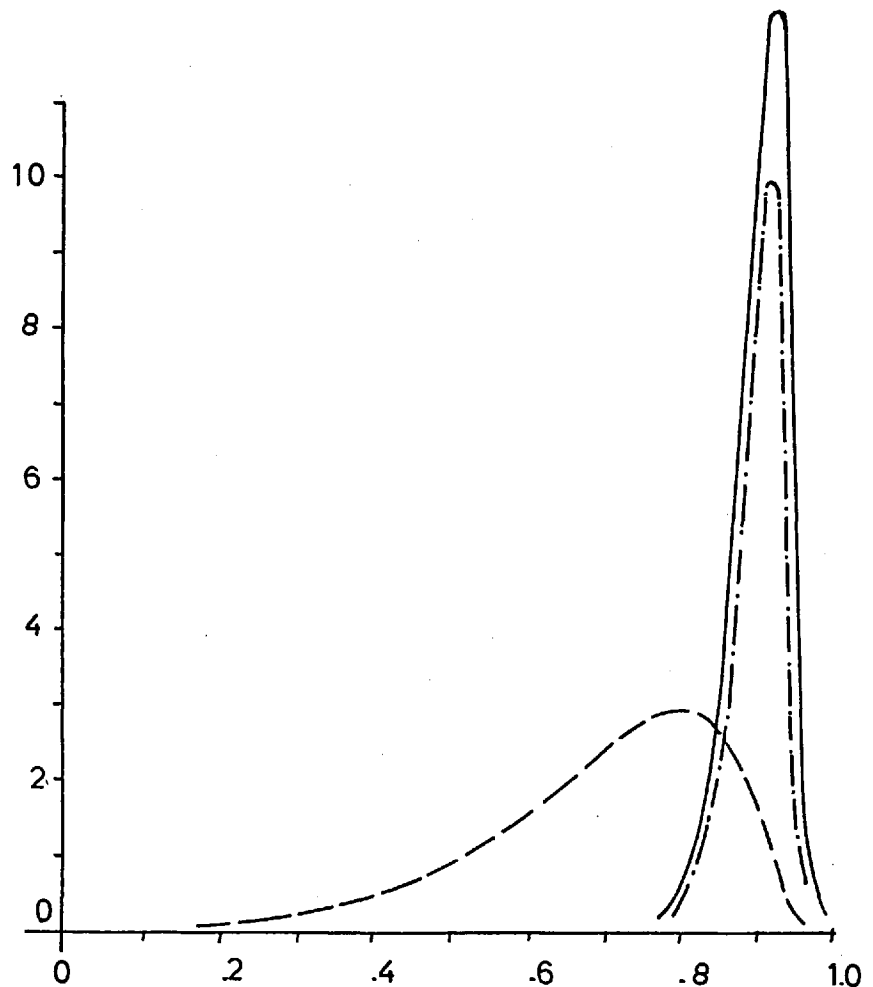


Fig. (5.21) - Type-III

$\mu = 10.0$
 $\sigma = 3.00$
 $\varphi = 0.90$
 $\theta = 0.65$
 $\eta_0 = 0.0$
 $x_0 \rightarrow N(\mu, \sigma^2)$

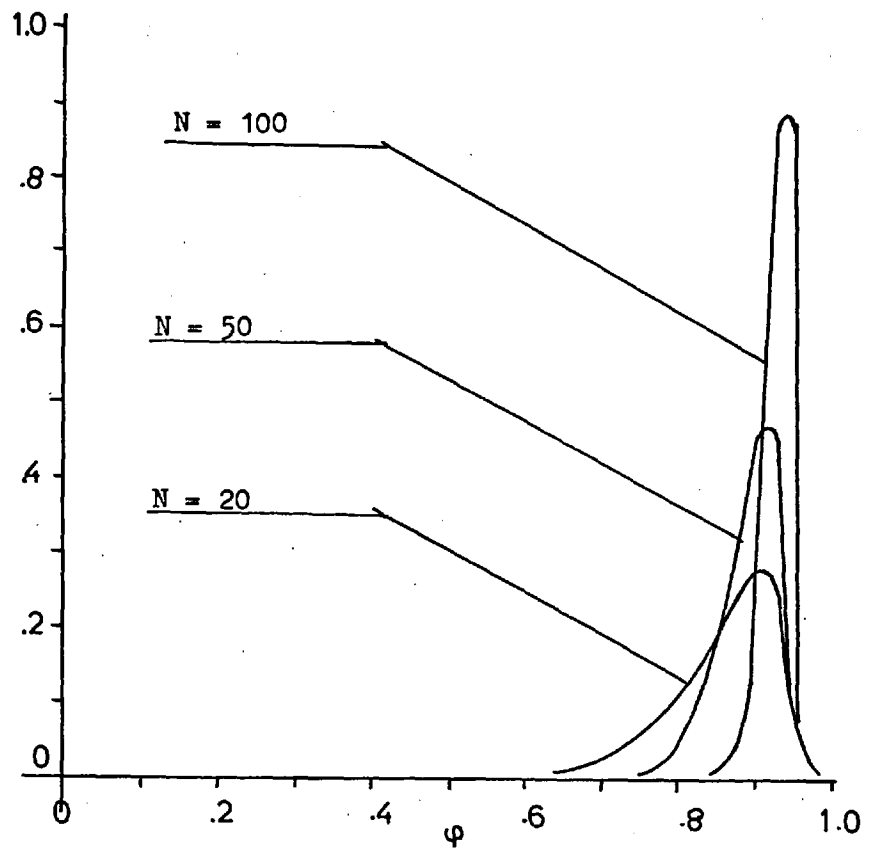


Fig. (5.22) - Type-III

$\mu = 10.0$
 $\sigma = 3.00$
 $\varphi = 0.90$
 $\theta = 0.65$
 $\eta_0 = 0.0$
 $x_0 \rightarrow N(\mu, \sigma^2)$

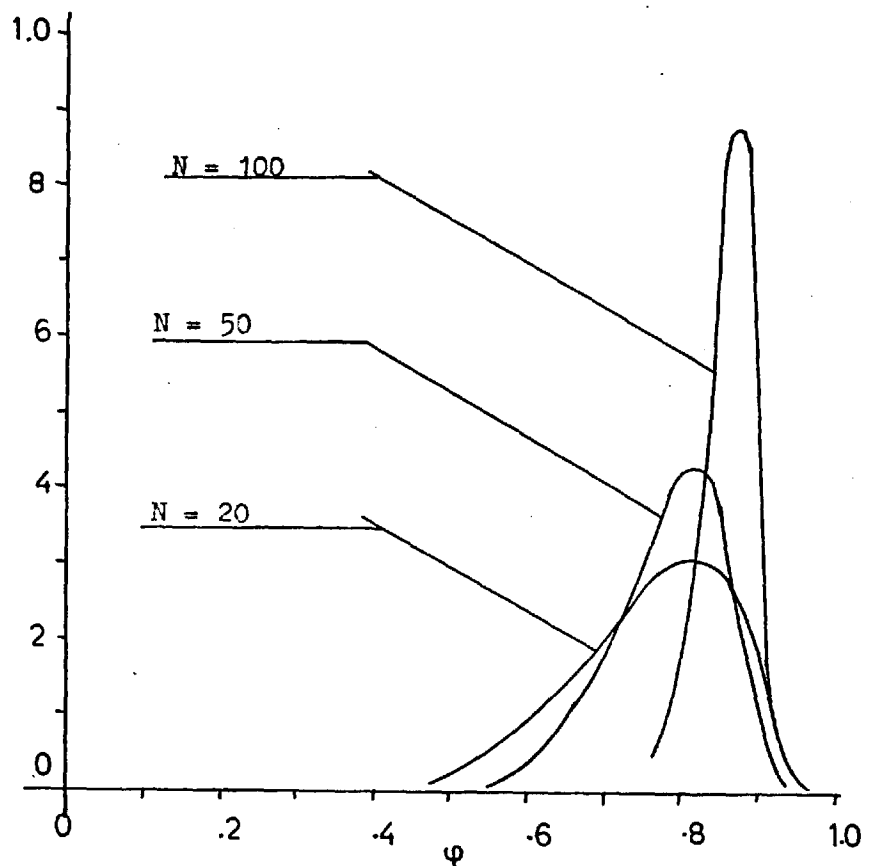


Fig. (5.23) - Type-III

$\mu = 10.0$
 $\sigma = 3.00$
 $\varphi = 0.90$
 $\theta = 0.70$
 $\eta_0 = 0.0$
 $x_0 \rightarrow N(\mu, \sigma^2)$

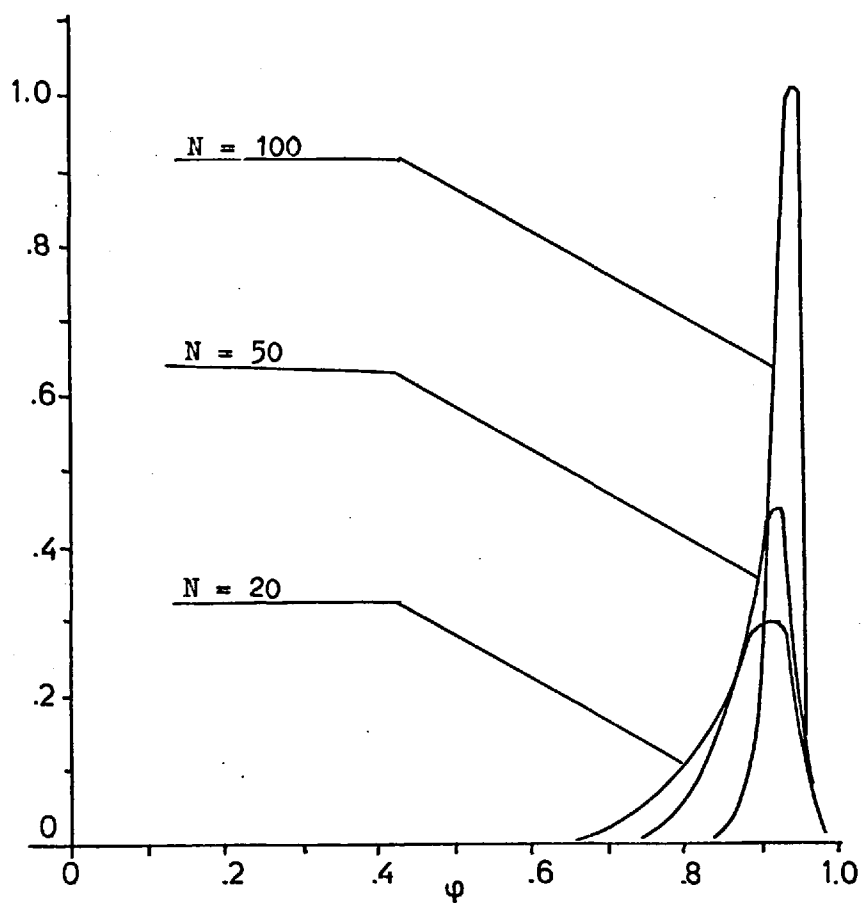


Fig. (5.24) - Type-III

$\mu = 10.0$
 $\sigma = 3.00$
 $\varphi = 0.90$
 $\theta = 0.50$
 $\eta_0 = 0.0$
 $x_0 \rightarrow N(\mu, \sigma^2)$

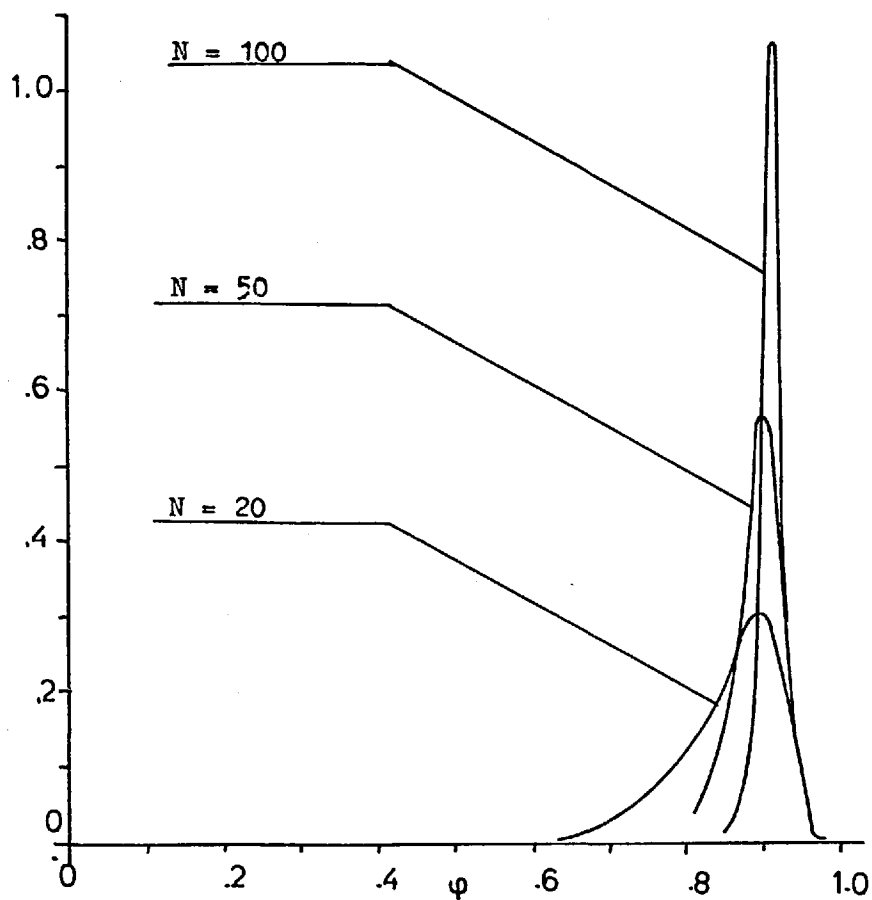


Fig. (5.25) -

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 20$$

$$\eta_0 = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$

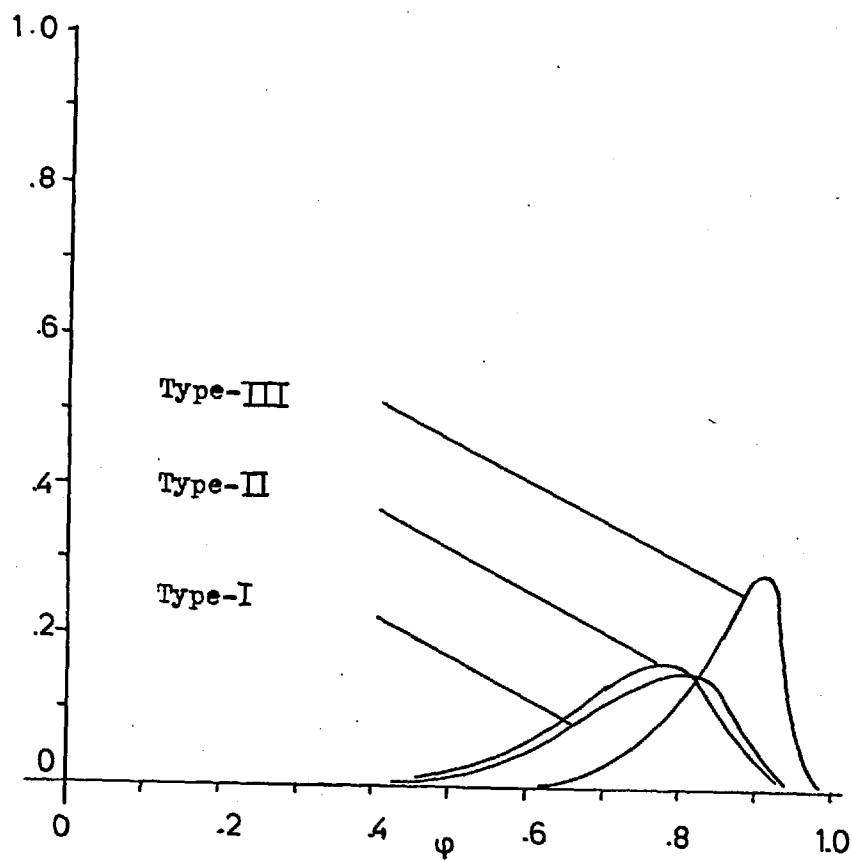


Fig. (5.26) -

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.65$$

$$N = 40$$

$$\eta = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$

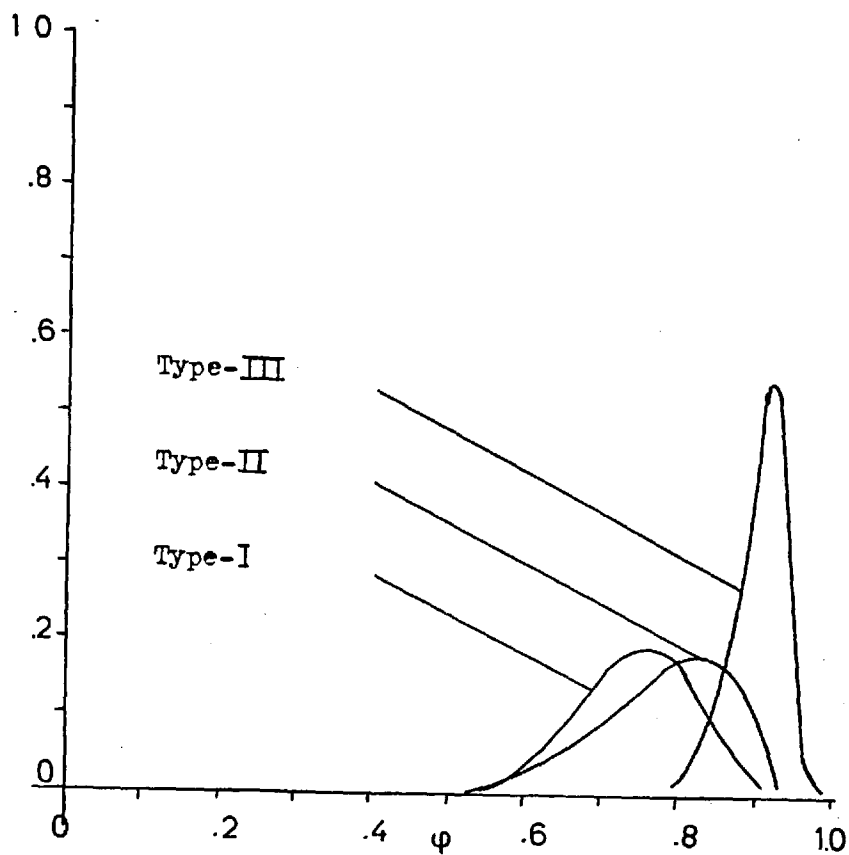


Fig. (5.27) -

$$\mu = 10.0$$

$$\sigma = 3.00$$

$$\varphi = 0.90$$

$$\theta = 0.80$$

$$N = 20$$

$$\eta_0 = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$

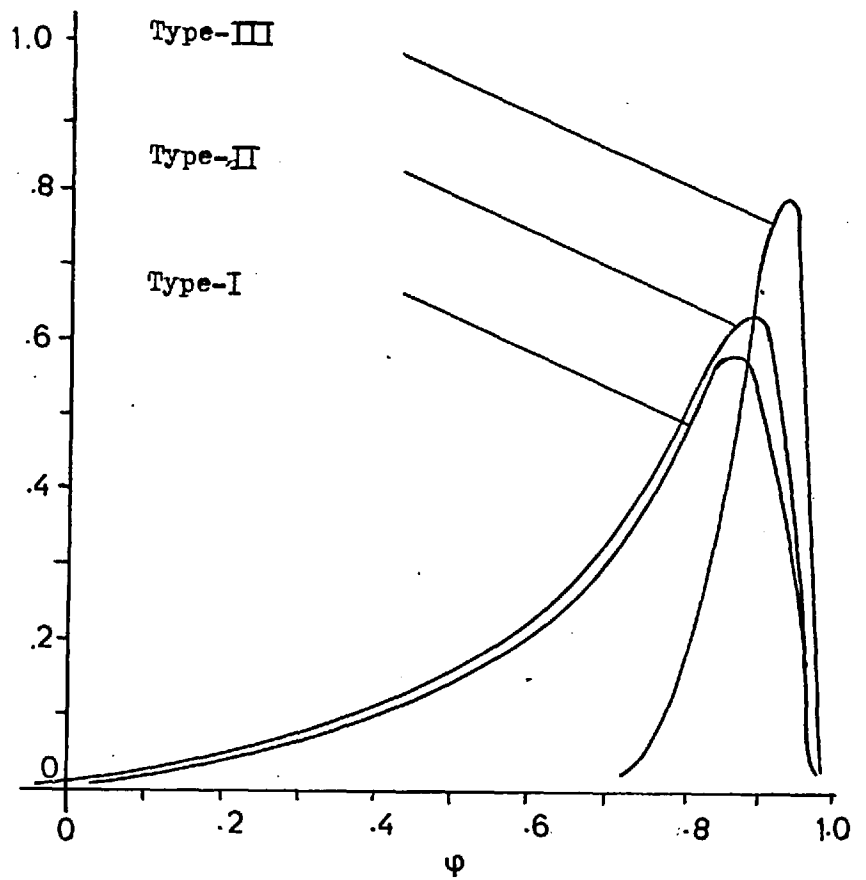


Fig. (5.28) -

$$\mu = 10.0$$

$$\sigma = 3.00$$

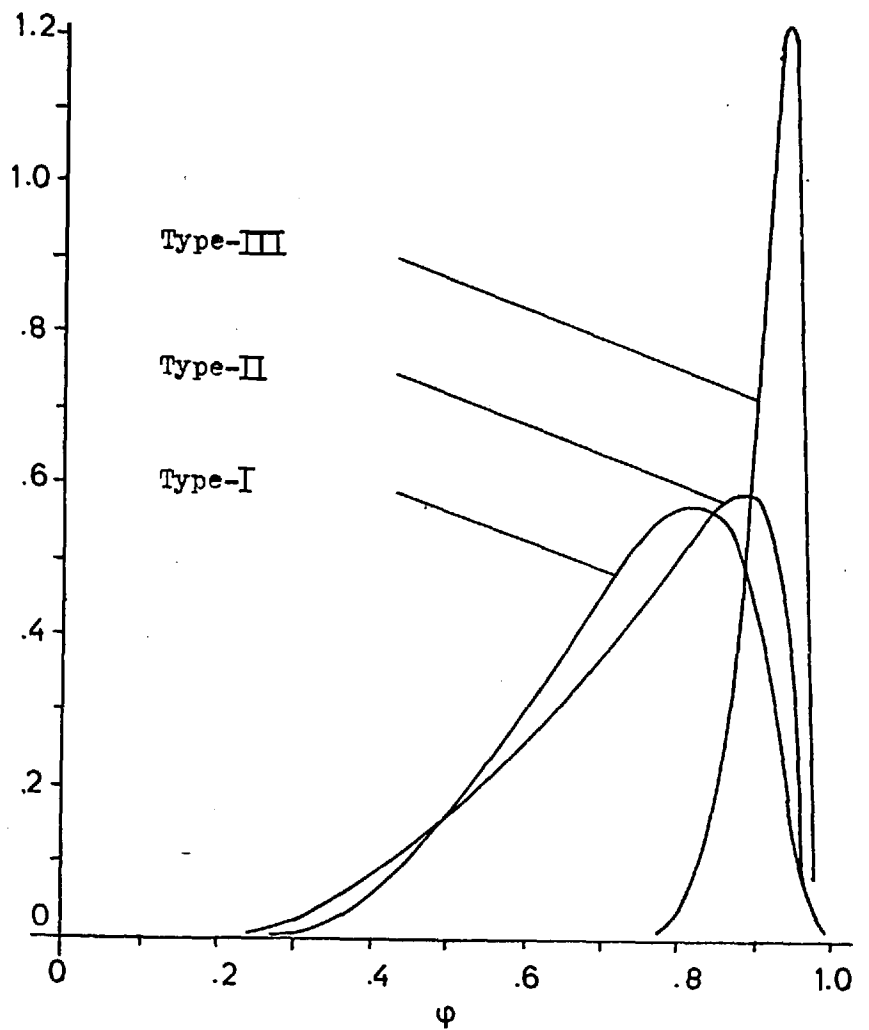
$$\varphi = 0.90$$

$$\theta = 0.80$$

$$N = 40$$

$$\eta_0 = 0.0$$

$$x_0 \longrightarrow N(\mu, \sigma^2)$$



SAMPLE SIZE	POSTERIOR μ			POSTERIOR σ		
	Type-I (φ, μ, σ)	Type-II (φ, σ)	Type-III (φ)	Type-I (φ, μ, σ)	Type-II (φ, σ)	Type-III (φ)
10	.704	.720	.645	.131	.135	.200
20	.733	.746	.864	.101	.108	.067
30	.720	.755	.885	.094	.098	.057
40	.740	.779	.918	.078	.084	.033
50	.745	.743	.924	.069	.072	.032

Table (5.1) - Mean and variance of the posterior pdf the
3 Bayesian approaches.

CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

6.1. General Summary

During the course of this study various interrelated aspects of stochastic hydrology have been examined with particular reference to the small sample properties of several currently used simulation models. Among such properties are the small sample expectation of the estimates of the most commonly employed parameters which are the lag-one autocorrelation coefficient, the standard deviation, the variance and the Hurst coefficient. Moreover, the rescaled range which is a measure of the long-term behaviour of a historic record has been analytically investigated and the expected value of it has been derived in small samples for a number of processes. All of the analytical expressions derived in this thesis have been checked by simulation methods on a digital computer.

A simple stochastic model with desirable long-term properties has been proposed and fully developed for simulating annual streamflow sequences either at a single site or a number of sites on a catchment area so as to preserve spatial as well as temporal correlations. The quantitative equivalence of the WM process to the ARIMA(1,0,1) process has been shown and small sample expectation of WM parameters have been analytically derived.

The posterior pdf of parameter ϕ of the ARIMA(1,0,1) process, which is a prerequisite for finding the Bayesian estimate of the same parameter has been derived on the basis of various simplifying assumptions

6.2. General Conclusions

The primary theme of this thesis has thus been the problems associated with small samples of hydrological variates and the Hurst phenomenon and the interaction between both. Some theoretical contributions to both aspects have been derived which shed further light on the generation of synthetic streamflows. The primary conclusions that might be drawn can be listed as follows

1) In general, the small sample properties of some vital parameters used in hydrology have been closely investigated and advances in analytical results have been achieved which should advance the understanding ^{of} / the behaviour of small samples of hydrological variates to a greater extent.

2) A considerable advance over the existing analytical results on the range of cumulative departures from the mean has been achieved. General expressions have been derived for the expected values of the adjusted and rescaled ranges which can be applied to any stationary stochastic processes. In particular,

a) The only difference between the analytical expressions for the expected value of range and rescaled range, irrespective of whether adjusted or population, is a mere constant factor for small samples whose value converges to unity asymptotically for large samples; hence the asymptotic expectation of the range is equal to that of the rescaled range.

b) The statistical properties of the rescaled range are distribution free by virtue of the central limit theorem mentioned in chapter 2. Furthermore, the rescaled range is a function of a t-distribution whose statistical properties are dependent on the sample length only.

c) As can be seen from an inspection of the analytical expressions given in chapter 2 the expected value of an ideal reservoir size which is equal to the range, increases with increasing lag-one autocorrelation coefficient; hence it is smallest when the inflows into the reservoir are coming from a NIP. Finally, the asymptotic distribution function has been obtained for the rescaled population range. However, because of the small samples available in hydrology it is not of a great interest but represents only a start. The same distribution for small samples would serve as a test of significance in practice.

3) The small sample properties of the lag-k serial correlation coefficient has been derived with particular reference to the ARIMA(1,0,1) process.

Moreover, the small sample expectation for the standard deviation has been analytically derived for any normal and stationary hydrologic process. In short, these expressions will give the hydrologist the ability of maintaining small sample statistical resemblance between synthetic and historic sequences without resorting to the cumbersome Monte Carlo techniques for deriving the small sample properties.

4) The white Markov process which is almost identical to the ARIMA(1,0,1) process except for the multivariate case has been fully developed for generating synthetic annual streamflows. In particular,

a) The WM process is a simple and adequate approximation to dfGn.

b) The observed long-term persistence in geophysical records can be successfully modelled by the WM process in addition to the simultaneous preservation of short-term persistence.

c) The small sample properties of the process have been analytically derived.

d) The multivariate case equations of this process are different from that of the ARIMA(1,0,1) process but have the same difficulty as regards the solution of them.

5) The general expressions for the posterior pdf of the parameters of the ARIMA(1,0,1) process are obtained as a prerequisite for the Bayesian approach to the parameter estimation procedure. It has been shown that it does not make a great difference whether the mean μ is known or unknown with regard to the estimation of the model parameter ϕ .

The idea of incorporating information other than that contained in the historic record to obtain a better estimate of a parameter is attractive, but in this thesis, formidable theoretical difficulties are encountered which limit this approach.

6.3. Suggestions for Future Research

Related to the aspects treated in this thesis there appear to be various potential areas that need further study in order to understand the stochastic processes involved in a better way and their use in the design and operation of water resources systems. These research areas can be given as follows.

1) The small sample expectation of the range or the ideal reservoir size has been derived analytically for the case of full development only; in the light of the work presented in chapter 2 of this thesis, the same expectation could also be derived analytically for the case of partial development.

2) Although the asymptotic distribution of the rescaled population range is obtained in chapter 2, it is not adequate to provide a formulation of a test of significance for long-term persistence because of the shortness of hydrological records. Therefore, one would hope to derive the distribution of the rescaled range in small samples.

3) The analytical expression to define the expected value of h in small samples could be derived on the basis of the analytical expression for $E(R_n/S)$ which has already been derived in chapter 2 of this thesis.

4) In the absence of the pdf of R_n/S for small samples, by making use of $E(R_n/S)$ only, the process that gives rise to the observed sequence could be identified. To this purpose curves of $E(R_n/S)$ versus n for the same lag-one autocorrelation coefficient but different processes could be plotted and then the estimate of R_n/S is extracted from a given historic sequence; whichever curve is closer to the estimate of R_n/S will then suggest a generating process.

5) The WM process could be developed to simulate monthly and daily streamflows either by applying the model directly or by the disaggregation of the annual streamflows as suggested by Valencia and Schaake (1972).

6) A comparison of the Bayesian approach and the classical approach to

estimating the parameter φ could be achieved through various types of loss functions.

7) The prior pdf of φ could be obtained from information other than that based on world wide data. To this end the information derived from a study of physical properties of the catchment area could be incorporated into the prior pdf. Some work in this direction has been performed by Moss (1972).

8) The Bayesian approach presented in this thesis could be extended to cover the optimum estimation of μ , σ and φ .

Appendix 2.1

The Adjusted Range for Lag-One Markov Process

In this appendix the values of quantities of $V(\bar{x}_k)$, $\text{Cov}(\bar{x}_k, \bar{x}_n)$ and $V(\bar{x}_n)$ in Eq.(2.39) are calculated for the lag-one Markov process.

The autocorrelation function of this process is a function of the first order autocorrelation coefficient, ρ , as given in Eq.(2.57). Hence,

$$V(\bar{x}_k) = \frac{1}{k^2} E\left(\sum_{i=1}^k x_i\right)^2$$

On expanding the r.h.s of this expression,

$$V(\bar{x}_k) = \frac{1}{k^2} \left[\sum_{i=1}^k E(x_i^2) + \sum_{i=1}^k \sum_{\substack{j=1 \\ i \neq j}}^k E(x_i x_j) \right] \quad (\text{A2.1})$$

where $E(x_i^2) = \sigma^2$ is the variance and $E(x_i x_j) = \sigma^2 \rho^{|i-j|}$ is the covariance of the underlying process. Therefore, the substitution of these expressions into Eq.(A2.1) gives

$$V(\bar{x}_k) = \frac{\sigma^2}{k^2} \left[k + \sum_{i=1}^k \sum_{\substack{j=1 \\ i \neq j}}^k \rho^{|i-j|} \right]$$

Or, by considering the symmetric property of the autocorrelation matrix the second term in the brackets on the r.h.s can be written in a more compact form which yields

$$V(\bar{x}_k) = \frac{\sigma^2}{k^2} \left[k + 2 \sum_{i=1}^{k-1} (k-i) \rho^i \right]$$

Or, more explicitly

$$V(\bar{x}_k) = \frac{\sigma^2}{k^2} \left[k + 2k \sum_{i=1}^{k-1} \rho^i - 2 \sum_{i=1}^{k-1} i \rho^i \right] \quad (A2.2)$$

where the first summation term in the brackets is a geometric progression whose sum is given as

$$\sum_{i=1}^{k-1} \rho^i = \rho \frac{1 - \rho^{k-1}}{1 - \rho} \quad (A2.3)$$

The second summation term is related to the derivative of first one as follows

$$\sum_{i=1}^{k-1} i \rho^i = \rho \frac{\delta}{\delta \rho} \left(\sum_{i=1}^{k-1} \rho^i \right)$$

By substituting Eq.(A2.3) into this last expression

$$\sum_{i=1}^{k-1} i \rho^i = \rho \frac{\delta}{\delta \rho} \left(\frac{\rho - \rho^k}{1 - \rho} \right)$$

and taking the derivative the following result is obtained

$$\sum_{i=1}^{k-1} i \rho^i = \rho \frac{(1 - k \rho^{k-1})(1 - \rho) + \rho - \rho^k}{(1 - \rho)^2} \quad (A2.4)$$

Finally, the substitution of Eq.(A2.3) and Eq.(A2.4) into Eq.(A2.2) results in

$$V(\bar{x}_k) = \frac{\sigma^2}{k^2} \left\{ k + \frac{2\rho}{(1-\rho)^2} \left[k(1-\rho) - (1-\rho^k) \right] \right\} \quad (A2.5)$$

The calculations required for $V(\bar{x}_n)$ are the same provided that k is replaced by n . Hence

$$V(\bar{x}_n) = \frac{\sigma^2}{n^2} \left\{ n + \frac{2\rho}{(1-\rho)^2} \left[n(1-\rho) - (1-\rho^n) \right] \right\} \quad (A2.6)$$

Now, the ^{CO}variance between $\bar{x}_k$ and $\bar{x}_n$ is given as

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{1}{nk} \cdot E \left(\sum_{i=1}^k x_i \right) \left(\sum_{j=1}^n x_j \right) \quad (A2.7)$$

which can be reduced to

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n E(x_i x_j) \quad (A2.8)$$

or by substituting $E(x_i x_j) = \sigma^2 \rho^{|i-j|}$

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{\sigma^2}{nk} \sum_{i=1}^k \sum_{j=1}^n \rho^{|i-j|}$$

If this last expression is considered in a rectangular matrix form, then the following explicit result can be written

$$nk \text{Cov}(\bar{x}_k, \bar{x}_n) = \begin{cases} \sigma^2 \sum_{i=1}^{n-k-1} (n-k-i) (\rho^{k+i} + \rho^{n-k-i}) + (n-k) \sigma^2 \sum_{i=0}^{2k-n} \rho^{n-k-i} & \text{for } (n-k) < k \\ \sigma^2 \sum_{i=1}^{k-1} (k-i) (\rho^{n-k+i} + \rho^{k-i}) + k \sigma^2 \sum_{i=0}^{n-2k} \rho^{k+i} & \text{for } (n-k) > k \end{cases} \quad (A2.9)$$

Appendix 2.2

The Adjusted Range for the ARIMA(1,0,1) Process

From the autocorrelation function of this process given by Eq.(1.21) and Eq.(1.22), the terms $V(\bar{x}_k)$, $\text{Cov}(\bar{x}_k, \bar{x}_n)$ and $V(\bar{x}_n)$ can be calculated as follows. As is obvious from these expressions the autocorrelation structure of the ARIMA(1,0,1) process is a function of two parameters namely ϕ and θ or equivalently ϕ and ρ . Therefore, the expected adjusted range will also be a function of ϕ and ρ .

Firstly from Eq.(A2.1) by substituting $E(x_i^2) = \sigma^2$ and $E(x_i x_j) = \sigma^2 \rho \phi^{|i-j|}$ for $i \neq j$, the following expression can be written,

$$V(\bar{x}_k) = \frac{\sigma^2}{k^2} \left\{ k + 2\rho \sum_{i=1}^{k-1} (n-i) \phi^{i-1} \right\}$$

By manipulating the necessary expansions as in the previous appendix (2.1)

$$V(\bar{x}_n) = \frac{\sigma^2}{n^2} \left\{ k + \frac{2\rho}{(1-\phi)^2} \left[k(1-\phi) - (1-\phi^k) \right] \right\} \quad (\text{A2.10})$$

In the same way, replacing k by n

$$V(\bar{x}_k) = \frac{\sigma^2}{k^2} \left\{ n + \frac{2\rho}{(1-\phi)^2} \left[n(1-\phi) - (1-\phi^n) \right] \right\} \quad (\text{A2.11})$$

and the covariance of Eq.(A2.8) can, in the same manner as in appendix (2.1) be obtained

$$\text{Cov}(\bar{x}_k, \bar{x}_n) = \frac{\sigma^2}{k^2} \left[k + \rho \sum_{i=1}^k \sum_{\substack{j=1 \\ i \neq j}}^n \phi^{|i-j|-1} \right]$$

or in an explicit form

$$nk\text{Cov}(\bar{x}_k, \bar{x}_n) = \begin{cases} \rho\sigma^2 \sum_{i=1}^{n-k-1} (n-k-i)(\varphi^{k+i-1} + \varphi^{n-k-i-1}) + \rho\sigma^2(n-k) \sum_{i=1}^{2k-n} \varphi^{n-k+i-1} & \text{for } (n-k) < k \\ \rho\sigma^2 \sum_{i=1}^{k-1} (k-i)(\varphi^{n-k+i-1} + \varphi^{k-i-1}) + k\rho\sigma^2 \sum_{i=1}^{n-2k} \varphi^{k+i-1} & \text{for } (n-k) > k \end{cases} \quad (\text{A2.12})$$

Appendix 2.3

Linear Independence of the Sample Mean and Sample Standard Deviation

The purpose of this appendix is to show that the mean and the standard deviation of a sequence of finite length are linearly independent under some assumptions and restrictions. The first restriction is that the underlying r.v. is standardized and normally distributed.

If two r.v.'s are independent then functions of them are also independent. As a result if $\bar{x}_n$ and S^2 are proved to be linearly independent, then in turn $\bar{x}_n$ and S are independent. The covariance between $\bar{x}_n$ and S^2 is given as

$$\text{Cov}(\bar{x}_n, S^2) = E(\bar{x}_n S^2)$$

where $\bar{x}_n$ is a r.v. defined by

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$$

and S^2 is also a r.v. defined by

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

The product of these last two equations yields

$$\bar{x}_n S^2 = \frac{1}{n^2} \left[\sum_{i=1}^n x_i \right] \left[\sum_{i=1}^n (x_i - \bar{x}_n)^2 \right]$$

the expansion of which leads to

$$\bar{x}_n S^2 = \frac{1}{n^2} \left[\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n x_i^2 \right) - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^3 \right] \quad (A2.13)$$

The terms in the brackets can be expanded as follows

$$\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i^2\right) = \sum_{i=1}^n x_i^3 + 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n x_i x_j^2 \quad (\text{A2.14})$$

and

$$\left(\sum_{i=1}^n x_i\right)^3 = \sum_{i=1}^n x_i^3 + 3 \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n x_i^2 x_j + 6 \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n \sum_{\substack{k=1 \\ i \neq j \neq k}}^n x_i x_j x_k \quad (\text{A2.15})$$

If the expected forms of Eq.(A2.14) and Eq.(A2.15) are substituted in Eq.(A2.13) and then the expectation of both sides is taken, it is seen that this expectation is zero, because all the terms on the r.h.s. of Eq.(A2.15) are of third order, and as long as the variable, x , is Gaussian the expectation of these odd order moments are all equal to zero, hence

$$E(\bar{x}_n S^2) = 0$$

from which it can be deduced that

$$E(\bar{x}_n S) = 0 \quad (\text{A2.16})$$

This is valid for the case of underlying Gaussian r.v's only.

Appendix 2.4

The Equivalence of Range and Rescaled Range

In the following, the limit, as n tends to infinity, of the constant which constitutes the relationship between the range and adjusted range as given in Eq.(2.95) will be evaluated. This limit operation can be written as

$$L = \lim_{n \rightarrow \infty} \frac{\sqrt{2n}}{(n-1)} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \quad (\text{A2.17})$$

In order to find this limit the Gamma functions, Γ , must be replaced by their approximations given by Stirling's formula (Feller, 1968), that is

$$\Gamma(\alpha) = \sqrt{2\pi} e^{\alpha} \alpha^{-1/2}$$

Hence, the ratio of the two Gamma functions in Eq.(A2.17) becomes,

$$\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} = \frac{\sqrt{2\pi} e^{n/2-1/2} (\frac{n+1}{2})^{n/2}}{\sqrt{2\pi} e^{n/2} (\frac{n}{2})^{n/2-1/2}}$$

Or

$$\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} = \left[\frac{n}{2e} \left(1 + \frac{1}{n}\right)^n \right]^{1/2} \quad (\text{A2.18})$$

Substitution of this expression into Eq.(A2.17) gives,

$$L = \lim_{n \rightarrow \infty} \left[\frac{\sqrt{2n}}{(n-1)} \cdot \frac{\sqrt{n}}{\sqrt{2e}} \left(1 + \frac{1}{n}\right)^{n/2} \right]$$

As n tends to infinity the term $\left(1 + \frac{1}{n}\right)^n$ converges to the exponential, e , hence

$$L = \lim_{n \rightarrow \infty} \left\{ \frac{1}{(1 - \frac{1}{n})} \cdot \frac{1}{\sqrt{e}} \sqrt{(1 + \frac{1}{n})^n} \right\} = 1$$

Therefore, from Eq.(2.95) it follows that asymptotically

$$E\left(\frac{R_n}{S}\right) = E(R_n)$$

which is also valid for the rescaled population range case.

Appendix 2.5

The Asymptotic Distribution of the Rescaled Population Range

The distribution sought is a mixture of the asymptotic distribution of population range given by Feller (1951) and the distribution of the standard deviation. These two r.v's are assumed to be independent from each other on the basis of appendix (2.3).

The asymptotic distribution of population range is given by Feller (1952) as

$$f(r';n)=(2/\pi)^{1/2} r'^{-1} L(r'/2n^{1/2}) \quad (A2.19)$$

where r' denotes the population range and

$$L(z) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2 z^2) \quad (A2.20)$$

The derivative of Eq.(A2.20) is

$$L'(z) = 8z \sum_{k=1}^{\infty} (-1)^{k-1} k^2 \exp(-2k^2 z^2) \quad (A2.21)$$

The rescaled population range, R_p/S , is denoted by a new r.v. y which is given as,

$$y = \frac{r'}{S}$$

or

$$y = \frac{r}{v} \quad (A2.22)$$

where $r = \sqrt{n} r'$ and $v = \sqrt{n} S$. The r.v. v is distributed according to

$$g(v) = 2av^{n-1}e^{v^2/2} \quad (A2.23)$$

where a is a constant given by

$$a = 1/2^{n/2} \Gamma(\frac{n}{2}) \quad (A2.24)$$

First of all it is necessary to find the distribution function of r by the method of transformation of r.v.'s given by Papoulis (1965). If a r.v. ϵ is transformed into a new r.v. as $y = a\epsilon$ then the pdf, $f_Y(y)$ of y is obtained from the pdf, $f_E(\epsilon)$ of ϵ as

$$f_Y(y) = \frac{1}{a} f_E\left(\frac{y}{a}\right)$$

Therefore, by virtue of $r = \sqrt{n} r'$ and Eq.(A2.19) the following can immediately be written:

$$f(r;n) = (2/\pi)^{1/2} \frac{1}{\sqrt{n}} \left(\frac{r}{\sqrt{n}}\right)^{-1} L'(r/2n)$$

or, in more compact form

$$f(r;n) = (2/\pi)^{1/2} r^{-1} L'(r/2n) \quad (A2.25)$$

Also, by considering the distribution of a ratio, in the case of independence of the numerator and the denominator, as given by Papoulis (1965), the distribution of y given in Eq.(A2.22) becomes

$$f(y) = \int_0^{\infty} v \cdot 2\alpha v^{n-1} \frac{e^{-v^2/2}}{(2/\pi)^{1/2}} (yv)^{-1} L'(yv/2n) dv$$

or by substituting Eq.(A2.21) into this last expression

$$f(y) = 8\alpha \left(\frac{2}{\pi}\right)^{1/2} \frac{1}{n} \sum_{k=1}^{\infty} (-1)^{k-1} k^2 \int_0^{\infty} v^n \exp\left(-\frac{n^2 + k^2 y^2}{2n^2} v^2\right) dv$$

Finally, the integral is evaluated to give

$$f(y) = \frac{8\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} n^n \sum_{k=1}^{\infty} (-1)^{k-1} k^2 (n^2 + k^2 y^2)^{-(n-1)/2} \quad (A2.26)$$

where $0 < y < +\infty$

This gives the asymptotic distribution of the rescaled population range, R_p/S . The expectation of r.v. y , can be found as follows,

$$E(y) = \int_0^{\infty} y f(y) dy = \frac{8\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} n^n \sum_{k=1}^{\infty} (-1)^{k-1} k^2 \int_0^{\infty} (n^2 + k^2 y^2)^{-(n-1)/2} y dy \quad (A2.27)$$

The integral to be evaluated is

$$I = \int_0^{\infty} y (n^2 + k^2 y^2)^{-(n-1)/2} dy$$

which, by the following change of variables becomes,

$$u^2 = n^2 + k^2 y^2 \quad ; \quad u du = k^2 y dy \quad \text{and} \quad y dy = \frac{u du}{k^2}$$

Hence,

$$I = \frac{1}{k^2} \int_n^{\infty} u^{-n} du = \frac{1}{k^2} \left[-\frac{1}{n-1} u^{-n+1} \right]_n^{\infty}$$

or,

$$I = \frac{n^{-n+1}}{k^2(n-1)}$$

By substituting this into Eq.(A2.27)

$$E(y) = \frac{8\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} \frac{n}{n-1} \sum_{k=1}^{\infty} (-1)^{k-1} \quad (A2.28)$$

The summation in Eq.(A2.28) is given as equal to 1/2 by Feller (1968), therefore,

$$E(y) = \frac{4\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} \frac{n}{n-1} \quad (A2.29)$$

For large values of n, this turns out to be

$$E(y) = 2\sqrt{\frac{2}{\pi} n} \quad (A2.30)$$

Appendix 3.1

Small Sample Expectation of the Estimate of Variance

The expression for the unbiased estimate of variance in the case of NIP is given in the main text by Eq.(3.7), that is,

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (A3.1)$$

By expanding the square term on the r.h.s. and then manipulating the summation operation the following result emerges

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2)$$

which by virtue of $n\bar{x} = \sum_{i=1}^n x_i$ takes the form

$$S^2 = \frac{1}{(n-1)} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]$$

By expanding the second term

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n x_i x_j$$

and taking the expectation of both sides

$$E(S^2) = \frac{1}{n} \sum_{i=1}^n E(x_i^2) - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n E(x_i x_j) \quad (A3.2)$$

Substitution of

$$E(x_i^2) = \sigma^2 - \mu^2$$

and

$$E(x_i x_j) = \text{Cov}(x_i, x_j) + \mu^2$$

into Eq.(A3.2) gives,

$$E(S^2) = \sigma^2 - \mu^2 - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n \text{Cov}(x_i, x_j) - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n \mu^2 \quad (\text{A3.3})$$

By considering that

$$\sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n \mu^2 = n(n-1)\mu^2$$

and

$$\sigma^2 \rho_{ii-jj} = \text{Cov}(x_i, x_j)$$

Eq.(A3.3) becomes,

$$E(S^2) = \sigma^2 \left[1 - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ i \neq j}}^n \rho_{ii-jj} \right] \quad (\text{A3.4})$$

This general formulae can be applied to any kind of normal stationary process provided that its autocorrelation function ρ_{ii-jj} is known for that process.

In the case of a symmetric autocorrelation function the above expression can be re-written in a more compact form as,

$$E(S^2) = \sigma^2 \left[1 - \frac{2}{n(n-1)} \sum_{i=1}^n (n-i) \rho_i \right] \quad (A3.5)$$

Eq.(A3.5) is valid whatever the underlying marginal distribution is, because no assumption as regards the distribution function of the r.v., x_i , is made.

Appendix 3.2

Small sample Expectation of the Lag-k Serial Correlation Coefficient for the ARIMA(1,0,1) Process

a) Open-Series Definition of $\hat{\rho}_k$

Kendall (1954) has given the result of the small sample expectation of $\hat{\rho}_k$ without a derivation and has applied it to various normal stationary processes. The general expression for $E(\hat{\rho}_k)$ is given by Kendall as

$$E(\hat{\rho}_k) = \frac{E(A)}{E(B)} - \frac{\text{Cov}(a,b)}{E^2(B)} + \frac{E(A)\text{Var}b}{E^3(B)} \quad (\text{A3.6})$$

the derivation of which has been given in section (3.7) of the main text.

When the open-series definition of the estimate given in Eq.(3.57) is considered then the following expressions are valid,

$$A = \frac{1}{(n-k)} \sum_{i=1}^{n-k} x_i x_{i+k} - \frac{1}{(n-k)^2} \left(\sum_{i=1}^{n-k} x_i \right) \left(\sum_{i=1}^{n-k} x_{i+k} \right) \quad (\text{A3.7})$$

$$B = \frac{1}{(n-k)} \sum_{i=1}^{n-k} x_i^2 - \frac{1}{(n-k)^2} \left(\sum_{i=1}^{n-k} x_i \right)^2 \quad (\text{A3.8})$$

and

$$C = \frac{1}{(n-k)} \sum_{i=1}^{n-k} x_{i+k}^2 - \frac{1}{(n-k)^2} \left(\sum_{i=1}^{n-k} x_{i+k} \right)^2 \quad (\text{A3.9})$$

As pointed out in section (3.7) it was assumed that $B = C$ and hence

two variates remain to be considered; let these be denoted by A and B. Only the expected values and the covariance of these two variables are required in the application of Eq.(A3.6). Let $n - k = v$ in the above expressions and by taking the expectation of both sides of Eq.(A3.7)

$$E(A) = \frac{1}{v} \sum_{i=1}^v E(x_i x_{i+k}) - \frac{1}{v^2} E \left[\left(\sum_{i=1}^v x_i \right) \left(\sum_{i=1}^v x_{i+k} \right) \right] \quad (A3.10)$$

The second expectation of the r.h.s. can explicitly be thought of like a two dimensional, $(v \times v)$, array as

$$E \left[\left(\sum_{i=1}^v x_i \right) \left(\sum_{i=1}^v x_{i+k} \right) \right] = \sigma^2 \begin{bmatrix} E(x_1 x_{k+1}) + E(x_1 x_{k+2}) + \dots + E(x_1 x_{k+v}) \\ +E(x_2 x_{k+1}) + E(x_2 x_{k+2}) + \dots + E(x_2 x_{k+v}) \\ \vdots \\ +E(x_v x_{k+1}) + E(x_v x_{k+2}) + \dots + E(x_v x_{k+v}) \end{bmatrix} \quad (A3.11)$$

where for a normal stationary process

$$E(x_i x_{i+k}) = \text{Cov}(x_i x_{i+k}) + E(x_i) E(x_{i+k}) = \text{Cov}(x_i x_{i+k}) + \mu^2 \quad (A3.12)$$

The substitution of Eq.(A3.12) into Eq.(A3.11) can be imagined as the summation of two arrays one consisting of the covariances and the other of the means only as, if Eq.(A3.11) is succinctly denoted by L then

$$L = \begin{bmatrix} \text{Cov}(x_1, x_{k+1}) + \dots + \text{Cov}(x_1, x_{k+v}) \\ + \text{Cov}(x_2, x_{k+1}) + \dots + \text{Cov}(x_2, x_{k+v}) \\ \vdots \\ + \text{Cov}(x_v, x_{k+1}) + \dots + \text{Cov}(x_v, x_{k+v}) \end{bmatrix} + \begin{bmatrix} \mu^2 + \mu^2 \dots + \mu^2 \\ + \mu^2 + \mu^2 \dots + \mu^2 \\ \vdots \\ + \mu^2 + \mu^2 \dots + \mu^2 \end{bmatrix} \quad (\text{A3.13})$$

where each one of the arrays has a dimension of $(v \times v)$. Furthermore, if the very well known expression given as

$$\rho_k = \frac{\text{Cov}(x_i, x_{i+k})}{\sigma^2}$$

or,

$$\text{Cov}(x_i, x_{i+k}) = \sigma^2 \rho_k \quad (\text{A3.14})$$

is substituted into the first array on the r.h.s. of Eq.(A3.13) this gives

$$L = \sigma^2 \begin{bmatrix} \rho_k + \rho_{k+1} \dots + \rho_{k+v-1} \\ + \rho_{k-1} + \rho_k \dots + \rho_{k+v-2} \\ \vdots \\ + \rho_{k+1-v} + \rho_{k+2-v} \dots + \rho_k \end{bmatrix} + \begin{bmatrix} \mu^2 + \mu^2 \dots + \mu^2 \\ + \mu^2 + \mu^2 \dots + \mu^2 \\ \vdots \\ + \mu^2 + \mu^2 \dots + \mu^2 \end{bmatrix} \quad (\text{A3.15})$$

On the main diagonal of the first array there are v times ρ_k elements, on the first off-diagonals there are $(v-1)$ times ρ_{k-1} and ρ_{k+1} elements, and in general on the i -th off-diagonals there exist $(v-i)$ times ρ_{k-i} and ρ_{k+i} elements; hence, the overall summation is

$$L = \sigma^2 \left[v\rho_k + (v-1)(\rho_{k+1} + \rho_{k-1}) + (v-2)(\rho_{k+2} + \rho_{k-2}) + \dots + (\rho_{k+v-1} + \rho_{k-v+1}) \right] + v^2 \mu^2$$

Or more succinctly,

$$L = \sigma^2 \left[v\rho_k + \sum_{i=1}^{v-1} (v-i)(\rho_{k+i} + \rho_{k-i}) \right] + v^2 \mu^2 \quad (\text{A3.16})$$

On the other hand, the first summation term, on the r.h.s. of Eq.(A3.10) can be re-written by making use of Eq.(A3.12) and Eq.(A3.14), which leads to

$$\sum_{i=1}^v E(x_i x_{i+k}) = v\sigma^2 \rho_k + v\mu^2 \quad (\text{A3.17})$$

Finally, the substitution of Eq.(A3.16) and Eq.(A3.17) into Eq.(A3.10) leads to

$$E(A) = \sigma^2 \rho_k + \mu^2 - \frac{\sigma^2}{v^2} \left[v\rho_k + \sum_{i=1}^{v-1} (v-i)(\rho_{k+i} + \rho_{k-i}) \right] - \mu^2$$

or

$$E(A) = \frac{\sigma^2}{v} \left[(v-1)\rho_k - \frac{1}{v} \sum_{i=1}^{v-1} (v-i)(\rho_{k+i} + \rho_{k-i}) \right] \quad (\text{A3.18})$$

One important remark is that Eq.(A3.18) is valid even in the case of a non-zero mean.

Similar algebraic calculations can be carried out in deriving the expectation of r.v. B given in Eq.(A3.8), which leads to

$$E(B) = 1 - \frac{1}{v} - \frac{2}{v^2} \sum_{i=1}^{v-1} (v-i) \rho_i \quad (A3.19)$$

Eq.(A3.18) and Eq.(A3.19) are general forms applicable to any type of normal stationary stochastic processes provided that the autocorrelation structure of the process is given. Here the ARIMA(1,0,1) process is considered whose autocorrelation function has been given earlier in Chapter 1 by Eq.(1.21) and Eq.(1.22). Hence,

$$\rho_i = \rho_1 \varphi^{i-1} \quad (A3.20)$$

By making the necessary substitution and after some tedious algebra,

$$\begin{aligned} E(A) = & \frac{1}{v} \left\{ v \rho_k - \frac{\rho}{v} \left\{ v \varphi^{k-1} \frac{(1-\varphi^v)}{(1-\varphi)} - \left[\frac{\varphi^k}{(1-\varphi)} \frac{1-\varphi^{k-1}}{(1-\varphi)} - (v-1) \varphi^{v-1} \right] \right\} \right. \\ & - \frac{\rho}{v} \left[\frac{v(1-\varphi^{k-1})}{(1-\varphi)} - \frac{(k-1)}{(1-\varphi)} - \frac{\varphi - \varphi^k}{(1-\varphi)} \right] - \frac{1}{v} (v-k) \\ & \left. - \frac{\rho}{v} (v-k) \frac{1-\varphi^{k-1}}{(1-\varphi)} - \frac{1}{(1-\varphi)} \left[\frac{1-\varphi^{v-k-1}}{(1-\varphi)} - (v-k-1) \varphi^{v-k-1} \right] \right\} \quad (A3.21) \end{aligned}$$

By ignoring terms higher in order than n^{-1} Eq.(A3.21) becomes

$$E(A) = \rho \varphi^{k-1} - \frac{2\rho}{v(1-\varphi)} - \frac{1}{v} \quad (A3.22)$$

Again in the case of the ARIMA(1,0,1) process after tedious algebra Eq.(A3.19), in general, becomes

$$E(B) = \frac{1}{v} \left\{ (v-1) - \frac{2}{v} \left[v\rho \frac{1-\varphi^{v-1}}{(1-\varphi)} - \frac{\rho}{(1-\varphi)} \frac{1-\varphi^{v-1}}{1-\varphi} - v\varphi^{v-1} + \varphi^{v-1} \right] \right\} \quad (A3.23)$$

or, by ignoring higher order terms Eq.(A3.23) reduces to

$$E(B) = 1 - \frac{1}{v} - \frac{2\rho}{v(1-\varphi)} \quad (A3.24)$$

An algebraic division of Eq.(A3.22) by Eq.(A3.24) leads to

$$\frac{E(A)}{E(B)} = \rho \varphi^{k-1} - \frac{(1-\rho \varphi^{k-1})(1+2\rho-\varphi)}{v(1-\varphi)} \quad (A3.25)$$

which is the first term in the r.h.s. of Eq.(A3.6). The other two terms necessary in the algebraic calculation of $E(\hat{\rho}_k)$ are given by Kendall (1954) as

$$\text{Cov}(a,b) = -\frac{2}{v} \sum_{l=-\infty}^{+\infty} \rho_l \rho_{l+k} \quad (A3.26)$$

and

$$\text{Var } b = \frac{2}{v} \sum_{i=-\infty}^{\infty} \rho_i^2 \quad (\text{A3.27})$$

Both of these two last expressions are valid for any type of normal stationary process; for the ARIMA(1,0,1) process they become,

$$\text{Cov}(a,b) = \frac{2}{v} \left[\rho \varphi^{k-1} + k \rho^2 \varphi^{k-2} + 2 \rho^2 \frac{\varphi^k}{(1-\varphi^2)} \right] \quad (\text{A3.28})$$

and

$$\text{Var } b = \frac{2}{v} \left[1 + \frac{2\rho}{(1-\varphi^2)} \right] \quad (\text{A3.29})$$

Finally, when all of the relevant equations are substituted into the r.h.s. of Eq.(A3.6) and the necessary algebra performed then the expected value of the serial correlation coefficient for small sample and the open-series definition of $\hat{\rho}_k$ given by Eq.(3.57) becomes

$$E(\hat{\rho}_k) = \rho \varphi^{k-1} - \frac{1}{v} \left[\frac{(1-\varphi+2\rho)(1+\varphi-2\rho^2\varphi^{k-1})}{(1-\varphi^2)} - \rho \varphi^{k-1} + 2k\rho^2\varphi^{k-2} \right] \quad (\text{A3.30})$$

However, hydrologists are mostly interested in the first order serial correlation coefficient; hence, by setting $k = 1$ Eq.(A3.30) reduces to

$$E(\hat{\rho}) = \rho - \frac{1}{v} \left[\frac{(1-\varphi^2) + \rho(1+2\varphi-4\rho) + \varphi\rho(2\rho-\varphi)}{(1-\varphi^2)} + \rho \right] \quad (\text{A3.31})$$

The verification of Eq.(A3.31) against the corresponding formula given for the lag-one Markov process by Kendall (1954) can be achieved simply by substituting $\psi = \rho$.

b) Circular and Maximum Likelihood Definitions of $\hat{\rho}_k$.

Here, the small sample expectation of $\hat{\rho}_k$ will be presented for the circular definition given in Eq.(3.57) and the maximum likelihood definition given in Eq.(3.58) by pursuing a similar sort of algebraic approach as taken in the first part of this appendix.

The only difference from the open-series approach is that here A is given as

$$A = \frac{1}{n} \sum_{i=1}^n x_i x_{i+k} - \frac{1}{n^2} \left(\sum_{i=1}^n x_i \right)^2 \quad (A3.32)$$

which is quite distinct from Eq.(A3.7). When the necessary operations are executed as previously, then the expected value of A, in general, becomes,

$$E(A) = \frac{1}{n} \left[(n-k) \rho_k + k \rho_{n-k} \right] - \frac{1}{n^2} \left[n + 2 \sum_{i=1}^{n-1} (n-i) \rho_i \right] \quad (A3.33)$$

and in particular for the ARIMA(1,0,1) process Eq.(A3.33) turns out to be

$$E(A) = \frac{1}{n} \left[(n-k) \rho \psi^{k-1} + k \rho \psi^{n-k-1} \right] \quad (A3.34)$$

$$- \frac{1}{n^2} \left[n + 2 \rho \frac{1 - \psi^{n-1}}{1 - \psi} - \frac{\rho}{(1 - \psi)} \left[\frac{1 - \psi^{v-1}}{(1 - \psi)} - (v-1) \psi^{v-1} \right] \right]$$

By ignoring the higher terms Eq.(A3.34) can be written in a more compact form as,

$$E(A) = \frac{1}{n} \left[(n-k) \rho \varphi^{k-1} + k \varphi^{n-k-1} \right] - \frac{1}{n} - \frac{2\rho}{n(1-\varphi)} \quad (A3.35)$$

The same expressions, given in the first part of this thesis, for $E(B)$, $\text{Var}(b)$ and $\text{Cov}(a,b)$ are applicable also in this part; hence the substitution of Eq.(A3.35) and other relevant equations derived in the first part, into Eq.(A3.6) yields

$$E(\hat{\rho}_k) = \rho \varphi^{k-1} - \frac{1}{n} \left[\frac{(1-\varphi+2\rho)(1-\rho \varphi^{k-1})}{(1-\varphi)} + k \rho (\varphi+2\rho) \varphi^{k-2} - k \rho \varphi^{n-k-1} + \frac{4\rho^2 \varphi^{k-1}}{(1-\varphi^2)} (\varphi-\rho) \right] \quad (A3.36)$$

For the case of $k = 1$ Eq.(A3.36) becomes

$$E(\hat{\rho}) = \rho - \frac{1}{n} \left[\frac{(1-\varphi+2\rho)(1-\rho)}{(1-\varphi)} + \frac{4\rho^2(\varphi-\rho)}{(1-\varphi^2)} + 3\rho \right] \quad (A3.37)$$

Appendix 5.1

The likelihood Function of the ARIMA(1,0,1) Process

The purpose of this appendix is to write the general expression of the likelihood function given by Eq.(5.28) in a more compact form. The exponent in Eq.(5.28) can be written as

$$P = M_1 + M_2 + M_3 \quad (A5.1)$$

where M_1 , M_2 and M_3 are

$$M_1 = \frac{(x_0 - \mu)^2}{2\sigma^2} + \frac{\eta_0^2}{2\sigma_\eta^2} \quad (A5.2)$$

$$M_2 = \frac{1}{2\sigma^2\sigma_\eta^2} \quad (A5.3)$$

$$\begin{aligned} M_3 = & \sum_{i=1}^n (x_i - \mu)^2 + 2(\theta - \varphi) \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^j (x_i - \mu)(x_{i+j} - \mu) \\ & + (\theta - \varphi)^2 \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{j-i} (x_i - \mu) \right]^2 - 2A \sum_{i=1}^n \theta^{i-1} (x_i - \mu) \\ & - 2A(\theta - \varphi) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} (x_{i-1} - \mu) + A^2 \frac{1 - \theta^{2n}}{1 - \theta^2} \end{aligned} \quad (A5.4)$$

respectively. Eq.(A5.1) can be shortened by introducing new variables as

follows. Firstly, with the new variables defined in Eq.(A5.1) the likelihood function becomes

$$l(x_1, x_2, x_3, \dots, x_n | \eta, x_0, \mu, \sigma, \varphi, \theta) \propto \frac{1}{\sigma^{n+2} \sigma_\eta^{n+1}} \exp(-P) \quad (A5.5)$$

Now each one of the terms on the r.h.s. of Eq.(A5.4) will be considered one by one. Hence,

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \quad (A5.6)$$

where only μ is a r.v. and the rest is constant for a given sample, therefore the following constants are introduced as

$$a_1 = \sum_{i=1}^n x_i^2$$

$$b_1 = -2 \sum_{i=1}^n x_i$$

$$c_1 = n$$

Substituting these in Eq.(A5.6) gives

$$\sum_{i=1}^n (x_i - \mu)^2 = a_1 + b_1 \mu + c_1 \mu^2 \quad (A5.7)$$

Just in the same way the next term in Eq.(A5.4) can be expanded as

$$\begin{aligned}
 \sum_{j=1}^n \theta^{j-1} \sum_{i=1}^{n-j} (x_i - \mu)(x_{i+j} - \mu) &= \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^{n-j} x_i x_{i+j} \\
 - \mu \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^{n-j} (x_i + x_{i+j}) &+ \mu^2 \sum_{j=1}^{n-1} (n-j) \theta^{j-1} \quad (A5.8)
 \end{aligned}$$

By introducing another set of constants as

$$\begin{aligned}
 a_2 &= \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^{n-j} x_i x_{i+j} \\
 b_2 &= - \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^{n-j} (x_i + x_{i+j}) \\
 c_2 &= \sum_{j=1}^{n-1} (n-j) \theta^{j-1}
 \end{aligned}$$

and substituting them in Eq.(A5.8) the result is

$$\sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^{n-j} (x_i - \mu)(x_{i+j} - \mu) = a_2 + b_2 \mu + c_2 \mu^2 \quad (A5.9)$$

The summation in the third term in Eq.(A5.4) can be expanded as

$$\begin{aligned}
 \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{j-i} (x_i - \mu) \right]^2 &= \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{2(j-i)} (x_i - \mu) + 2 \sum_{k=1}^j \sum_{\substack{l=1 \\ k \neq l}}^j \theta^{2j-k-l} (x_k - \mu)(x_l - \mu) \right] \\
 &= \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{2(j-i)} x_i^2 + 2 \sum_{k=1}^j \sum_{\substack{l=1 \\ k \neq l}}^j \theta^{2j-k-l} x_k x_l \right] \\
 &\quad - 2\mu \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{2(j-i)} x_i + \sum_{k=1}^j \sum_{\substack{l=1 \\ k \neq l}}^j \theta^{2j-k-l} (x_k + x_l) \right] \\
 &\quad + \mu^2 \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{2(j-i)} + 2 \sum_{k=1}^j \sum_{\substack{l=1 \\ k \neq l}}^j \theta^{2j-k-l} \right] \tag{A5.10}
 \end{aligned}$$

Hence the constants in this case become

$$\begin{aligned}
 a_3 &= \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{2(j-i)} x_i^2 + 2 \sum_{k=1}^j \sum_{\substack{l=1 \\ k \neq l}}^j \theta^{2j-k-l} x_k x_l \right] \\
 b_3 &= -2 \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{2(j-i)} x_i + \sum_{k=1}^j \sum_{\substack{l=1 \\ k \neq l}}^j \theta^{2j-k-l} (x_k + x_l) \right] \\
 c_3 &= \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{2(j-i)} + 2 \sum_{k=1}^j \sum_{\substack{l=1 \\ k \neq l}}^j \theta^{2j-k-l} \right]
 \end{aligned}$$

By substituting these last three expressions in Eq.(A5.10)

$$\sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{j-i} (x_i - \mu) \right]^2 = a_3 + b_3 \mu + c_3 \mu^2 \tag{A5.11}$$

The fourth term in the exponent of Eq.(A5.4) is

$$\begin{aligned}
A \sum_{i=1}^n \theta^{i-1} (x_i - \mu) &= \left[\varphi(x_0 - \mu) - \theta \sigma \eta_0 \right] \left[\sum_{i=1}^n \theta^{i-1} x_i - n\mu \right] \\
&= n\varphi\mu^2 + (n\theta\sigma\eta_0 - n\varphi x_0 - \varphi \cdot \sum_{i=1}^n \theta^{i-1} x_i) \mu \\
&\quad + (\varphi x_0 \sum_{i=1}^n \theta^{i-1} x_i - \theta\sigma\eta_0 \sum_{i=1}^n \theta^{i-1} x_i)
\end{aligned} \tag{A5.12}$$

The constants are denoted by

$$\begin{aligned}
a_4 &= \varphi x_0 \sum_{i=1}^n \theta^{i-1} x_i - \theta\sigma\eta_0 \sum_{i=1}^n \theta^{i-1} x_i \\
b_4 &= n\theta\sigma\eta_0 - n\varphi x_0 - \varphi \sum_{i=1}^n \theta^{i-1} x_i \\
c_4 &= n\varphi
\end{aligned}$$

Therefore,

$$A \sum_{i=1}^n \theta^{i-1} (x_i - \mu) = a_4 + b_4 \mu + c_4 \mu^2 \tag{A5.13}$$

The fifth term in Eq.(A5.4) is given by

$$A \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} (x_{i-1} - \mu) = \left[\varphi(x_0 - \mu) - \theta\sigma\eta_0 \right] \left[\sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} x_{i-1} \right]$$

$$- \mu \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} \Big]$$

$$= (\varphi x_0 - \theta \sigma \eta_0) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} x_{i-1}$$

$$+ (\theta \sigma \eta_0 - \varphi x_0) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} - \varphi \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} x_{i-1}$$

$$+ (\varphi \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i}) \mu^2 \quad (A5.14)$$

where the constants are given by

$$a_5 = (\varphi x_0 - \theta \sigma \eta_0) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} x_{i-1}$$

$$b_5 = (\theta \sigma \eta_0 - \varphi x_0) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} - \varphi \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} x_{i-1}$$

$$c_5 = \varphi \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i}$$

Therefore, from Eq.(A5.14)

$$A \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} (x_{i-1} - \mu) = a_5 + b_5 \mu + c_5 \mu^2 \quad (A5.15)$$

And the final term in Eq.(A5.4) is given by

$$\begin{aligned} \frac{1 - \theta^{2n}}{1 - \theta^2} A^2 &= \frac{1 - \theta^{2n}}{1 - \theta^2} \left[\varphi(x_0 - \mu) - \theta \sigma \eta_0 \right]^2 \\ &= \left[(\varphi^2 x_0^2 - 2\theta \sigma \eta_0 \varphi x_0 + \theta^2 \sigma^2 \eta_0^2) + 2\varphi(\theta \sigma \eta_0 - \varphi x_0) + \varphi^2 \mu^2 \right] \frac{1 - \theta^{2n}}{1 - \theta^2} \end{aligned}$$

Therefore, the constants are given as

$$a_6 = (\varphi^2 x_0^2 - 2\theta \sigma \eta_0 \varphi x_0 + \theta^2 \sigma^2 \eta_0^2) \frac{1 - \theta^{2n}}{1 - \theta^2}$$

$$b_6 = 2\varphi(\theta \sigma \eta_0 - x_0 \varphi) \frac{1 - \theta^{2n}}{1 - \theta^2}$$

$$c_6 = \varphi^2 \frac{1 - \theta^{2n}}{1 - \theta^2}$$

Hence

$$A \frac{1 - \theta^{2n}}{1 - \theta^2} = a_6 + b_6 \mu + c_6 \mu^2 \quad (A5.16)$$

and the first term of Eq.(A5.1) can be simplified as

$$\frac{(x_0 - \mu)^2}{2} = \frac{x_0^2}{2} - x_0 \mu + \frac{\mu^2}{2}$$

The constants are

$$a_7 = x_0^2 / 2$$

$$b_7 = -x_0$$

$$c_7 = 1/2$$

Therefore Eq.(A5.2) becomes

$$\frac{(x - \mu)^2}{2} = a_7 + b_7 \mu + c_7 \mu^2 \quad (A5.16)$$

After all the above manipulations and simplifications Eq.(A5.1) can be re-written in a more compact form as

$$\begin{aligned} P = & \frac{a + b_7 \mu + c_7 \mu^2}{\sigma^2} + \frac{\eta_0^2}{2\sigma_\eta^2} + \frac{1}{2\sigma^2\sigma_\eta^2} \left[a_1 + b_1 \mu + c_1 \mu^2 + \right. \\ & + 2(\theta - \varphi)(a_2 + b_2 \mu + c_2 \mu^2) + (\theta - \varphi)^2(a_3 + b_3 \mu + c_3 \mu^2) - 2(a_4 + b_4 \mu + c_4 \mu^2) \\ & \left. - 2(\theta - \varphi)(a_5 + b_5 \mu + c_5 \mu^2) + (a_6 + b_6 \mu + c_6 \mu^2) \right] \quad (A5.17) \end{aligned}$$

Appendix 5.2

Three-Parameter Process - Type-1

The general expression for the posterior pdf for this case has already been represented in Section (5.5) by Eq.(5.34) which can be written as

$$P(\mu, \sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+2} \sigma_\eta^{n+1}} \exp(-P) \quad (A5.18)$$

where P has been given by Eq.(A5.17).

The purpose of this appendix is to integrate Eq.(A5.18) with respect to μ (or to v) and then with respect to σ , so that the posterior pdf of φ , will be attained. For the former integration everything save μ is assumed constant therefore Eq.(A5.17) can be brought into a quadratic form which is given as

$$\begin{aligned} P = & \frac{1}{2\sigma^2 \sigma_\eta^2} \left[\sigma_\eta^2 c_7 + c_1 + 2(\theta - \varphi)c_2 + (\theta - \varphi)^2 c_3 - 2c_4 - 2(\theta - \varphi)c_5 + c_6 \right] \mu^2 \\ & + \left[\sigma_\eta^2 b_7 + b_1 + (\theta - \varphi)b_2 + (\theta - \varphi)^2 b_3 - 2b_4 - 2(\theta - \varphi)b_5 + b_6 \right] \mu \\ & + \left[\sigma_\eta^2 a_7 + \sigma_\eta^2 a_0 + a_1 + (\theta - \varphi)a_2 + (\theta - \varphi)^2 a_3 - 2a_4 - 2(\theta - \varphi)a_5 + a_6 \right] \end{aligned} \quad (A5.19)$$

By introducing a new set of constants as

$$C_1 = c_1 + 2(\theta - \varphi)c_2 + (\theta - \varphi)^2 c_3 - 2c_4 - 2(\theta - \varphi)c_5 + c_6 + \sigma_\eta^2 c_7$$

$$B_1 = b_1 + 2(\theta - \varphi)b_2 + (\theta - \varphi)^2 b_3 - 2b_4 - 2(\theta - \varphi)b_5 + b_6 + \sigma^2 b_7$$

$$A_1 = a_1 + 2(\theta - \varphi)a_2 + (\theta - \varphi)^2 a_3 - 2a_4 - 2(\theta - \varphi)a_5 + a_6 + \sigma^2 a_7 + \sigma^2 \eta_0$$

As a result Eq.(A5.19) can be re-written as

$$P = \frac{1}{2\sigma^2\sigma_\eta^2} (A_1 + B_1\mu + C_1\mu^2) \quad (A5.20)$$

where A_1 , B_1 and C_1 are the functions of σ , φ , θ , x_0 , η_0 and the given data. Eq.(A5.20) can be written in a more convenient form as

$$P = \frac{C_1}{2\sigma^2\sigma_\eta^2} \left[(\mu - E_1)^2 + D_1 \right] \quad (A5.21)$$

where

$$E_1 = - \frac{B_1}{2C_1}$$

$$D = \frac{A_1}{B_1} - \left(\frac{B_1}{2C_1} \right)^2$$

Therefore, the posterior pdf of Eq.(A5.18) can be written as

$$P(\mu, \sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+2}\sigma_\eta^{n+1}} \exp \left\{ - \frac{C_1}{2\sigma^2\sigma_\eta^2} \left[(\mu - E_1)^2 + D_1 \right] \right\} \quad (A5.22)$$

or, in terms of the inverse of the coefficient of variation, $v = \mu / \sigma$,
Eq.(A5.22) can be expressed as

$$P(v, \sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+2} \sigma_\eta^{n+1}} \exp \left\{ -\frac{C_1}{2 \sigma^2 \sigma_\eta^2} \left[(v \sigma - E_1)^2 + D_1 \right] \right\} \quad (A5.23)$$

where v varies between $-\infty$ to $+\infty$. Now, Eq.(A5.23) must be integrated with respect to v , which gives

$$P(\sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+2} \sigma_\eta^{n+1}} \exp \left\{ -\frac{C_1 D_1}{2 \sigma^2 \sigma_\eta^2} \right\} \int_{-\infty}^{+\infty} \exp \left\{ -\frac{C_1 (\sigma v - E_1)^2}{2 \sigma^2 \sigma_\eta^2} \right\} dv \quad (A5.24)$$

A change of variable is introduced as

$$z = \frac{(\sigma v - E_1)}{\sqrt{2} \sigma \sigma_\eta} \sqrt{C_1}$$

from which the following results

$$v = \frac{\sqrt{2} \sigma_\eta}{\sqrt{C_1}} z + \frac{E_1}{\sigma}$$

and

$$dv = \frac{\sqrt{2}}{\sqrt{C_1}} \sigma_\eta dz$$

By substituting this new variable in Eq.(A5.24) it becomes

$$P(\sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta) \sigma_\eta}{\sqrt{C_1} \sigma^{n+2} \sigma_\eta^{n+1}} \exp \left\{ -\frac{C_1 D_1}{2 \sigma^2 \sigma_\eta^2} \right\} \int_{-\infty}^{+\infty} e^{-z^2} dz$$

or

$$P(\sigma, \varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{\sqrt{C_1} \sigma^{n+2} \sigma_\eta^{n+1}} \exp \left\{ -\frac{C_1 D_1}{2 \sigma^2 \sigma_\eta^2} \right\} \quad (A5.25)$$

The posterior pdf of φ can be arrived at after integrating Eq.(A5.25) with respect to σ . Therefore

$$P(\varphi|\eta_0, x_0, \theta, D) \propto \frac{P(\varphi|\theta)}{\sqrt{C_1} \sigma_\eta^{n+1}} \int_0^{+\infty} \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{C_1 D_1}{2 \sigma^2 \sigma_\eta^2}\right\} d\sigma \quad (\text{A5.26})$$

where C_1 and D_1 are independent of σ . By considering a change of variable as

$$z_1 = \frac{1}{\sigma^2} \quad ; \quad \sigma = z_1^{-1/2} \quad ; \quad d\sigma = -\frac{1}{2} z_1^{-3/2} dz_1$$

and by incorporating these in Eq.(A5.26)

$$P(\varphi|\eta_0, x_0, \theta, D) \propto \frac{P(\varphi|\theta)}{\sqrt{C_1} \sigma_\eta^{n+1}} \int_0^{+\infty} z_1^{(n-1)/2} \exp\left\{-\frac{C_1 D_1}{2 \sigma_\eta^2} z_1\right\} dz_1 \quad (\text{A5.27})$$

By using the following standard integration

$$\int_0^{\infty} x^n \exp(-ax) dx = \frac{\Gamma(n+1)}{a^{n+1}} \quad (\text{A5.28})$$

Eq.(A5.27) becomes

$$P(\varphi|\eta_0, x_0, \theta, D) \propto \frac{P(\varphi|\theta)}{\sqrt{C_1} (C_1 D_1)^{(n+1)/2}} \quad (\text{A5.29})$$

where ,

$$F_1 = C_1 D_1 = A_1 - \frac{B_1^2}{4C_1} \quad (\text{A5.30})$$

Therefore, finally

$$P(\varphi | \eta_0, x_0, \theta, D) \propto \frac{P(\varphi | \theta)}{C^{1/2}} F_1^{-(n+1)/2} \quad (A5.31)$$

Eq.(A5.31) can be reduced by substituting $\theta = 0$, $\eta_0 = 0$ and $\varphi = \rho$ in all the previous equations and then Eq.(A5.31) takes the following form for the Markov process

$$P(\rho | x_0, D) \propto \frac{P(\rho)}{C^{1/2}} \left[A_1 - \frac{B_1^2}{4C_1} \right]^{-(n+1)/2} \quad (A5.32)$$

where

$$C_1 = n(1 - \rho)^2 + (1 - \rho^2) \quad (A5.33)$$

$$B_1 = -2(1 - \rho) \left[x_0(1 + \rho) + \sum_{i=1}^n (x_i - \rho x_{i-1}) \right] \quad (A5.34)$$

$$A_1 = (1 - \rho^2) x_0^2 + \sum_{i=1}^n (x_i - \rho x_{i-1})^2 \quad (A5.35)$$

The derivation of Eq.(A5.32) has first been given by Thornber (1967).

However, apart from the calculations given here for the posterior pdf of φ a similar sort of algebraic simplification could be achieved for finding the posterior pdf of μ or σ , by integrating Eq.(A5.18) with respect to the remaining parameters.

Appendix 5.3

Two-Parameter Process - Type-2

a) Known mean

If one of the three parameters, namely, μ , σ , φ were known, then, there remains two unknown parameters which can be treated by Bayes theorem to find their posterior pdf. There are three alternative ways of picking up two parameters among three values, only two of these alternatives will be considered here; one is in the case of known mean and the other one is the known standard deviation. In the case of known mean the population mean μ is considered to be fixed while σ and φ are assumed to be r.v's. The behaviour of the posterior pdf for φ is then examined. The joint posterior pdf of σ and φ is easily seen from Eq.(A5.18) to be

$$P(\sigma, \varphi | \eta_0, x_0, \mu, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+2} \sigma^{n+1}} \exp(-P) \quad (\text{A5.36})$$

A new compact form of P can be developed by introducing a set of constants which include μ as a fixed value. These constants are as follows

$$\alpha_1 = (x - \mu)^2$$

$$\alpha_2 = \sum_{i=2}^n (x_i - \mu)^2$$

$$\alpha_3 = 2(\theta - \varphi) \sum_{j=1}^{n-1} \theta^{j-1} \sum_{i=1}^j (x_i - \mu)(x_{i+j} - \mu)$$

$$\alpha_4 = (\theta - \varphi)^2 \sum_{j=1}^{n-1} \left[\sum_{i=1}^j \theta^{j-i} (x_i - \mu) \right]^2$$

$$\alpha_5 = -2A \sum_{i=1}^n \theta^{i-1} (x_i - \mu)$$

$$\alpha_6 = -2A(\theta - \varphi) \sum_{j=2}^n \theta^{j-1} \sum_{i=2}^j \theta^{j-i} (x_{i-1} - \mu)$$

$$\alpha_7 = A^2 \frac{1 - \theta^{2n}}{1 - \theta^2}$$

and as a result Eq.(A5.17) can be re-written as

$$P = \frac{1}{2\sigma^2\sigma_\eta^2} (\sigma_\eta^2\alpha_1 + \sigma\eta_0 + \sum_{i=2}^7 \alpha_i) \quad (A5.37)$$

By inserting this expression in Eq.(A5.36) the following result is obtained

$$P(\sigma, \varphi | \eta_0, x_0, \mu, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+2}\sigma_\eta^{n+1}} \exp - \frac{\sigma_\eta^2\alpha_1 + \sigma\eta_0 + \sum_{i=2}^7 \alpha_i}{2\sigma^2\sigma_\eta^2} \quad (A5.38)$$

The simple case where η_0 is assumed to be equal to zero is now considered.

Hence Eq.(A5.38) reduces to

$$P(\sigma, \varphi | x_0, \mu, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+2}\sigma_\eta^{n+1}} \exp - \frac{\sigma_\eta^2\alpha_1 + \sum_{i=2}^7 \alpha_i}{2\sigma^2\sigma_\eta^2} \quad (A5.39)$$

or by integrating with respect to σ

$$P(\varphi | x_0, \mu, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+1}} \int_0^\infty \frac{1}{\sigma^{n+2}} \exp - \frac{\sigma_\eta^2\alpha_1 + \sum_{i=2}^7 \alpha_i}{2\sigma^2\sigma_\eta^2} \quad (A5.40)$$

With the introduction of a new constant as

$$B = \frac{\sigma_{\eta}^2 \alpha_1 + \sum_{i=2}^2 \alpha_i}{2}$$

Eq.(A5.40) becomes

$$P(\varphi | x, \mu, \theta, D) \propto \frac{P(\varphi | \theta)}{\sigma^{n+1}} \int_0^{\infty} \frac{1}{\sigma^{n+2}} \exp - \frac{B}{\sigma^2 \sigma^2} d\sigma \quad (A5.41)$$

Now, by changing variables as

$$z = \frac{B}{\sigma^2 \sigma_{\eta}^2} \longrightarrow \sigma = \frac{B^{1/2}}{\sigma_{\eta}} z^{-1/2}$$

$$d\sigma = - \frac{B^{-1/2}}{2 \sigma_{\eta}} z^{-3/2}$$

Again by the use of the standard integration given in Eq.(A5.28)

$$P(\varphi | x, \mu, \theta, D) \propto \frac{P(\varphi | \theta)}{B^{(n+1)/2}} \quad (A5.42)$$

This is the general form for the posterior pdf of the ARIMA(1,0,1) process when the necessary substitutions are done the corresponding form of Eq.(A5.42) for the lag-one Markov process can be obtained.

b) Known Standard deviation

This is the case of fixed standard deviation where the mean μ and φ are treated as r.v's. Therefore, the integration of Eq.(A5.22) with respect to μ will lead to the posterior pdf of φ given that σ is fixed. So the result can straightforwardly be obtained from Eq.(A5.25) as

$$P(\varphi | \eta_0, x_0, \sigma, \theta, D) \propto \frac{P(\varphi | \theta)}{C^{1/2} \sigma^{n+2} \sigma_{\eta}^{n+1}} \exp \left\{ -\frac{C_1 D_1}{2 \sigma^2 \sigma_{\eta}^2} \right\} \quad (A5.43)$$

where all of the necessary constants appearing in Eq.(A5.43) have already been given in appendix (5.1)

REFERENCES

- AITCHISON, J., BROWN, J.A.C. (1957), The Lognormal Distribution With Special Reference to Its Use in Economics, Cambridge, University Press, pp 176.
- ANDERSON, R.L. (1942), Distribution of the Serial Correlation Coefficient, Ann. Math. Stat., 13, 1-13.
- ANIS, A.A. (1955), The Variance of the Maximum of Partial Sums of a Finite Number of Independent Normal Variates, Biometrika, 42, 96-101.
- ANIS A.A. (1956), On the Moments of Maximum of Partial Sums of a Finite Number of Independent Normal Variates, Biometrika, 43, 79-84.
- ANIS, A.A., LLOYD, E.H. (1953), On the Range of Partial Sums of a Finite Number of Independent Normal Variates, Biometrika, 40, 35-42.
- BARLEY, G.V., HAMMERSLEY, J.M. (1946), The Effective Number of Independent Observations in an Autocorrelated Series, J. Roy. Stat. Soc., 8(1-B), 184-197.
- BARNES, F.B. (1954), Storage Required for a City Water Supply, J. Instn. Engrs. Austral., 26, 198-203.
- BENJAMIN, J.R., CORNELL, C.A. (1970), Probability Statistics and Decision for Civil Engineers, New York, N.Y., McGraw-Hill Book Co., pp 452.
- BENSON, M.A., MATALAS, N.C. (1967), Synthetic Hydrology Based on Regional Statistical Parameters, Water Resour. Res., 3(4), 931-935.
- BOX, G.E.P., JENKINS, G.M. (1968), Some Recent Advances in Forecasting and Control, App. Stat., 17(2), 91-109.
- BOX, G.E.P., JENKINS, G.M. (1970), Time Series Analysis Forecasting and Control, Holden Day, San Francisco, pp 498.
- BROPHY, J.J. (1970), Statistical Randomness and $1/f$ Noises, Navel Research Reviews, 8-21.
- CARLSON, R.F., MACCORMICK, A.J.A., WATTS, D.G. (1970), Application of Linear Random Models to Four Annual Streamflow Series, Water Resour. Res., 6(4), 1070-1078.
- CONOVER, W.J. (1971), Introduction to Bayesian Methods Using the Thomas-Fiering Model Water Resour. Res., 7(2), 406-409.

- CRAMER, H. (1946), Mathematical Methods of Statistics,
Princeton University Press, Princeton pp 575.
- DITLEVSEN, O. (1971), Extremes and First Passage Times With Applications
in Civil Engineering,
Ph. D. Thesis, Copenhagen, Tech. Univ. of Denmark
- FELLER, W. (1951), The Asymptotic Distribution of the Range of Sums of
Independent Random Variables,
Ann. Math. Stat., 22, 427-432.
- FELLER, W. (1968), An Introduction to Probability Theory and Its
Application,
John Wiley and Sons, Inc., pp 509.
- FIERING, M.B. (1967), Streamflow Synthesis,
London, The McMillan Press Ltd., pp 139.
- GARCIA, L.E., DAWDY, D.R., MEJIA, J.M. (1972), Long Memory Monthly
Streamflow Simulation by a Broken Line Model,
Water Resour. Res., 8(4), 1100-1105.
- HAZEN, A. (1914), Storage to be Provided in Impounding Reservoirs for
Municipal Water Supply,
Trans. Am. Soc. Civ. Engrs., 77, 1539-1669.
- HOEL, P.G. (1947), Introduction to Mathematical Statistics
New York John Wiley Co. pp 427.
- HURST, H.E. (1951) Long Term Storage Capacity of Reservoirs,
Trans. Am. Soc. Civ. Engrs., 116, 770-808.
- HURST, H.E., (1956), Method of Using Long Term Storage in Reservoirs,
Proc. Instn. Civ. Engrs., 1, 519-542.
- HURST, H.E. (1957), A Suggested Statistical Model of Some Time Series
Which Occur in Nature,
Nature, 180, p.494.
- JEFFREYS, H. (1961), Theory of Probability,
Oxford, Clarendon pp 459.
- JENKINS, G.M., WATTS, D.G. (1968), Spectral Analysis and Its Application,
San Francisco, Holden Day Inc., pp 525.
- KENDALL, M.G. (1954), Note on the Bias in the Estimation of Autocorrelation,
Biometrika, 42, 403-404.
- KIRBY, W. (1972), Computer-Oriented Wilson Hilferty Transformation that
Preserves the First Three Moments and Lower Bound of the
Icarson Type 3 Distribution,
Water Resour. Res., 8(5), 1251-1254.
- LANGLEIR, W.B. (1956), Discussion of "Methods of Using Long Term Storage
in Reservoirs", by H.E. Hurst,
Proc. Instn. Civ. Engrs., 1, 281-339.

- LENTON, R.L., RODRIGUEZ-ITURBE, I., SCHAAKE, J.C. (1973), A Bayesian Approach to Autocorrelation Estimation in Hydrologic Autoregressive Models, Report No. 163, Ralph M. Parson Lab., MIT, pp 121.
- LEVY, P. (1965), Processus Stochastique et Mouvement Brownien, Paris, Gauthier Villards, Imprimeur Editeur, pp 438.
- MASS ET AL, (1962), Design of Water Resources Systems, Cambridge, Mass., Harvard Univ. Press, pp. 660.
- MANDELBROT, B.B. (1965), Une Class de Processus Homothetiques a Soi; Application a la Loi Climatologique de H.E. Hurst, Compt. Rend. Acad. sci. Paris, 260, 3274-3277.
- MANDELBROT, B.B. (1971), A Fast Fractional Gaussian Noise Generator, Water Resour. Res., 7(3), 543-553.
- MANDELBROT, B.B. (1972), Statistical Methodology for Non-Periodic Cycles: from the Covariance to R/S Analysis, Annals of Economic and Social Measurement, 1(3), 259-285.
- MANDELBROT, B.B., VAN NESS, J.W. (1968), Fractional Brownian Motions, Fractional Noises and Applications, Soc. Ind. Appl. Math. Rev., 10(4), 422-437.
- MANDELBROT, B.B., WALLIS, J.R. (1968), Noah, Joseph and Operational Hydrology, Water Resour. Res., 4(5), 909-918.
- MANDELBROT, B.B., WALLIS, J.R. (1969a), Computer Experiments With Fractional Gaussian Noises. Part 1-Averages and Variances, Water Resour. Res., 5(1), 228-241.
- MANDELBROT, B.B., WALLIS, J.R. (1969b), Computer Experiments With Fractional Gaussian Noises. Part 2-Rescaled Ranges and Spectra, Water Resour. Res., 5(1), 242-259.
- MANDELBROT, B.B., WALLIS, J.R. (1969c), Computer Experiments With Fractional Gaussian Noises. Part 3-Mathematical Appendix, Water Resour. Res., 5(1), 260-267.
- MANDELBROT, B.B., WALLIS, J.R. (1969d), Some Long-Run Properties of Geophysical Records, Water Resour. Res., 5(2), 321-340.
- MANDELBROT, B.B., WALLIS, J.R. (1969e), Robustness of the Rescaled Range R/S in the Measurement of Non-Cyclic Long-Run Statistical Dependence, Water Resour. Res., 5(5), 967-988.
- MARRIOTT, F.H.C., POPE, J.A. (1954), Bias in the Estimation of Autocorrelations, Biometrika, 42, 390-402.
- MATALAS, N.C. (1966), Time Series Analysis, Water Resour. Res., 3(3), 817-829.

- MATALAS, N.C. (1967), Mathematical Assessment of Synthetic Hydrology, Water Resour. Res., 3(4), 931-935.
- MATALAS, N.C., HUZZEN, C.S. (1967), A Property of the Range of Partial Sums, Proc. International Hydrology Symposium, Fort Collins, Colorado State University, Vol. 1, pp 253-257.
- MATALAS, N.C., LANGBEIN, W.B. (1962), Information Content of the Mean, Jour. Geophys. Res. 67(9), 3441-3448.
- MATALAS, N.C., WALLIS, J.R. (1971a), Correlation Constrains for Generating Processes. Proc. International Symposium on Mathematical Models in Hydrology, Warsaw, Int. Ass. Sci. Hydrol., in press.
- MATALAS, N.C., WALLIS, J.R. (1971b), Statistical Properties of Multivariate Noise Processes, Water Resour. Res., 7(6), 1460-1468.
- MEJIA, J.M., RODRIGUEZ-ITURBE, I., DAWDY, D.R. (1972), Streamflow Simulation. 2- The Broken Line Process as a Potential Model for Hydrologic Simulation, Water Resour. Res., 8(4), 931-941.
- MORAN, P.A.P. (1959), The Theory of Storage, London, Methuen and Co. Ltd., pp 111.
- MORAN, P.A.P. (1964), On the Range of Cumulative Sums, Ann. Inst. Stat. Math., 16, 109-112.
- MOSS, M.E. (1972), Reduction of Uncertainties in Autocorrelation by the Use of Physical Models, Proc. International Symposium on Uncertainties in Hydrologic and Water Resource Systems, University of Arizona, Tucson Vol. 3, 1361-1362.
- O'CONNELL, P.E. (1971), A Simple Stochastic Modelling of Hurst's Law, Proc. International Symposium on Mathematical Models in Hydrology, Warsaw, Int. Ass. Hydrol. Sci., in press.
- O'CONNELL, P.E. (1974), Stochastic Modelling of Long-Term Persistence in Streamflow Sequences, Ph. D. Thesis, London, Imperial College pp 284.
- PAPOULIS, A. (1965), Probability, Random Variables and Stochastic Processes, McGraw-Hill Co., pp 583.
- QUENOUILLE, M.H. (1956), Notes on Bias in Estimation, Biometrika, 43, 353-360.
- RIPPL, W. (1883), The Capacity of Storage Reservoirs for Water Supply, Proc. Instn. Civ. Engrs., 71, 270-278.

- SOLARI, M.E., ANIS, A.A. (1957), The Mean and Variance of the Maximum of the Adjusted Partial Sums of a Finite Number of Independent Normal Variates,
Ann. Math. Stat., 28, 706-716.
- SUDLER, C.E. (1927), Storage Required for the Regulation of Streamflow,
Trans. Am. Soc. Civ. Engrs., 91, 622-704.
- THOMAS, H.A., BURDEN, R.P. (1963), Statistical Analysis of the Reservoir Yield Relation,
Report, Chap. 1, pp. 1-21, Cambridge, Mass., Harvard Water Resources Group.
- THOMAS, H.A., FIERING, M.B. (1962), Mathematical Synthesis of Streamflow Sequences for the Analysis of River Basins by Simulation,
A. Maas et al, "Design of Water Resources Systems", Chapter 12, Cambridge, Mass. Harvard University Press.
- THORNER, H. (1967), Finite Sample Monte Carlo Studies: An Autoregressive Illustration,
Journal Am. Stat. Assoc., Vol. 62, 801-818.
- TSCHANNERL, G. (1970), A Decision Framework for the efficient Use of Data Sources,
Center for Population Studies Harvard University Cambridge, Massachusetts, No. 4.
- VALENCIA, D., SCHAAKE, J.C.Jr. (1972) Disaggregation Processes in Stochastic Hydrology,
Water Resour. Res., 9(3), 580-585.
- WALLIS, J.R., MATALAS, N.C. (1970), Small Sample Properties of H and K-Estimators of the Hurst Coefficient h.
Water Resour. Res., 6(6), 1583-1594.
- WALLIS, J.R., MATALAS, N.C. (1971), Correlogram Analysis Revisited,
Water Resour. Res., 7(6), 1448-1594.
- WALLIS, J.R., MATALAS, N.C. (1972), Sensitivity of Reservoir Design to the Generating Mechanism of Inflows,
Water Resour. Res., 8(3), 634-641.
- WALLIS, J.R., O'CONNELL, P.E. (1972), Small Sample Estimation of ρ ,
Water Resour. Res., 8(3), 707-712.
- WALLIS, J.R., O'CONNELL, P.E. (1973), Firm Reservoir Yield: How Reliable are Historic Hydrological Records,
Bull. Int. Ass. Hydrol. Sci., 18(3), 347-365.
- YEVJEVICH, V.M. (1963), Fluctuations of Wet and Dry Years: Part I - Research Data Assembly and Mathematical Models,
Hydrology Paper 1, Colorado State Univ., Fort Collins, pp.55.
- YEVJEVICH, V.M. (1967), Mean range of Linearly Dependent Normal Variables With Application to Storage Problems,
Water Resour. Res., 3(3), 663-671.

ZELLNER, A. (1971), An Introduction to Bayesian Inference in Econometrics,
John Wiley and Sons. pp 385.