ANALYSIS OF DROUGHT CHARACTERISTICS BY THE THEORY OF RUNS

by

Pedro Guerrero-Salazar
and
Vujica Yevjevich

September 1975

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

80

ANALYSIS OF DROUGHT CHARACTERISTICS BY THE THEORY OF RUNS

by

Pedro Guerrero-Salazar
and
Vujica Yevjevich

September 1975

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

80

ANALYSIS OF DROUGHT CHARACTERISTICS BY THE THEORY OF RUNS

by

Pedro Guerrero-Salazar*

and

Vujica Yevjevich**

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO

September 1975 No. 80

*Previously, Ph.D. graduate student at Colorado State University. Presently, associate professor of Civil Engineering at COPPE (Coordinacao dos Programas de Pos-Graduacao em Engenharia), the Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.

**Professor of Civil Engineering and Professor-in-Charge of Hydrology and Water Resources Program, Civil Engineering Department, Colorado State University, Fort Collins, Colorado, USA.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

This paper results from the research in the Hydrology and Water Resources Program, Department of Civil Engineering, at Colorado State University, made possible by the financial support of the U.S. National Science Foundation under the grant GK-11564 (Large Continental Droughts), and GK-31512X (Stochastic Processes in Water Resources) with V. Yevjevich as the principal investigator. The financial support under this project that gave the opportunity for advanced studies are gratefully acknowledged.

The doctoral dissertation by Pedro Guerrero, with V. Yevjevich the advisor, served as the basic material for shaping this paper. Thanks are expressed to Dr. Duane C. Boes and Dr. Mohammed M. Siddiqqi, professors in the Department of Statistics of Colorado State University, for their advice in statistical developments. Dr. Carl C. Nordin of the U.S. Geological Survey and Dr. David Woolhiser of the U.S. Agricultural Research Service were very helpful with their comments during different stages of the study. Dr. N.T. Kot egoda, from the University of Birmingham, England, on sabbatical leave with Colorado State University, reviewed the material of this paper in detail, giving useful suggestions, which is gratefully acknowledged.

ABSTRACT

Methodologies for analysis of droughts are presented on the basis of objective definitions of droughts for stationary and periodic-stochastic processes. Droughts of stationary series are studied by means of the theory of runs. Distributions of the longest run-length and the largest run-sum in a series of a given length, and distributions of the run-length and the run-sum of infinite series for various cases of univariate and bivariate series are investigated. Exact, approximate or experimentally obtained expressions are presented for univariate and bivariate independent and dependent series. For the bivariate series all combinations of serially independent and dependent, and mutually independent and dependent series are studied. Where exact or approximate analytical solutions could not be obtained, the data generation method is used, with results checked by using particular cases for which the exact solutions are available. Frequency distributions of various drought characteristics associated with the runs, obtained by the generation method for the bivariate case, are fitted by discrete or continuous probability distribution functions, respectively for the run-length and the run-sum.

Multiple regression analysis is used to obtain useful relationships between the parameters of fitted distribution functions and the parameters of time series dependence, cross dependence and the truncation levels of the basic series.

Periodic-stochastic series are studied by defining drought and its parameters for this particular type of hydrologic processes. New approaches and techniques are presented with a case study illustrating the power of these new approaches.

PREFACE

Pressure for a higher standard of living and the increase of world population continuously require more food, energy, raw materials, industrial production and various services. The inevitable result is the increase in pressure with time on all types of world-wide available water resources. Because these renewable natural resources on continental areas are constant, in their averages, regardless of their space and time variations, sooner or later the increase in water demand faces space and time shortages because of stochastic variations in water supply and demand. The experiences and investigations show that the risks of water shortage increases rapidly with an increase of utilization of the total available water resources in an area. Particularly sensitive in this regard is the food production as the most important commodity of a world living on the margins of balance between food supply and food demand. Usually water shortages of drought proportions have the largest impact on the agricultural production.

Confusion governs the selection of random variables which are used to define the concepts of water shortages, deficits and droughts. Differences between water demands and water supplies, as periodic-stochastic processes, are crucial in defining the shortages, deficits and droughts. Difficulties often arise with the meaning of the terms such as water demand, requirement, use, consumption, deliveries, rights, and accompanying factors. It is rare to meet two individuals of different professional backgrounds who have the same connotation of the term "drought."

International organizations (such as UNO, UNDP, FAO, UNESCO, WMO, regional UN commissions, scientific and professional associations) and national and regional organizations are concerned with both the broad and the specific problems related to drought phenomenon and its consequences. International conferences are held on population, environmental control, food production, food distribution, eventual international food storage, and on similar subjects which are strongly related to droughts. Characteristics of these meetings are discussions in generalities, often without sufficient scientific information for claims, positions and proposals. Feeding the world population and the establishment of world-wide food storage centers are ever-increasingly important issues of a very sensitive character. Only the most correct information, on an advanced scientific level, can replace the subjective approaches by a more objective analysis and decision making process.

Three characteristics related to drought consequences and drought control technology can be distinguished at present:

(1) An unusually high emphasis is given to atmospheric circulation in search for explanations and predictions of droughts and related agricultural food production. This emphasis may enhance the understanding of atmospheric processes but definitely lacks predictability of droughts of long duration, large water deficits and extensive areal coverages.

(2) Great attention is paid to droughts of semi-arid and arid regions of presently marginal agricultural production, while a surprisingly small attention is given to drought risks and necessary drought control technology to mitigate its consequences in the semi-arid regions of presently substantial world food production (US Midwest, USSR steppe, Canadian prairies, Argentinian pampas, Australian wheat regions, and similar areas). Droughts in the marginal regions cause stress on several millions of people, while droughts in the large food-producing regions do not only disrupt the world food prices but also involve the fate of hundreds of millions of people.

(3) It is a common and necessary expectation to search for new agricultural technologies and new arable lands in order to increase the food production. This line of activity is and should be the principal thrust for an increase in food supply. However, stabilization of food production by using the presently available technologies and lands already under cultivation, and finding solutions for random fluctuations in food supply, represent a task as important as the search for new technology and new lands. In several aspects, this stabilization and solutions for fluctuations in food production may be as important and productive as the search for new technology and new lands. Understanding the drought phenomenon, and particularly finding the best mix of drought control measures specific to each region, for solving the problems of stabilization in food supply, including the establishment of food storage centers, are the challenging tasks to a multidisciplinary scientific approach.

Random variables must be well selected if they are to be meaningfully used for definitions of water shortages, deficits and droughts. Soil moisture, precipitation, evaporation, groundwater levels, river runoff, state of water storage in reservoirs and lakes, snow and ice accumulation and melting, and similar variables are periodic-stochastic space-time processes, which must be used either individually or in combinations, and according to the problem at hand, for the definition of the three concepts of shortages, deficits and droughts. It seems that as many definitions of these three concepts are available as there are investigators. This creates confusion among the users of information on droughts. In general, droughts are associated with water deficits of long duration, high intensity of deficits, and large areal coverage, usually involving all water resources variables and users, having significant economic and social consequences. Deficits can be related to the lack of water at a given place for a given time interval, with the relatively moderate consequences. Shortages are a small negative difference between water demand and water supply, with readily acceptable consequences. Definitions of the three concepts of droughts, deficits and shortages, acceptable to a majority of professionals in the world, need a universal acceptance.

Droughts are a creeping-type disaster phenomenon. In studying physical aspects of droughts, the following properties of drought-defining variables are of practical significance: duration of shortages, total water deficits over this duration, areal coverage by this total deficits, intensity of largest shortages, and similar random variables. These variables are best described by joint or marginal probability distributions of individual variables. The properties of these random variables are related either to population or to samples of various sizes. Assuming a multivariate or a univariate of water supply variable(s) as the input process, and a multivariate or a univariate of water demand variable(s) as the output process of agricultural and water resources systems, the crossing of these two time processes provides the necessary information for computing or estimating the probabilities of drought properties. Furthermore, the economic drought properties, as functions of a mutually dependent set of random variables, therefore also as random variables, are necessary for solutions of drought problems.

In contrast to atmospheric circulation approach to drought investigations, investigations of probability distributions of drought properties should be realistically based on past records of selected climatic and hydrologic random variables, under the following two basic hypotheses:

(1) Inferences on population characteristics of drought properties, based on drought-defining periodic-stochastic variables, are subject to sampling errors (often with historic non-homogeneity and systematic errors in samples, which must be first identified and removed), requiring the unbiased and most efficient estimation techniques; and

(2) General climate and resulting hydrologic periodic-stochastic processes over the next 150-200 years will have essentially the same population characteristics (structures and parameters) as the records of the past 150-200 years demonstrate; this assumption has a strong support, namely that of a temporary stationarity of annual values of these periodic-stochastic processes, regardless of a continuous production of papers with the claims of expected sudden changes in the climate.

Reliable probabilistic characteristics of drought properties are fundamental as the information for any advanced approach to technologic, economic and social aspects in drought investigations and related decision making. Economic aspects are basically of two types:
(a) measurement of and modeling the economic damages and regional consequences due to droughts; and
(b) economic benefit-to-cost analysis for optimization in selecting a mix of drought control measures.

In connecting probabilities of physical drought properties to economic drought impacts, especially in the agricultural production, new indices are needed on droughts if information produced should seriously affect the decision making process. Furthermore, a relationship exists between physical drought properties, loss of agricultural production and the population involved. This then requires additional indices and mathematical modeling in order to take into account all factors. Social consequences of droughts, with all the political implications, represent a synthesis of drought analysis and drought control. They are less prone to be measured by indices or by mathematical modeling, usually being analyzed by descriptive methods.

Drought investigations cannot be productive without using advanced methodologies in selecting drought control measures, as the drought control technology, by optimizations and particularly well

designed decision making process. For a future development of such methodologies, the following assumptions are necessary:

(1) Drought control measures may be divided into internal measures to a water user and to external measures to all or most of water users. Internal measures are such as moisture or water conservation inside a production unit, various types of adjustments to water shortages, replacements, changes in the production mix and technology, and similar measures. External measures are basically water storage and regulation outside the production units, uni-directional water transfer, water interchange between adjacent regions, and weather modification. Furthermore, insurance against drought losses and storage of various products in water surplus times for water deficit times complement the classification of drought control measures in their most general treatment.

(2) Because of large varieties and a range of levels of drought control measures, it should be rarely expected that only a single measure would result as an economic and social optimum. More often than not, a mix of most of relevant drought control measures would come out to be a global optimum for a given region.

(3) Treatment of drought control measures is an interdisciplinary and multidisciplinary problem, subject to a most effective treatment only by a team of specialists and generalists.

(4) The systems analysis is a good approach to major drought problems, not only for drought description, responses to it, determination of its loss function and the selection of an optimal mix of drought control measures, but also for incorporating inputs from various disciplines for both a large-scale and a small-scale approach to drought investigation problems.

The contributions to drought investigations until 1968 have been presented in the form of annotated references in the publication "Drought Bibliography," prepared by Wayne C. Palmer and Lyle M. Denny, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Data Service, NOAA Technical Memorandum EDS 20, Silver Spring, Maryland, June 1971. Though it does not contain all the literature on a world-wide basis, this bibliography gives a good insight to problems treated, approaches used, and indirectly to the state-of-the-art of various aspects of droughts.

Research on continental droughts has been going on for more than a decade at Colorado State University in the Hydrology and Water Resources Program of its Civil Engineering Department. Different aspects of large droughts, involving long duration, significant water deficits, large areal coverage, and economic impacts on a region have been investigated. The present paper "Analysis of Drought Characteristics by the Theory of Runs" is a continuation of research carried out previously by using the probability theory, mathematical statistics, and stochastic processes under a strict objective definition of drought characteristics.

The paper first reviews the state-of-present-knowledge of droughts of both univariate and bivariate processes. However, the main emphasis and contribution are on drought characteristics for bivariate processes, mainly concerned with droughts of two representative variables. These two variables may be the time series at two selected points, average characteristics of time processes of drought defining variables of two areas or regions, water yields of two river basins, two reservoirs, two aquifers, or their combinations. The major thrust of the paper is intended to contribute to a future methodology of studying large continental droughts using the water supply and demand variables which best define a given drought problem.

Vujica Yevjevich

September 1975
Fort Collins, Colorado

Chapter I
INTRODUCTION

1-1 An Overall Review of Drought Definitions

It is difficult to come out with a universal and commonly accepted definition of a drought. Several authors have tried to define a drought under different conditions, such as the agricultural drought, climatological drought, hydrological drought, etc. (Subrahmanyam, 1967).

A drought is defined in this study on the basis of differences between the processes of water supply and water demand. The supply processes or supply time series may be the precipitation over an area, the streamflow at a given point of a river, moisture in the soil, storage of water in an aquifer or reservoir, and similar hydrologic variables. The demand process or demand time series may be a single-purpose water use, such as water used for agriculture, for continuous or supplemental irrigation, hydropower, water supply, low flow augmentation for quality control, or the demand process may result from a combination of various water uses. When the demand exceeds the supply, the water shortage occurs, and this is the general condition for drought initiation.

Natural and artificial water retentions affect highly the initiation and duration of a drought. The retention occurs naturally in the soil in case of dry farming, or it can be artificial as in case of reservoirs for runoff regulation. Natural storage is considered in this study as a part of water supply. Artificial storage is considered both as a part of water supply when it already exists and as a drought alleviation measure when it is only planned.

The drought analysis is based on time series of water supply and water demand. It is sometimes claimed that reliable data both on water supply and water demand are difficult to obtain even in developed countries. With sufficient efforts, regardless of the relatively scarce data, it is feasible in most cases to gather sufficient information on water supply and water demand for investigation of drought related problems. The periodicity of the year in various parameters of water supply and water demand makes the analysis of droughts somewhat difficult, so that the study of droughts with time intervals of less than a year warrants a special attention.

A drought is defined here as the deficiency in water supply over significant time to meet the water demand for various human activities. This deficiency is mainly produced both by the random character of natural processes that control the distribution of water in space and time on the earth's surface, and by randomness in water demand.

The existence of variety of climates over the earth surface implies that droughts should vary according to climatic characteristics. The climates as classified by Thornthwaite (1948) are arid, semiarid, semihumid and humid. The climate determines the natural biological cover. Combined with human activities it produces the water demand, which differs from region to region and from one time interval to another. The long-term stochastic fluctuations with large variations around the mean of available water makes the problem of long and large droughts much more important in arid and semiarid regions than in semihumid or humid regions.

An objective definition of droughts, based on the theory of runs, may be used for stationary time series (Yevjevich, 1967, 1972b). For the univariate case and discrete time series of water supply, a selected arbitrary variable value or truncation level X_o may represent the water demand, as shown in Fig. 1-1. The

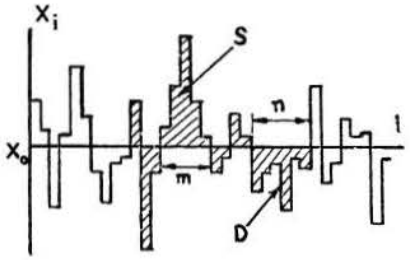

Fig. 1-1 Definitions of Positive Run-Length, m, Positive Run-Sum, S, Negative Run-Length, n, and Negative Run-Sum, D, for a Discrete Series, x_i.

discrete series truncated by this constant x_o gives two new truncated series of positive and negative differences. A sequence of consecutive negative deviations preceded and followed by positive deviations is called a negative run-length (n in Fig. 1-1); it may be associated with the duration of a drought. In this context, the definition was used by Llamas and Siddiqui, 1969; Saldarriaga and Yevjevich, 1970; Millan and Yevjevich, 1971; and Millan, 1972. The sum of all negative deviations over such a run-length is called the negative run-sum (D in Fig. 1-1), and the ratio of the negative run-sum and the negative run-length is called the negative run-intensity (D/n, Fig. 1-1).

For a two-dimensional process $\{X_i, Y_i\}$, with distribution $F(x,y)$, the following concepts can be used (Yevjevich, 1972b). Two crossing or truncation levels are now used, denoted by x_o and y_o (Fig. 1-2), which are not necessarily of the same

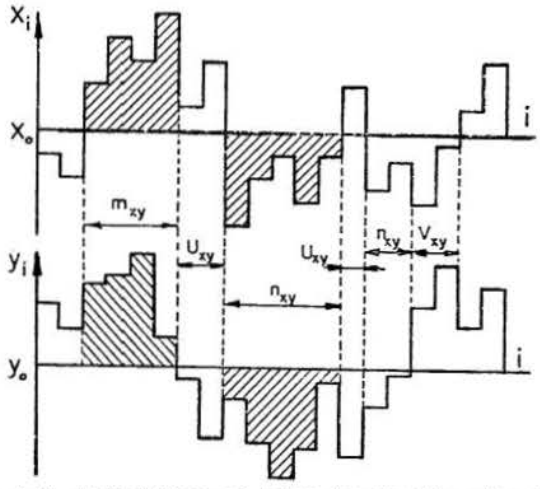

Fig. 1-2 Definitions of Joint Run-Lengths of a Two-Dimensional Process, with Two Constant Crossing Levels, x_o and y_o.

probability for each marginal distribution. Four events are obtained as shown in Fig. 1-2: both deviations are positive which define the joint positive run-length (m_{xy}); both deviations are negative which define the joint negative run-length n_{xy}; x_i are positive and y_i negative deviations which define the joint positive-negative run-length (U_{xy}); and x_i are negative and y_i positive deviations which define the joint negative-positive run-length (V_{xy}). The joint run-sum is defined as the sum of deviations of both the run-sum in x_i and the corresponding run-sum in y_i over the corresponding joint run-length. Consequently, there are four different types of run-sums, one for each of the four types of joint run-lengths. The joint run-intensities are defined as the joint distribution of the intensity in x and the intensity in y over the joint run-length.

For the case of hydrologic periodic-stochastic series, the theory of runs cannot be used directly and simply as in the case of stationary stochastic processes, because of the periodicity involved. In this case criteria must be developed concerning the parameters of drought magnitude, duration and volume. For the unidimensional case, the drought magnitude criteria can be defined as the minimum of the mean monthly difference between supply and demand over the duration of a drought.

1-2 Objectives of Investigations

The first objective of this study is to determine the joint probability distribution of hydrologic droughts for two hydrologic time series, concurrently observed at two locations. The second objective is to find the relations of characteristics of probability distributions of joint drought occurrence at two locations and the statistical parameters of the corresponding two hydrologic time series. Since the theory of bivariate runs has not been developed yet (Yevjevich, 1972b), this study is a contribution towards this goal. The third objective is initiate a development of a methodology of studying droughts of hydrologic periodic-stochastic processes, exemplified here by monthly time series.

1-3 Organization of the Study

The study of droughts of the bivariate stationary case is presented in Chapter II by giving the exact analytical expressions for the simple cases and by analytical approximations for the more complex cases. The experimental (Monte Carlo) approach, which was used for cases for which even the approximate expressions are not available, is presented in Chapter III. Results of the experimental approach are given in Chapter IV. Discrete density functions are fitted to frequency distributions of run-lengths, while continuous density functions of the Pearson family of functions, and series expansion approach, are used to fit the frequency distributions of run-sums. This approach allows the parameters of distributions to be expressed in terms of basic statistics of the two underlying hydrologic time series by using the multiple regression equations. Since the theory of runs of stationary series is not adequate for the analysis of droughts of periodic-stochastic processes, the runs of these processes are discussed in Chapter V, with an example.

Chapter II

ANALYTICAL INVESTIGATION OF DROUGHTS OF STATIONARY TIME SERIES USING NEGATIVE RUNS

The theory of runs as used here to investigate the droughts of stationary stochastic processes has been a topic of inquiry for a long time. Reviewing the statistical literature one observes that several definitions of runs are used.

2-1 Definitions of Runs

Three definitions have been proposed in literature for runs called here: classical, recurrence and Mood's definitions.

Classical definition of runs. This definition is probably given first by De Moivre, Uspensky (1937), among others. It is defined as a success-run of length r in a series of independent trials when a success occurs at least r times in succession. In Feller's words (1957), it is an uninterrupted sequence of either exactly r or of at least r successes. According to Feller, this definition has the following drawback. If exactly r successes are required, a success at the (n+1)-th trial may make null the run completed at the n-th trial. On the other hand, if at least r successes are required, every run may be prolonged indefinitely, and the occurrence of a run does not reestablish the initial situation.

Recurrence or Feller's definition of runs. A run of length r (Feller, 1957) to be used in recurrence theory is uniquely defined with the counting starting every time a run occurs. Namely, a sequence of n events of 0 and 1 contains as many runs of 0 of the length r as there are non-overlapping and uninterrupted blocks containing exactly r events of 0. This definition is not well suited for the analysis of droughts, since it does not say when a run starts or when it finishes, because a run-length of three zeros, for example, may be preceded or succeeded by zeros.

Mood's definition of runs. Mood's (1940) definition seems the most suited for the analysis of droughts because a run is defined as a succession of similar events preceded and succeeded by different events, with the number of elements in a run referred to as its length, as shown in Fig. 1-1.

The above distinction of various definitions of runs is needed because the articles in the statistical literature sometimes treat runs without clarifying in which sense the term "run" is used. A reader may be often misled. Mood's definition of runs is the definition used throughout this study only.

Runs as they are used in statistics are characterized as a philosophy and a technique (Wolfowitz, 1943). The ordering of observations according to some characteristic is always involved, and the results of this ordering is again ordered according to some other characteristic. In the case of hydrologic applications, the characteristic which defines runs is the occurrence of series values above or below a certain level. This level does not need to be the same for all time positions.

2-2 Approaches to Analysis of Run-Length

In the application of the theory of runs to hydrologic problems two approaches have been followed in various studies of run lengths: the integration approach and the combinatorial approach.

The integration approach refers to runs of an infinite population, which in the case of stationary and ergodic series is synonymous with the first run. In this context the term infinite population will be used. The combinatorial approach treats the runs in a sample of given size.

For the case of run-length, the integration approach is based on finding the probability

$$P\ (\text{run-length} = k) = P\ (x_i > C;\ x_{i+1} \leq C;\ldots;\ x_{i+k} \leq C;\ x_{i+k+1} > C).$$

If the joint distribution of the x_i's is known, the integration approach gives the required probability. If the time process is independent, the computation is simple because the product of the marginal probabilities give the probability of the run-length. A drawback in the integration approach is that it does not permit the computation of the probability of a run-length equal to k in n trials, which the combinatorial approach does. Furthermore, the analytical expressions for the other types such as the run-sum, and the run-intensities are very complex to integrate for the dependent bivariate cases.

Probabilities of various runs are studied in this chapter by using the theory of runs for the case of infinite population and for both the univariate and the bivariate cases. The exact analytical solutions are obtained only for simple basic processes, while approximations are obtained for more complex cases. The data generation or Monte Carlo approach is used for those cases for which neither the exact analytical nor approximate analytical solutions are feasible.

For the combinatorial approach the run sample statistics studied differ according to the objective for which the run theory is used. Such statistics are the total number of positive and negative runs regardless of their length, the total number of runs of a given kind, the longest run-length of either kind, the longest run-length of a given kind, the largest run-sum, the other run-sums, the run-intensities, and any other statistic of interest. For drought purposes, types of common interest such as the longest and the second longest negative run-length, and the largest and the second largest run-sum, are investigated in this paper.

The combinatorial approach in the case of run-lengths makes use of a transformation to a zero-one process. Whenever a value is below the truncation level the new random variable is one and whenever a value is greater than the level the new variable is zero. Taking advantage for the independent case of the fact that the new variable has a Bernoulli distribution of events 0 and 1, the combinatorial approach may be used. For the independent case, as shown later, it is simple to obtain the probability: P (run-length = k in n trials).

The combinatorial approach is adequate for those hydrologic problems which relate to the probability of extreme events in a sample, for example a drought duration of a given probability to occur in the life of a project of n years. This approach is used in this paper to obtain the analytical approximations or exact expressions for the most simple cases of

underlying stochastic processes. The results are also used to check the experimental or Monte Carlo method of deriving the properties of runs in the sample of a given size for more complex cases.

The empirical method of studying droughts for stationary time series is discussed by Saldarriaga and Yevjevich (1970) for runs of infinite series. The sample data obtained by the empirical techniques are used to determine the probabilities of durations of droughts. The empirical procedure is as follows. Run-lengths are measured with respect to a given truncation level and the relative frequencies of run-lengths that are greater than a given duration are computed. These frequencies provide the estimates of probabilities. This enables the study of drought measures with droughts not to be exceeded, on the average, in a given number of years. These frequencies are used as probabilities of droughts of a given duration, and as probabilities of all events equal to or greater than a given duration. Because sample sizes of hydrologic data are small, large sampling errors are common in the estimates of these probabilities. Drought probabilities are studied analytically making use of statistics of the basic processes. These have smaller sampling variations than the above computed frequencies. A convenient analytical method is the theory of runs. Run-length properties are distribution free in comparison to run-sum and run-intensity properties which are dependent on the type of the underlying distribution.

2-3 Probabilities of Longest Run-Length in a Sample of Size n for Univariate Independent Process

The study of the longest run-length in a sample of size n for independent series was initiated by De Moivre (1738) when finding the probability of a sequence of r successes in n trials. Following Whitworth (1896, Propositions XXVIII and LII) an experiment succeeds m times and fails n times, the probability that the longest run-length of successes is less or equal to k in $m+n$ trials is the coefficient of x^m in the expansion of the expression

$$\frac{1}{\binom{m+n}{n}} \left(\frac{1-x^{k+1}}{1-x}\right)^{n+1} \tag{2-1}$$

This expression resulted from the number of ways in which m items can be distributed into $n+1$ different compartments with no compartment to be either empty or to have more than $k+1$ items, which is the coefficient of x^m in the expansion of the expression

$$\left(\frac{1-x^{k+1}}{1-x}\right)^{n+1} \tag{2-2}$$

Similarly, Bateman (1948) presents the number of ways of arranging r_i elements ($i=1,2$) into t parts none of which exceeds k in magnitude. In the same way, Mosteller (1941) presents the special case of the probability of one or more runs not less than k in length amongst all runs of values below the median. For Mosteller, the coefficient of x^n in

$$(x + x^2 + \ldots + x^{k-1})^{r_1} \tag{2-3}$$

gives the number of ways of partitioning n elements into r_1 partitions in such a way that no partition contains k or more elements and none is void. Rewriting the above expression as

$$x^{r_1}\left[(1-x^{k-1})\right]^{r_1} \sum_{t=0}^{\infty} \binom{r_1 - 1+t}{r_1 - 1} x^t, \tag{2-4}$$

the coefficient of x^n becomes

$$\sum_{j=0}^{r_1} (-1)^j \binom{r_1}{j}\binom{n-j(k-1)-1}{r_1 - 1}, \tag{2-5}$$

or as Bateman presented it

$$\sum_{j=0}^{t} (-1)^j \binom{t}{j}\binom{n-jk-1}{t-1}, \tag{2-6}$$

which is identical to the coefficient of x^{r_i} in the expansion of the equation

$$x^t \left(\frac{1-x^k}{1-x}\right)^t.$$

Furthermore, the number of ways of arranging r_i elements into t parts of magnitude k is

$$f_i(t,k) = \sum_{j=1}^{t} (-1)^{j+1}\binom{t}{j}\left[\binom{r_i-j(k-1)-1}{t-1} - \binom{r_i-jk-1}{t-1}\right]. \tag{2-7}$$

An explicit expression for the probability distribution of the longest run-length of a given kind in a series of n independent trials was given by Bateman (1948). A sequence of r elements is studied, of which r_1 are of one type and r_2 of another type, with $r_1+r_2 = r$. For example, a sequence of r years of annual precipitation is studied of which r_1 years are deficit years and r_2 are surplus years, with $r_1+r_2 = r$. The total number of possible combinations ${}_rC_{r_1}$ which can be formed from the r elements constitutes the fundamental probability set. The subset of all combinations each containing at least one run-length of a given kind and of a given length g_d can be determined by considering the partitions of r_1 elements having k as the greatest part, where $k = 1,2,\ldots,g_d$ and finding the number of ways in which they can be combined to form a combination with at least one part equal to g_d and no part greater than g_d. This may be achieved simply by considering the different ways in which such partitions of r_1 form groups of length $2t$ or $2t+1$, where $t=1,2,\ldots,r_1-g_d+1$ for $r_1 \geq r_2$. There will be no loss of generality in assuming $r_1 \geq r_2$.

The number of sequences of $2t$ groups with at least one group containing g_d elements and no group containing more than g_d elements, designated by $N(2t, g_d|r_1,r_2)$ is

$$N(2t,g_d|r_1,r_2) = 2f_1(t,g_d)\binom{r_2-1}{t-1}. \qquad (2\text{-}8)$$

The factor 2 is introduced to allow for the sequence to begin with either a deficit or a surplus. In the same way, the number of sequences of $2t+1$ groups of which the largest has g_d elements is

$$N(2t+1,g_d|r_1,r_2) = f_1(t+1,g_d)\binom{r_2-1}{t-1} + f_1(t,g_d)\binom{r_2-1}{t}. \qquad (2\text{-}9)$$

The enumeration of the required subset is completed by summing $N(2t,g_d|r_1,r_2)$ and $N(2t+1,g_d|r_1,r_2)$ over all groups, i.e., from t=1 to $t = r_1-g_d+1$. Denoting this subset by $N(g_d|r_1,r_2)$ then

$$N(g_d|r_1,r_2) = \sum_{t=1}^{r_1-g_d+1}\left\{2f_1(t,g_d)\binom{r_2-1}{t-1} + f_1(t+1,g_d)\binom{r_2-1}{t-1} + f_1(t,g_d)\binom{r_2-1}{t}\right\}. \qquad (2\text{-}10)$$

Factorizing and simplifying terms, Eq. 2-10 becomes

$$N(g_d|r_1,r_2) = \sum_{t=1}^{r_1-g_d+1} f_1(t,g_d)\left[\binom{r_2+1}{t}\right]. \qquad (2\text{-}11)$$

Hence in a sequence of r elements, r_1 of which are deficit and r_2 are surplus, with $r_1+r_2 = r$ and $r_1 \geq r_2$, the probability that the longest deficit run consists of g_d elements is

$$P[G_d = g_d|r_1,r_2] = \frac{N(g_d|r_1,r_2)}{\binom{r}{r_1}}. \qquad (2\text{-}12)$$

The probability of the longest negative run-length being equal to or longer than a given value, say g_d, is

$$P[G_D \geq g_d|r_1,r_2] = \frac{\sum_{t=1}^{r_1-g_d+1}\binom{r_2+1}{t}\sum_{j=1}^{t}(-1)^{j+1}\binom{t}{j}\binom{r_1-j(g_d-1)-1}{t-1}}{\binom{r}{r_1}} \qquad (2\text{-}13)$$

Equation 2-13 presented by Bateman (1948) is a more general equation than that given by Mosteller (1941). Mosteller considered the case of runs above and below the median, where $r_1 = r_2 = r/2$ n, for a sample of even size, and derived the probability of obtaining at least one run equal to or longer than a given length.

Equation 2-13 for these conditions becomes the Mosteller's equation. Replacing

$$\binom{t}{j}\binom{r_2+1}{t} \quad \text{by} \quad \binom{r_2+1}{j}\binom{r_2-j+1}{t-j},$$

in Eq. 2-13, interchanging the order of summation, and using the relation

$$\sum_{i=0}^{m}\binom{m}{k+i}\binom{n}{i} = \binom{m+n}{k+n},$$

then

$$P[G_d \geq g_d|r_1,r_2] = \frac{\sum_{j=1}^{r_1/g_d}(-1)^{j+1}\binom{r_2+1}{j}\binom{r_1+r_2-jg_d}{r_2}}{\binom{r}{r_1}}, \qquad (2\text{-}14)$$

because $m = r_1 - j(g_d-1) - 1$, $k = j-1$, $n = r_2 - j+1$, and $i = t-j$. If only r is given and the probability of a deficit to occur is constant and equal to p, with

$$P[r_1] = \binom{r}{r_1} p^{r_1}(1-p)^{r-r_1},$$

then

$$P[G_d \geq g_d] = \sum_{r_1=g_d}^{r} P[G_d \geq g_d|r_1,r]\,P[r_1]. \qquad (2\text{-}15)$$

The probability that a deficit occurs at least g times in succession in a series of n independent trials with the probability p of the deficit at any trials is the well known problem of the "runs of luck" solved by De Moivre (1738). The same problem has been solved using difference equations by Uspensky (1937), and is also given by Whitworth (1896), Cramer (1946) and others. This can also be obtained using Eq. 2-12 and summing up accordingly.

Making use of generating functions, denoting $P_{n,g} = P$, the longest run $\leq (g-1)$ in n trials, and $P(G_d \geq g_d) = 1 - P_{n,g}$, their generating function is

$$\psi(x) = \sum_{n=1}^{\infty} P_{n,g}\,x^n = \frac{1-\psi(x)}{1-x} = \frac{1 - p^g x^g}{1 - x + p^g q\, x^{g+1}}, \qquad (2\text{-}16)$$

so that the coefficient of the x^n term is the probability that the longest run is less than or equal to (g-1) in n trials. The proof is given by Uspensky (1937, pages 78-79) and also through combinatorial theory by Whitworth (1896, Proposition LIII).

The generating function $\psi(x)$ is a rational function and can be developed into a power series of x according to known rules. Uspensky shows that the coefficient of x^n is

$$P_{n,g} = \beta_{n,g} - p^r \beta_{n-g,g} , \qquad (2\text{-}17)$$

with

$$\beta_{n,g} = \sum_{\ell=0}^{n/g+1} (-1)^{\ell} \binom{n-\ell g}{g} (qp^g)^{\ell} \qquad (2\text{-}18)$$

and $\beta_{n-g,g}$ is obtained by substituting $n-g$ for n. David and Barton (1962) give a solution for $P_{n,g}$, based also on the combinatorial analysis, as

$$P_{n,g} = \sum_{r_1=0}^{g_d} P[G_d \le g_d | r_1, r] \; P[r_1] , \qquad (2\text{-}19)$$

and

$$P[G_d \le g_d | r_1, r] = \frac{\sum_{i=0}^{a} (-1)^i \binom{r_2+1}{i} \binom{n-i(m+1)}{r_2}}{\binom{r}{r_1}}$$

with $a = \min\{r_2+1, (\frac{n-r_2}{m+1})\}$,

and $n + 1 - r_2 \ge m + 1 \ge [\frac{n+r_2+1}{r_2+1}]$.

The parameters of the above sampling distributions of the longest run-length are not available except for special cases but only as approximations. Cramer (1946) gives the asymptotic mean (valid for large sample sizes) of the distribution of the longest run-length, g_d, for the sample of size n as

$$E[g_d] = - \frac{\log n}{\log(1-q)} + 0(1), \qquad (2\text{-}20)$$

with $q = P(x \le C)$, C the truncation level, and $0(1)$ an error term of the order of one.

Baticle (1946) studying the problem of repartitions gives asymptotic equations for parameters of the sampling distribution of the longest run of consecutive successes in n trials, valid for $(g/s) \to 0$, with g the length of the longest run, and s the total number of successes, as

$$E[\frac{g}{s}] = \frac{1}{n} \left[1 + \frac{1}{2} \quad \frac{1}{3} + \ldots\ldots + \frac{1}{n}\right] , \qquad (2\text{-}21)$$

and

$$E[(\frac{g}{s})^2] \approx \frac{2}{n(n+1)} [1 + f(n) + \frac{1}{2} (f(n))^2], \qquad (2\text{-}22)$$

with

$$f(n) = \frac{1}{2} + \frac{1}{3} + \ldots\ldots\ldots + \frac{1}{n} . \qquad (2\text{-}23)$$

Burr and Cane (1961) present approximations to the exact expression presented previously by Whitworth and Mosteller. Another approximation presented by David and Barton (1962) is

$$P[G_d \le g_d | r_1, r] \approx e^{-[(r_2+1) \frac{r_1^{m+1}}{n^{m+1}}]} , \qquad (2\text{-}24)$$

which is valid for large g_d and $r \ge 20$.

2-4 Probabilities of Longest Run-Length in a Sample of Size n for Univariate Dependent Process

Approximation of the first-order linear autoregressive model by Markov chains. The case of the uni-dimensional dependent time series can be solved for the first-order linear autoregressive model,

$$x_{i+1} = \rho x_i + \sqrt{1-\rho^2} \; \varepsilon_{i+1} ,$$

where ρ is the first serial correlation coefficient of the standardized series x; and ε_i is a sequence of independent identically distributed variables. This model is approximated either by a first-order Markov chain or better by a second-order Markov chain. The approximation for the first-order Markov model is then

$$P[x_{i+1} \le C | x_i \le C, \ldots, x_{i-n} \le C]$$

$$= P[x_{i+1} \le C | x_i \le C][1 + \phi(\rho^2)], \qquad (2\text{-}25)$$

with $\phi(\rho^2)$ an error term. Millan (1972) found that for $\rho \le 0.4$ the approximation is good. In the case of a first-order Markov chain used to approximate the first-order autoregressive model, the transition probabilities may be obtained by using the autoregressive model, namely

$$P_1 = P[x_{i+1} \le C | x_i \le C] = \frac{P[x_{i+1} \le C, \; x_i \le C]}{P[x_i \le C]} ,$$

$$1 - P_1 = P[x_{i+1} > C | x_i \le C] = \frac{P[x_{i+1} > C, x_i \le C]}{P[x_i \le C]} , \qquad (2\text{-}26)$$

with the joint probabilities obtained from tables for the case of a normal distribution. The transition probability values are

$$P_1 = P[x_i \le C | x_{i-1} \le C] , \quad Q_1 = 1 - P_1 ,$$

$$P_2 = P[x_i \le C | x_{i-1} > C] , \quad Q_2 = 1 - P_2 , \qquad (2\text{-}27)$$

Development of probability distribution of the longest run-length for simple Markov chains. Bateman (1948) obtained the distribution of the longest run in n trials regardless of its kind. The probability distribution of the longest run of a given kind, say the negative run, in a sample of size n, as developed in this paper, is outlined below. Considering the partitions of r_1 and r_2, for each partition within a given number of $2t$ or $2t+1$ groups, the multiplying probabilities are the same as the number of transitions from $(x_i > C)$ to $(x_{i-1} \le C)$, and the opposite.

Thus for a given sequence of $2t$ groups beginning with $(x_i \le C)$ there are $2t-1$ transitions, t from $(x_i \le C)$ to $(x_{i-1} > C)$ and $t-1$ from $(x_{i-1} > C)$ to $(x_i \le C)$, while the remaining $r_1 - t$ and $r_2 - t$ cases are continuations of $(x_i \le C)$ and $(x_i > C)$, respectively. The probability of obtaining a given sequence of $2t$ groups is

$$PP_1^{r_1-t} P_2^{t-1} Q_1^t Q_2^{r_2-t} + QQ_1^{t-1} Q_2^{r_2-t} P_1^{r_1-t} P_2^t , \qquad (2\text{-}28)$$

which may be written as

$$\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)\left(\frac{P_2Q_1}{Q_2P_1}\right)^t P_1^{r_1} Q_2^{r_2} . \qquad (2\text{-}29)$$

In the same way, the probability of obtaining a sequence of $t+1$ groups of $(x_i \le C)$ and t of $(x_i > C)$ is

$$\frac{P}{P_1}\left(\frac{P_2Q_1}{Q_2P_1}\right)^t P_1^{r_1} Q_2^{r_2} , \qquad (2\text{-}30)$$

and t groups of $(x_i \le C)$ and $t+1$ of $(x_i > C)$ is

$$\frac{Q}{Q_2}\left[\frac{P_2Q_1}{Q_2P_1}\right]^t P_1^{r_1} Q_2^{r_2} \qquad (2\text{-}31)$$

The joint probability distribution of $2t$ and g is

$$P[2t,g|r_1,r_2] = \frac{\phi(t,t,g)\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)\left(\frac{P_2Q_1}{Q_2P_1}\right)^t}{\sum_t\sum_g\left(\frac{P_2Q_1}{Q_2P_1}\right)^t\{\phi(t,t,g)\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)+\phi(t+1,t,g)\frac{P}{P_1}+\phi(t,t+1,g)\frac{Q}{Q_2}\}} \qquad (2\text{-}32)$$

Similarly for $2t+1$

$$P[2t+1,g|r_1,r_2] = \frac{\left(\phi(t+1,t,g)\frac{P}{P_1}+\phi(t,t+1,g)\frac{Q}{Q_2}\right)\left(\frac{P_2Q_1}{Q_2P_1}\right)^t}{\sum_t\sum_g\left(\frac{P_2Q_1}{Q_2P_1}\right)^t\{\phi(t,t,g)\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)+\phi(t+1,t,g)\frac{P}{P_1}+\phi(t,t+1,g)\frac{Q}{Q_2}\}} , \qquad (2\text{-}33)$$

in which

$$\phi(t,t,g) = f_1(t,g)\binom{r_2-1}{t-1};$$

$$\phi(t+1,t,g) = f_1(t+1,g)\binom{r_2-1}{t-1} ; \text{ and} \qquad (2\text{-}34)$$

$$\phi(t,t+1,g) = f_1(t,g)\binom{r_2-1}{t};$$

$$f_i(t,s) = \sum_{j=1}^{t}(-1)^{j+1}\binom{t}{j}\left\{\binom{r_i-j(s-1)-1}{t-1}-\binom{r_i-js-1}{t-1}\right\} .$$

The probability distribution of g is obtained by summing over all t from $t = 1$ to $r_1 - g+1$, or

$$P[g|r_1,r_2] = \frac{\sum_{t=1}^{r_1-g+1}\{\phi(t,t,g)\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)+\phi(t+1,t,g)\frac{P}{P_1}+\phi(t,t+1,g)\frac{Q}{Q_2}\}\left(\frac{P_2Q_1}{Q_2P_1}\right)^t}{\sum_t\sum_g\left(\frac{P_2Q_1}{Q_2P_1}\right)^t\{\phi(t,t,g)\left(\frac{P}{P_2}+\frac{Q}{Q_1}\right)+\phi(t+1,t,g)\frac{P}{P_1}+\phi(t,t+1,g)\frac{Q}{Q_2}\}} . \qquad (2\text{-}35)$$

Since

$$P[R_1 = r_1] = \binom{r}{r_1} p^{r_1}(1-p)^{r-r_1},$$

then

$$P[G=g] = P[G=g|r_1,r_2]\,P[R_1=r_1]. \qquad (2\text{-}36)$$

To obtain the cumulative distribution function of the longest run of one kind in a series of Markov chain trials, a summation is made from $g = 1$ to $g = g_d$, so that

$$P[g\le g_d|r_1,r_2] = \sum_{g=1}^{g_d} P[G = g|r_1,r_2] ,$$

and

$$P[g\le g_d] = \sum_{r_1=0}^{g_d}\sum_{g=1}^{g_d} P[G = g|r_1,r_2]\,P[R_1=r_1] , \qquad (2\text{-}37)$$

with

$$P = \frac{P_2}{1-P_1+P_2} , \qquad (2\text{-}38)$$

used throughout this development. This condition is arrived at by using the relation

$$P[E_i] = P[E_{i-1}E_i] + P[\bar{E}_{i-1}E_i] ,$$

on the assumption that $P(E_i) = P$ and $P(\bar{E}_i) = Q$ for all i. It is assumed here that the probability of the event E occurring at the i-th trial, when nothing is known about the results of the preceding trials, is independent of i. This in effect implies that the start of the sequence of observations is a randomly selected point in a longer sequence following the same probability laws.

Millan (1972), working independently, obtained the conditioned distribution of the longest run-length in a series of dependent trials (Markov chain type) of size n, making use of the developments of Gabriel (1959) and Whitworth (1896), which are a different approach than the one used in this study, as

$$P[g \le g_d] = \left\{\sum_{s=1}^{n}\sum_{c=1}^{c_1}\frac{L(s,g,a)+L(s,g,a+1)}{\binom{s-1}{a-1}+\binom{s-1}{a}}\binom{s}{a}\binom{n-s-1}{b-1}\cdot\left[\frac{1-P_1}{1-P_2}\right]^b\left[\frac{P_2}{P_1}\right]^a P_1^{\,s}(1-P_2)^{n-s}\right\} P[R_1=r_1]$$

$$+\left\{\sum_{s=0}^{n}\sum_{c=1}^{c_1}\frac{L(s,g,a)}{\binom{s-1}{a-1}}\binom{s-1}{b-1}\binom{n-s}{a}\left[\frac{1-P_1}{1-P_2}\right]^a\cdot\left[\frac{P_2}{P_1}\right]^b P_1^{\,s}(1-P_2)^{n-s}\right\}\cdot P[R_2=r_2] , \qquad (2\text{-}39)$$

in which

$$L(s,m,e) = \sum_{i=0}^{a}(-1)^i\binom{e}{i}\binom{s-m\ i-1}{e-1} , \qquad (2\text{-}40)$$

with

$$a = \min\{e, \left(\frac{s-e}{m}\right)\} ,$$

and

$$s - e + 1 \geq m \geq \left[\frac{s+e-1}{e}\right] .$$

$L(s,m,e)$ represent the number of ways in which s elements can be arranged into e intervals, each of which contains at least one element and the largest of which contains m or less elements. Equation 2-37 becomes, then, the expression for the probability distribution of the longest run of a given kind, say the negative run, in a sample of size n for a simple Markov chain, which also can be used as an approximation for the first-order linear autoregressive models.

2-5 Probabilities of Longest Run-Length in a Sample of Size n for Bivariate Cases

For the two-dimensional or bivariate cases, a similar approach to the one used for univariate series is followed for two series in four alternatives: (1) serially and mutually independent; (2) serially independent but mutually dependent; (3) serially dependent but mutually independent; and (4) both serially and mutually dependent. All four alternatives are studied even though only the second and fourth cases are likely to occur in hydrologic problems. Furthermore, for each of these four alternatives there are four types of run-lengths, as defined previously: negative-negative, negative-positive, positive-negative and positive-positive. Only the negative-negative and the negative-positive run-lengths are treated in this paper, since the other two run-lengths are the opposites to these two types and their properties can be analogously developed.

Bivariate case with serially and mutually independent series. Consider a sequence of a two-dimensional process (X_i, Y_i), $i = 1,2,\ldots,n$, with two series mutually and serially independent, each having the same normal distribution. Given two levels of truncation, C_1 and C_2, the four possible events can be transformed to a new random variable with values 0 or 1 as follows:

$$\begin{aligned} P(X_i \leq C_1 \,,\, Y_i \leq C_2) &= P(X_i' = 1 \,,\, Y_i' = 1) \,, \\ P(X_i \leq C_1 \,,\, Y_i > C_2) &= P(X_i' = 1 \,,\, Y_i' = 0) \,, \\ P(X_i > C_1 \,,\, Y_i \leq C_2) &= P(X_i' = 0 \,,\, Y_i' = 1) \,, \\ P(X_i > C_1 \,,\, Y_i > C_2) &= P(X_i' = 0 \,,\, Y_i' = 0) \,. \end{aligned} \tag{2-41}$$

Since X and Y are mutually independent, the joint probabilities are the product of marginal probabilities, i.e.,

$$P(X_i \leq C_1, Y_i \leq C_2) = P(X_i \leq C_1)\, P(Y_i \leq C_2).$$

For the case of the negative-negative run-length, a new random variable is defined as $Z = X'Y'$, which has a value of 1 only when $X' = 1$ and $Y' = 1$, otherwise its values are zeros. The problem is reduced to obtaining the probability of the longest run-length of ones in n trials of the new random variable Z. The solutions of this case are given by Eqs. 2-15 and 2-19.

Similarly, for the case of the negative-positive run-length, a new random variable is defined as $V = X'(1-Y')$, which has a value of 1 only for $X' = 1$ and $Y' = 0$, otherwise its value is zero. The problem of obtaining the probability of the longest negative-positive run-length in n trials of a bivariate process (X_i, Y_i), whose series are mutually and serially independent, is reduced to the problem of obtaining the longest run-length of ones in n trials of the random variable V.

Instead of a transformation to the univariate process with only two outcomes, an alternative for the case of two series serially and mutually independent is to make the transformation to the univariate process with four outcomes and obtaining the expressions for the longest run-length of one kind following the developments of David and Barton (1962). Consider a series of n trials, with r_i of the i-th kind of a total of four kinds so that $\sum_{i=1}^{4} r_i = n$. David and Barton (1962) give a solution for the probability of the longest run-length irrespective of its kind in a similar manner to obtaining the probability of the longest run of one color in a collection of balls of two colors. Consider a linear array of r_i trials split into t_i groups, none larger than g, for $i = 1,2,3,4$, with all arrangements of the t_i groups of the different kinds, so that no two groups of like type are adjacent. Denoting this number by $C(t_1,t_2,t_3,t_4)$, it is clear that of the $r!/r_1!\, r_3!\, r_4!$ possible arrangements of all the possible trials, the number of arrangements with no run longer than g is

$$G_g(r_1,r_2,r_3,r_4) = \sum_{t_i} C(t_1,t_2,t_3,t_4) \prod_{i=1}^{4} G(r_i,t_i,g), \tag{2-42}$$

with the summation being over all t_i's. It can be recognized that $C(t_1,t_2,t_3,t_4)$ is the coefficient of $x_1^{t_1} x_2^{t_2} x_3^{t_3} x_4^{t_4}$ in the expansion of the expression

$$\frac{1}{j - \sum_{i=1}^{4} \frac{x_i}{1+x_i}}, \tag{2-43}$$

so that the distribution is theoretically obtained. It should be noted also that $G(r_i,t_i,g)$ is the coefficient of x^{t_i} in the expansion of

$$(x + x^2 + \ldots\ldots + x^g)^{t_i} , \tag{2-44}$$

and that

$$\frac{G_g(r_1,r_2,r_3,r_4)}{\frac{r!}{r_1!r_2!r_3!r_4!}} = P\left[\begin{array}{l|l} \text{longest run of} & R_1=r_1, R_2=r_2, \\ \text{either kind} \leq g & R_3=r_3, R_4=r_4 \end{array}\right]. \tag{2-45}$$

An alternative to the computation of the C function is to consider that $G(r_i,t_i,g_D)$ is the coefficient of Z^{r_i} in $[G_g(Z_i)]^{t_i}$ and $G_g(r_1,r_2,r_3,r_4)$ is the coefficient of $Z_1^{r_1} Z_2^{r_2} Z_3^{r_3} Z_4^{r_4}$ in the expansion of

$$\left[1 - \sum_{i=1}^{4} \frac{G_g(Z_i)}{1+G_g(Z_i)}\right] . \tag{2-46}$$

David and Barton report that it is easier to evaluate the C functions.

To obtain the probability of the longest run-length of one kind, conditioned to the knowledge of the total numbers of each of the four kinds, a linear array of the r_i trials split into t_i groups is considered, with t_i not larger than g for $i = 1,2,3,4$. All arrangements of the t_i groups of different kinds are obtained so that no two groups of like kind are adjacent. Denote this number by $C(t_1,t_2,t_3,t_4)$. It is clear that from all the possible $r!/r_1!\, r_2!\, r_3!\, r_4!$ arrangements of all trials, the number with no run longer than g is $G'_g(r_1,r_2,r_3,r_4)$, and is equal to

$$G'_g(r_1,r_2,r_3,r_4)$$

$$= \sum_{t_i} C(t_1,t_2,t_3,t_4)G(r_i,t_i,g) \prod_{J\neq i}^{4} G(r_i,t_i,r_i).$$

With the same definition of $G(r_i,t_i,g)$ as in Eq. 2-42 then

$$P\left[\begin{matrix}\text{longest run of a}\\ \text{given kind} \le g\end{matrix}\,\middle|\, \begin{matrix}R_1=r_1,R_2=r_2,\\ R_3=r_3,R_4=r_4\end{matrix}\right] = \frac{G'_g(r_1,r_2,r_3,r_4)}{\dfrac{r!}{r_1!r_2!r_3!r_4!}} \qquad (2\text{-}47)$$

This alternative has the disadvantage of difficult computations in comparison with the changing variable approach as showed earlier in this text.

Bivariate case of two series serially independent but mutually dependent. Consider a sequence of the bivariate process (X_i,Y_i), $i = 1,2,\ldots,n$ with the series mutually dependent but serially independent following the normal distribution. Given the two levels of truncation, C_1 and C_2, there are four types of run-lengths, similar as earlier stated. Furthermore, since X and Y are mutually dependent, their joint probabilities follow a bivariate normal distribution and can be easily obtained.

As before, the probability of the longest negative-negative run-length in n trials can be obtained by using a new random variable $Z = X'Y'$ and determining the probability of the longest run composed of 1 of the new random variable. Similarly, the probability of the longest negative-positive run-length in n trials can be obtained by using the new random variable $V = X'(1-Y')$, and determining the probability of the longest run of 1 of this new random variable.

Bivariate case of two series serially dependent but mutually independent. As for the case of both series serially and mutually independent, this case can be treated similarly with the only difference that the joint probabilities of X and Y, which are the product of the marginal probabilities

$$P(X_i \le C_1,\ Y_i \le C_2) = P(X_i \le C_1)\ P\ (Y_i \le C_2),$$

take into account the serial dependence by means of

$$P(X_{i+1}\le C_1) = P(X_{i+1}\le C_1 | X_i \le C_1)P(X_i \le C_1)$$

$$+ P(X_{i+1}\le C_1 | X_i > C_1)P(X_i > C_1),$$

and similarily for Y_i. However, the use of a Markov chain instead of Markov models is an approximation, so that the solution for this case is an approximation to the true solution. The approximation is good for values of $\rho \le 0.4$. The probabilities of the longest negative-negative run-length, and the longest negative-positive run-length in n trials are obtained by using the transformed random variable, $Z = X'Y'$ and $V = X'(1-Y')$, respectively.

Bivariate case for two series serially and mutually dependent. The analytical treatment of this case is more complex than for the other three cases. An approximate solution for simple cases is presented here.

Consider a sequence of a bivariate process (X_i,Y_i), $i = 1,2,\ldots,n$, whose series are mutually and serially dependent, each normally distributed. Given the two levels of truncation, C_1 and C_2, the four types of run-lengths can be investigated by using the approximation through a four-state Markov chain, and with the scheme of transition probabilities given in Table 2-1 for X_i and Y_i, or X'_i and Y'_i variables, respectively.

To obtain the transition probabilities of the four-state Markov chain, knowledge is required of the first-order linear autoregressive models, with their parameters ρ_1 and ρ_2, respectively, and the correlation coefficient ρ between X and Y, assuming the distribution of the independent stochastic components are normal.

Table 2-1 Scheme for Transition Probabilities of Four-State Markov Chains of X_i and Y_i, or X'_i and Y'_i.

		$X_{i+1}\le C_1$ or $X'_{i+1}=1$	$X_{i+1}\le C_1$ or $X'_{i+1}=1$	$X_{i+1}> C_1$ or $X'_{i+1}=0$	$X_{i+1}> C_1$ or $X'_{i+1}=0$
		$Y_{i+1}\le C_2$ or $Y'_{i+1}=1$	$Y_{i+1}> C_2$ or $Y'_{i+1}=0$	$Y_{i+1}\le C_2$ or $Y'_{i+1}=1$	$Y_{i+1}> C_2$ or $Y'_{i+1}=0$
$X_i\le C_1$ or $X'_i=1$	$Y_i\le C_2$ or $Y'_i=1$	a_1	a_2	a_3	a_4
$X_i\le C_1$ or $X'_i=1$	$Y_i> C_2$ or $Y'_i=0$	b_1	b_2	b_3	b_4
$X_i> C_1$ or $X'_i=0$	$Y_i\le C_2$ or $Y'_i=1$	c_1	c_2	c_3	c_4
$X_i> C_1$ or $X'_i=0$	$Y_i> C_2$ or $Y'_i=0$	d_1	d_2	d_3	d_4

The feasibility of using the transformed random variables, $Z = X'Y'$ and $V = X'(1-Y')$, requires (1) that the marginal distributions of X and Y be Markov chains, and (2) that the transformed random variables are also Markov chains. Once these requirements are satisfied, it is feasible to use the univariate approximation in determining the probabilities of longest run-length for series serially and mutually dependent. The above requirements can be investigated using the theory on Markov chain lumpability developed by Kemeny and Snell (1960). A lumped process is defined as the process which can be reduced from a process with a large number of states to a process with a small number of states. The disadvantage is that lumpability conditions are very restrictive and could be applied only in a few cases.

Given an r-states Markov chain with transition matrix P, let $A = (A_1, A_2, \ldots, A_t)$ be a partition of the set of states. Also let $p_{iA_j} = \sum_{k \epsilon A_j} p_{ik}$ represent the probability of moving from state s_i into set A_j in one step for the original Markov chain. Then, a necessary and sufficient condition for a Markov chain to be lumpable with respect to a partition $A = (A_1, A_2, \ldots, A_s)$ is that for every pair of sets A_i and A_j, p_{kA_j} must have the same value for every s_k in A_i.

For a Markov chain to be lumpable and to obtain the lumped transition matrix, the following procedure may be followed. Assume that the original Markov chain with transition matrix P has r states, while the desired lumped chain has s states, with $s < r$. Let U be a $s \times r$ matrix whose i-th row is the probability vector having equal components for states in A_i, and 0 for the remaining states. Also let V be a $r \times s$ matrix with the j-th column a vector with value unity in the components corresponding to states in A_j and 0 otherwise. If the Markov chain with transition matrix P is lumpable with respect to the partition A, then the following condition needs to be satisfied (Kemeny and Snell, 1960)

$$VUPV = PV \ . \tag{2-48}$$

The lumped transition matrix is given by

$$\hat{P} = UPV \ . \tag{2-49}$$

For the case of investigating the lumpability conditions for the process X of Table 2-1, then

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & 0 & 0 \\ 0 & 0 & u_{23} & u_{24} \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ d_1 & d_2 & d_3 & d_4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \tag{2-50}$$

For X to be Markov chain, the four-state Markov chain must satisfy the four conditions:

$$\begin{aligned} &a_1 + a_2 = b_1 + b_2 \ , \qquad a_3 + a_4 = b_3 + b_4 \ , \\ &c_1 + c_2 = d_1 + d_2 \ , \text{ and } \quad c_3 + c_4 = d_3 + d_4 \ . \end{aligned} \tag{2-51}$$

Similarly, for Y of Table 2-1 to be a Markov chain, the four-state Markov chain must satisfy the four conditions:

$$\begin{aligned} &a_1 + a_3 = c_1 + c_3 \ , \qquad a_2 + a_4 = c_2 + c_4 \ , \\ &b_1 + b_3 = d_1 + d_3 \ , \text{ and } \quad b_2 + b_4 = d_2 + d_4 \ . \end{aligned} \tag{2-52}$$

For the transformed random variable $Z = X'Y'$ to be a Markov chain, the four-state Markov chain must satisfy

$$a_4 = b_4 = c_4 \ . \tag{2-53}$$

Similarly, for the transformed random variable $V = X'(1-Y')$ the conditions are

$$a_3 = b_3 = d_3 \ . \tag{2-54}$$

Another way of approaching the problem of a sequence of a bivariate process $(X_i Y U_i)$, $i = 1,2,\ldots,n$, for the two series mutually and serially dependent, is by considering the marginal distributions of each process. For the process X, the corresponding Markov chain has the scheme of transition probabilities given in Table 2-2.

Table 2-2 Scheme of Transition Probabilities of Markov Chain for the Process X_i.

	$X_{i+1} \le C_1$, or $X'_{i+1}=1$	$X_{i+1} > C_1$, or $X'_{i+1}=0$
$X_i \le C_1$, or $X'_i=1$	p_{11}	p_{10}
$X_i > C_1$, or $X'_i=0$	p_{01}	p_{00}

For the process Y, the corresponding scheme of transition probabilities of the Markov chain are given in Table 2-3.

Table 2-3 Scheme of Transition Probabilities of Markov Chain for the Process Y_i.

	$Y_{i+1} \le C_2$, or $Y'_{i+1}=1$	$Y_{i+1} > C_2$, or $Y'_{i+1}=0$
$Y_i \le C_2$, or $Y'_i=1$	q_{11}	q_{10}
$Y_i > C_2$, or $Y'_i=0$	q_{01}	q_{00}

Furthermore, the joint probabilities can be obtained either by using a table of bivariate normal distribution or by integration as

$$\begin{aligned} P[X_i > C_1 \ , \ Y_i > C_2] &= P[X'_i = 0 \ , \ Y'_i = 0] = P_{00} \ , \\ P[X_i > C_1 \ , \ Y_i \le C_2] &= P[X'_i = 0 \ , \ Y'_i = 1] = P_{01} \ , \\ P[X_i \le C_1 \ , \ Y_i > C_2] &= P[X'_i = 1 \ , \ Y'_i = 0] = P_{10} \ , \\ P[X_i \le C_1 \ , \ Y_i \le C_2] &= P[X'_i = 1 \ , \ Y'_i = 1] = P_{11} \ . \end{aligned} \tag{2-55}$$

The matrices of transition probabilities for X_i and Y_i are obtained by means of

$$P[X'_{i+1} = k | X'_i = j] = \frac{P[X'_{i+1} = k, X'_i = j]}{P[X'_i = j]} . \quad (2\text{-}56)$$

To obtain the probability of the longest negative-negative run-length in n trials in this bivariate process, the new random variable $Z = X'Y'$ is expected to be also a Markov chain, with the scheme of transition probabilities given in Table 2-4.

Table 2-4 Scheme of Transition Probabilities of Markov Chain for the Process Z_i.

	$Z_{i+1} = 0$	$Z_{i+1} = 1$
$Z_i = 0$	A_1	A_2
$Z_i = 1$	B_1	B_2

The transition probabilities are obtained as

$$A_1 = P[Z_{i+1} = 0 | Z_i = 0] = 1 - P[Z_{i+1} = 1 | Z_i = 0]$$

$$= 1 - \frac{P[Z_{i+1}=1, Z_i=0]}{P[Z_i = 0]} = 1 - \frac{P[Z_{i+1}=1] - P[Z_{i+1}=1, Z_i=1]}{1-P[Z_i=1]}$$

$$= \frac{1-P[Z_i=1]-P[Z_{i+1}=1] + P[Z_{i+1} = 1 | Z_i=1]\ P[Z_i=1]}{1-P[Z_i=1]}$$

$$= \frac{1-P[X'_i=1, Y'_i=1]-P[X'_{i+1}=1, Y'_{i+1}=1]+P[X'_{i+1}=1, Y'_{i+1}=1, X'_i=1, Y'_i=1]}{1-P[X'_i=1,\ Y'_i=1]} . \quad (2\text{-}57)$$

The four-variate joint probability of Eq. 2-57 can be obtained by integrating the quadrivariate normal distribution for the parameters of the underlying model. Then $A_2 = 1 - A_1$. The probability B_1 can be obtained similarly by

$$B_1 = 1 - \frac{P[X'_{i+1} = 1, Y'_{i+1} = 1, X'_i = 1, Y'_i = 1]}{P[X'_i = 1, Y'_i = 1]} , \quad (2\text{-}58)$$

with $B_2 = 1 - B_1$.

With the transition probabilities of Z determined, the probability of the longest negative-negative run-length in n trials can be obtained by using these transition probabilities and the expressions developed for the univariate dependent case, Eqs. 2-35 and 2-37.

To obtain the probability of the longest negative-positive run-length in n trials, the new random variable $V = X'(1-Y')$ is expected also to be also a Markov chain. Its scheme of transition probabilities are shown in Table 2-5.

In a similar way to transition probabilities of Z, the transition probabilities of V are obtained as

$$F_1 = \frac{1-P[X'_i = 1, Y'_i=0]-P[X'_{i+1}=1, Y'_{i+1}=0]+P[X'_{i+1}=1, Y'_{i+1}=0,\ X'_i=1,\ Y'_i=0]}{1-P[X'_i=0,\ Y'_i = 0]} , \quad (2\text{-}59)$$

Table 2-5 Scheme of Transition Probabilities for Markov Chain of the Process V.

	$V_{i+1} = 0$	$V_{i+1} = 1$
$V_i = 0$	F_1	F_2
$V_i = 1$	G_1	G_2

and

$$G_1 = 1 - \frac{P[X'_{i+1}=1,\ Y'_{i+1}=0,\ X'_i=1,\ Y'_i=0]}{P[X'_i = 1,\ Y'_i = 0]} , \quad (2\text{-}60)$$

with $F_2 = 1-F_1$, and $G_2 = 1-G_1$.

With the transition probabilities of V known, the probabilities of the longest negative-positive run-length in n trials are obtained by using these transition probabilities and the expressions developed for the univariate dependent case, Eqs. 2-35 and 2-37.

2-6 Integration of Quadrivariate Normal Distribution

To integrate the quadrivariate normal distribution function, as needed for Eqs. 2-57 through 2-60, consider a multivariate n-dimensional stationary Gaussian process whose distribution is

$$dF = \frac{1}{(2\pi)^{n/2}\sqrt{|R|}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} X_j X_k \right\} \prod_{j=1}^{n} dx_j , \quad (2\text{-}61)$$

with all components $X_1, X_2, \ldots, X_n$ having zero mean and unit variance, $|R|$ the determinant of the correlation matrix of these components, and a_{jk} the elements of the inverse of this correlation matrix. Reviewing the literature on integration of the multivariate normal function, presented by Saldarriaga (1969, 1970), it was found that no explicit expression is available for the general solution of this integral. Solutions exist only for special cases. Saldarriaga (1969, 1970) gave a solution following Kendall (1941) for the probability of run-length for an infinite population in the case of dependent univariate case.

The tetrachoric series expansion, for the trivariate case, is given by Kendall (1941). It was extended by this writer to the quadrivariate case under study. For simplicity, the following notation is used: the first X_i in the first series is designated by 1, the second X_{i+1} in the first series is designated by 2, the first Y_i in the second series by 3, and the second Y_{i+1} in the second series by 4. Also whenever the integration goes from the truncation level to infinity the index (+) will be given to the corresponding variable, and (-) for integration from $-\infty$ to this level. The truncation levels are C_1 and C_2, respectively. Since the underlying model is a bivariate first-order linear autoregressive model, the serial correlation coefficients are ρ_1 and ρ_2, respectively for series 1 and 2, and ρ_c is their cross correlation coefficient. Then

$$P(1^+, 2^+, 3^+, 4^+) = \int_{C_1}^{\infty} \int_{C_1}^{\infty} \int_{C_2}^{\infty} \int_{C_2}^{\infty} dF . \quad (2\text{-}62)$$

Making use of Saldarriaga's developments, this expression can be evaluated by using Hermite polynomials

$$P[1^+,2^+,3^+,4^+] = f^2(C_1)\, f^2(C_2) \sum_{i=0}^{\infty} A(\rho,i)\Pi(H), \quad (2\text{-}63)$$

in which

$$A(\rho,i) = \frac{\rho_{12}^{\ell}\rho_{13}^{m}\rho_{14}^{n}\rho_{23}^{p}\rho_{24}^{q}\rho_{34}^{r}}{\ell!m!n!p!q!r!}, \quad (2\text{-}64)$$

$$i = \ell + m + n + p + q + r, \quad (2\text{-}65)$$

and

$$\Pi(H) = H_{S_1-1}(C_1)\, H_{S_2-1}(C_1)\, H_{S_3-1}(C_2)\, H_{S_4-1}(C_2), \quad (2\text{-}66)$$

with

$$\begin{aligned} S_1 &= i_{12} + i_{13} + \ldots + i_{1n} = \ell + m + n \\ S_2 &= i_{12} + i_{13} + \ldots + i_{2n} = \ell + p + q \\ &\ldots\ldots\ldots \\ &\ldots\ldots\ldots \\ S_n &= i_{1n} + i_{2n} + \ldots + i_{n-1,n}. \end{aligned} \quad (2\text{-}67)$$

Since the definition of the Hermite polynomials applies only to $r = 0,1,2,\ldots$, its values are

$$\begin{aligned} H_o(C_i) &= 1, \\ H_1(C_i) &= C_i, \\ H_2(C_i) &= C_i^2 - 1, \\ H_3(C_i) &= C_i^3 - 3C_i. \end{aligned} \quad (2\text{-}68)$$

For the case of $r = -1$, $H_{-1}(C_i)$ is defined by

$$H_{-1}(C_i) = \frac{1-F(C_i)}{f(C_i)}. \quad (2\text{-}69)$$

Equation 2-63 is an infinite series. However, in practical applications it is desirable to restrict the series to a few terms. A truncation of the series after $i = 2$ is used in this study, implying that terms of the third order and higher orders are negligible. The error introduced by the polynomial truncation is negligible for small values of ρ_i.

After developments and simplifications, Eq. 2-63 becomes

$$\begin{aligned} P[1^+,2^+,3^+,4^+] = f^2(C_1)f^2(C_2)\Bigg\{ & \left[\frac{1-F(C_1)}{f(C_1)}\right]^2\left[\frac{1-F(C_2)}{f(C_2)}\right]^2 + \rho_1\left[\frac{1-F(C_1)}{f(C_1)}\right]^2 \\ & + \left[\frac{1-F(C_1)}{f(C_1)}\right]\left[\frac{1-F(C_2)}{f(C_2)}\right]\rho(2+\rho_1+\rho_2) + \rho_2\left[\frac{1-F(C_1)}{f(C_1)}\right]^2 + \frac{1}{2}\left[\rho_1^2C_1^2\left[\frac{1-F(C_2)}{f(C_2)}\right]^2\right. \\ & + C_1C_2\left[\frac{1-F(C_1)}{f(C_1)}\right]\left[\frac{1-F(C_2)}{f(C_2)}\right][2+\rho_1^2\rho_2^2]\rho^2 + \rho_2^2\left[\frac{1-F(C_1)}{f(C_1)}\right]^2C_2^2 + \rho_1\rho_2 \\ & + \rho^2 + \rho^2C_1\rho_2 + C_1\left[\frac{1-F(C_2)}{f(C_2)}\right][2\rho_1\rho+\rho_1^2\rho+\rho_1\rho_2\rho] + C_2\left[\frac{1-F(C_1)}{f(C_1)}\right][2\rho_2\rho+\rho_2^2\rho+\rho_2\rho_1\rho] \\ & \left. + C_1\left[\frac{1-F(C_1)}{f(C_1)}\right][\rho^2\rho_1+\rho^2\rho_2] + C_2\left[\frac{1-F(C_2)}{f(C_2)}\right][\rho^2\rho_1+\rho^2\rho_2]\right\} \end{aligned} \quad (2\text{-}70)$$

The transformations that are used later are

$$P[1^-,2^+,3^+,4^+] = P[2^+,3^+,4^+] - P[1^+,2^+,3^+,4^+], \quad (2\text{-}71)$$

$$P[1^+,2^-,3^+,4^+] = P[1^+,3^+,4^+] - P[1^+,2^+,3^+,4^+], \quad (2\text{-}72)$$

$$P[1^+,2^+,3^-,4^+] = P[1^+,2^+,4^+] - P[1^+,2^+,3^+,4^+], \quad (2\text{-}73)$$

$$P[1^+,2^+,3^+,4^-] = P(1^+,2^+,3^+] - P[1^+,2^+,3^+,4^+], \quad (2\text{-}74)$$

and the necessary probabilities are obtained as differences of probabilities of four trivariate cases and one quadrivariate case. The definite expressions are of the same length as Eq. 2-70 and are obtained in the same way.

To obtain the probability of all four variables being negatives, the same procedure with the following changes is used, namely for

$$P[1^-,2^-,3^-,4^-] = \int_{-\infty}^{C_1}\int_{-\infty}^{C_1}\int_{-\infty}^{C_2}\int_{-\infty}^{C_2} dF, \quad (2\text{-}75)$$

by using the tetrachoric series expansion as,

$$P[1^-,2^-,3^-,4^-] = f^2(C_1)f^2(C_2) \sum_{i=0}^{\infty} A(\rho,i)\, \pi^C(H), \quad (2\text{-}76)$$

in which $A(\rho,i)$ is the same term as defined by Eq. 2-64

$$\pi^C(H) = [-H_{S_1-1}(C_1)][-H_{S_2-1}(C_1)][-H_{S_3-1}(C_2)][-H_{S_4-1}(C_2)]. \quad (2\text{-}77)$$

For this case, the negative Hermite polynomials are:

$$\begin{aligned} -H_{-1}(C_i) &= \frac{F(C_i)}{f(C_i)}, \\ -H_o(C_i) &= -1, \\ -H_1(C_i) &= -C_i. \end{aligned} \quad (2\text{-}78)$$

The truncation of the expansion of Eq. 2-76 is also made after $i = 2$, with the corresponding error involved. With the above considerations, Eq. 2-76 becomes

$$\begin{aligned} P[1^-,2^-,3^-,4^-] = f^2(C_1)f^2(C_2)\Bigg\{ & \left[\frac{F(C_1)}{f(C_1)}\right]^2\left[\frac{F(C_2)}{f(C_2)}\right]^2 + \rho_1\left[\frac{F(C_2)}{f(C_2)}\right]^2 \\ & + \left[\frac{F(C_1)}{f(C_1)}\right]\left[\frac{F(C_2)}{f(C_2)}\right][2\rho+\rho\rho_1+\rho\rho_2] + \rho_2\left[\frac{F(C_1)}{f(C_1)}\right]^2 + \frac{1}{2}\left[\rho_1^2C_1^2\left[\frac{F(C_2)}{f(C_2)}\right]^2\right. \\ & + C_1C_2\left[\frac{F(C_1)}{f(C_1)}\right]\left[\frac{F(C_2)}{f(C_2)}\right][2\rho^2+\rho\rho_1^2+\rho\rho_2^2] + \rho_2^2C_2^2\left[\frac{F(C_1)}{f(C_1)}\right]^2\Bigg] + \rho\rho_1+\rho^2+\rho^2\rho_1\rho_2 \\ & - C_1\left[\frac{F(C_2)}{f(C_2)}\right][2\rho_1\rho+\rho_1^2\rho+\rho\rho_2\rho_1] - C_2\left[\frac{F(C_1)}{f(C_1)}\right][2\rho_2\rho+\rho_2^2\rho+\rho_1\rho_2\rho] \\ & - [\rho^2\rho_1+\rho^2\rho_2]\left[-C_1\left[\frac{F(C_1)}{f(C_1)}\right] - C_2\left[\frac{F(C_2)}{f(C_2)}\right]\right]\Bigg\}. \end{aligned} \quad (2\text{-}79)$$

Similar transformations to Eq. 2-71 through 2-74 are used for the negative case, namely:

$$P(1^+,2^-,3^-,4^-) = P(2^-,3^-,4^-) - P(1^-,2^-,3^-,4^-), \quad (2\text{-}80)$$

$$P(1^-,2^+,3^-,4^-) = P(1^-,3^-,4^-) - P(1^-,2^-,3^-,4^-), \quad (2\text{-}81)$$

$$P(1^-,2^-,3^+,4^-) = P(1^-,2^-,4^-) - P(1^-,2^-,3^-,4^-), \quad (2\text{-}82)$$

$$P(1^-,2^-,3^-,4^+) = P(1^-,2^-,3^-) - P(1^-,2^-,3^-,4^-). \quad (2\text{-}83)$$

The final expressions for Eqs. 2-80 through 2-83 are similar to those of Eq. 2-79.

To obtain $P(1^-,2^-,3^+,4^+)$, the procedure is similar as followed in previous cases, namely

$$P[1^-,2^-,3^+,4^+] = \int_{-\infty}^{C_1}\int_{-\infty}^{C_1}\int_{C_2}^{\infty}\int_{C_2}^{\infty} dF$$

$$= f^2(C_1)f^2(C_2)\sum_{i=0}^{\infty} A(\rho,i)\pi^C(H_{C_1})\Pi(H_{C_2}), \quad (2\text{-}84)$$

in which $A(\rho,i)$, $\pi^C(H)$, $\Pi(H)$ are defined above. After replacing terms and simplifying,

$$P[1^-,2^-,3^+,4^+] = f^2(C_1)f^2(C_2)\left\{\left[\frac{F(C_1)}{f(C_1)}\right]^2\left[\frac{1-F(C_2)}{f(C_2)}\right]^2\right.$$

$$+ \rho_1\left[\frac{1-F(C_2)}{f(C_2)}\right]^2 - \left[\frac{F(C_1)}{f(C_1)}\right]\left[\frac{1-F(C_2)}{f(C_2)}\right]\rho(2+\rho_1+\rho_2)$$

$$+ \rho_2\left[\frac{F(C_1)}{f(C_1)}\right]^2 + \frac{1}{2}\left[\rho_1^2 C_1^2\left[\frac{1-F(C_2)}{f(C_2)}\right] - C_1\left[\frac{F(C_1)}{f(C_1)}\right]C_2\left[\frac{1-F(C_2)}{f(C_2)}\right]\right.$$

$$\rho^2(2+\rho_1^2+\rho_2^2) + \rho_2^2\left[\frac{F(C_1)}{f(C_1)}\right]^2 C_2^2\Big] + \rho_1\rho_2 + \rho^2 + \rho^2\rho_1\rho_2$$

$$+ C_1\left[\frac{1-F(C_2)}{f(C_2)}\right]\rho\rho_1(2+\rho_1+\rho_2) - \left[\frac{F(C_1)}{f(C_1)}\right]C_2\rho_2\rho(2+\rho_2+\rho_1)$$

$$\left.-\left[\frac{F(C_1)}{f(C_1)}\right]C_1\rho^2(\rho_1+\rho_2) + C_2\left[\frac{1-F(C_2)}{f(C_2)}\right]\rho^2(\rho_2+\rho_1)\right\}. \quad (2\text{-}85)$$

Similar expressions are obtained for $P(1^-,2^+,3^-,4^+)$, $P(1^+,2^+,3^-,4^-)$, $P(1^+,2^-,3^+,4^-)$, $P(1^+,2^-,3^-,4^+)$, and $P(1^-,2^+,3^+,4^-)$.

The sixteen expressions thus far developed for the joint probabilities were programmed for computation by a digital computer, and the transition probabilities of Table 2-1 were computed.

For testing the accuracy, the bivariate case is used, with the approximation available in the Handbook of Mathematical Functions (Abramowitz, 1955) for the bivariate standard normal distribution, namely

$$L(C_1,C_2,\rho) = Q(C_1)Q(C_2) + \sum_{i=1}^{\infty}\frac{Z^i(C_1)Z^i(C_2)}{i!}\rho^i, \quad (2\text{-}86)$$

in which

$$Z^n(C) = \frac{d^n}{dx^n}\,Z(x) = \frac{d^n}{dx}\left(\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right).$$

Although Eq. 2-86 is an infinite series, for this test the series was truncated at $i = 5$ and the approximation was compared with the values given in Tables for the bivariate normal distribution, finding the approximation adequate. For the case of $\rho_{1,2} > 0.4$, i should be greater than five. The bivariate case obtained from expressions for the sixteen joint probabilities were compared with bivariate case obtained by the above approximation. It was found that the approximation presented for the quadrivariate normal case was good for each ρ_1, ρ_2, and ρ less than 0.4. When ρ_i was greater than 0.4, it is noted that the row values of transition probabilities do not add up to unity.

Therefore, the approximations give the probabilities in case of the quadrivariate normal distribution, that are valid only for small values of ρ, ρ_1 and ρ_2. These results may be used to evaluate results obtained by Eqs. 2-57 through 2-60 as well as to obtain the transition probabilities of Table 2-1, thus producing approximations for probabilities of the longest negative-negative and the longest negative-positive run-length in n trials for the case of mutually and serially dependent series of the normal bivariate process.

2-7 Probabilities of Largest Run-Sums in a Sample of Size n.

The analytical treatment and necessary approximations for probabilities of the largest run-sum in n trials are presented here in the general form. The case of the largest run-sum not exceeding a given value in n trials is more complex than for the case of longest run-length of a univariate independent normal process. It is still more complex for a univariate dependent process. The run-sum case of the bivariate independent or dependent process is expected to be also very complex. The problem of probabilities of run-sum has not received very much attention in statistical literature. The problems of hydrologic droughts have stimulated this type of study recently. Because of various problems involved, the simple case of univariate independent standard normal process is used first in studying here in the probability of largest run-sum in samples of given sizes. The first study of the run-sum seems to have been done by Downer, Siddiqqi and Yevjevich (1967) for the run-sum distribution of an infinite stationary and ergodic population.

The following random variables are of interest in hydrology for the case of a univariate process, as defined by Millan and Yevjevich (1971):

- L_n, the longest negative run-length in n trials,
- S_n, the largest negative run-sum in n trials,
- $L_{n,s}$, the negative run-length corresponding to the largest negative run-sum in n trials and,
- $S_{n,\ell}$, the negative run-sum corresponding to the longest negative run-length in n trials, as well as,

the ratios $S_{n,\ell}/L_n$ and $S_n/L_{n,s}$ as measures of drought severity. First, the largest negative run-sum in n trials is investigated.

For the case of the uni-dimensional case of independent, identically distributed normal random variables, a truncation level is selected so that

$$q = P[X_i \leq C] = P[X_i' = 1] = F(C), \tag{2-87}$$

and

$$p = 1 - q = 1 - F(C) .$$

A negative run-sum corresponding to a negative run-length of size n is defined by

$$S_n = \sum_{i=1}^{n} (C - X_i) , \tag{2-88}$$

while the largest negative run-sum is

$$S_n = \max[\max(S_1), \max(S_2), \ldots, \max(S_\ell)], \tag{2-89}$$

where $\max(S_\ell) = S_{n,\ell}$ in the above notation.

The largest run-sum in n trials, S_n, is obtained from a_1 negative runs of length 1, a_2 negative runs of length 2, up to a_ℓ negative runs of length ℓ, where ℓ is the longest negative run-length in these n trials. Then $\sum_{i=1}^{\ell} a_i$ is the total number of negative runs in n trials. The maximum or largest run-sum of each of the run-lengths, $i = 1,2,\ldots,\ell$, is obtained and the maximum amongst them is the largest run-sum in n trials. Let define

$$P[X_i^* \leq x] = F^*(x) = \frac{F(C)-F(x)}{F(C)} = 1 - \frac{F(x)}{F(C)} , \text{ for } x \leq C, \tag{2-90}$$

with

$$P[X_i^* \leq x] = F^*(x) = 0, \text{ for } x > C, \tag{2-91}$$

where $F^*(x)$ is the truncated normal cumulative distribution function of X_i. The following notation is adopted, following Downer et al., (1969): M^* is the moment generating function of X_i; $\kappa^* = \log M^*$ is the cumulants generating function, and κ_r^* is the r-th cumulant of X_i, so that

$$M^*(v) = E[e^{vX_i}] = \int_{-\infty}^{\infty} e^{vx} dF^*(x) = -\frac{1}{q}\int_{-\infty}^{C} d^{vx} dF(x) , \tag{2-92}$$

with

$$\kappa^*(v) = \log M^*(v) = \log \sum_x p_x e^{vx} . \tag{2-93}$$

After replacements and simplifications, Eq. 2-93 becomes

$$\kappa^*(v) = \sum_{r=1}^{\infty} \frac{v^r}{r!} \mu_r' - \frac{1}{2}\left(\sum_{r=1}^{\infty} \frac{v^r}{r!} \mu_r'\right)^2 + \ldots = \sum_{r=1}^{\infty} \frac{v^r}{r!} \kappa_r^* , \tag{2-94}$$

as the cumulants generating function. In particular, equating like powers of $\frac{v^r}{r!}$ then

$$\kappa_1^* = \mu_1' = E[X] ,$$

$$\kappa_2^* = \mu_2' - \mu_1^2 = \mu_2 = \mathrm{Var}[X] ,$$

and

$$\kappa_3^* = \mu_3' - 2\mu_2' \mu_1' + 2\mu_1'^3 = \mu_3 . \tag{2-95}$$

Since for this approach knowledge of the distribution of the run-length is required and the distribution of the run-length corresponding to the largest run-sum is not known, then only the distribution of the run-sum corresponding to the longest run-length is looked for analytically.

Let $M(\mu,v)$ and $\kappa(\mu,v)$ be the joint moments and cumulants generating function of the longest negative run-length in n trials, g_d, and the corresponding run-sum, $S_{n,\ell}$. Since the joint moment generating function is defined for any two real numbers μ and v by

$$m_{G_n,S_n}(\mu,v) = E\,[e^{\mu G_n + vS_n}] = E\,\{[e^{\mu G_n}]\; E\,[e^{vS_n/G_n}]\} , \tag{2-96}$$

and making use of the expressions developed for the longest run-length in n trials earlier, Eqs. 2-12, 2-15 and 2-17,

$$m_{G_n,S_n}(\mu,v) = \sum_{g_d} \left\{ e^{\mu g_d}\; P[G_d = g_d]\; E\left[e^{rS_g | G_d = g_d} \right] \right\}. \tag{2-97}$$

Since the moment generating function of the sum of independent random variables is equal to the product of their moment generating functions, then

$$E[e^{vS_g}] = [M^*(v)]^{g_d} = e^{g_d \log M^*(v)} = e^{g_d\, \kappa^*(v)} , \tag{2-98}$$

or

$$m_{G_n,S_n}(\mu,v) = \left\{ \sum_{g_d} e^{\mu g_d} P[G_d = g_d] \right\} e^{g_d\, \kappa^*(v)} , \tag{2-99}$$

and

$$\kappa_{G_n,S_n}(\mu,v) = \log m_{G_n,S_n}(\mu,v). \tag{2-100}$$

The individual cumulant generating functions can be obtained by

$$\kappa_{G_n}(\mu) = \kappa_{G_n,S_n}(\mu,0), \tag{2-101}$$

and

$$\kappa_{S_n}(v) = \kappa_{G_n,S_n}(0,v). \tag{2-102}$$

The parameters of these distributions should be obtainable by differentiating at the origin. However, because the joint moment generating function cannot be reduced to a simple and recognizable expression, they cannot be obtained easily. Faced with these difficulties, the investigator can only use the experimental

method to obtain the required results. The purpose of the above development was to show that even in the case of run-sums in n trials for a univariate independent identically and normally distributed random variables is not simple. Therefore, the more complex cases of univariate dependent, and the two-dimensional independent and dependent cases, being still more complex, do not yield themselves to easy analytical solutions. The experimental statistical method seems the only alternative left at present, and it will be used in all these cases in the further text.

2-8 Run-Length Distributions for Infinite Populations of Univariate Cases

The runs of an infinite population are studied similarly as for the runs of given sample sizes.

Univariate independent process. The distribution of run-lengths of a uni-dimensional sequence of independent identically distributed normal variables is the same as the distribution of the number of trials required to obtain the first success in a sequence of repeated independent Bernoulli trials in which the probability of success at each trial is a constant, p. This distribution follows the well known geometric probability function. Downer, Siddiqqi, and Yevjevich (1967) studied the distribution of the positive and negative run-lengths and applied it to the normal variable. They used the data generation method to check the analytical solutions developed for the independent standard normal variable. For this distribution, the constant truncation level C of the stochastic process X_i was replaced by its probability

$$q = P[x_i \leq C], \text{ with } p = 1-q = P[x_i > C]. \quad (2\text{-}103)$$

By this replacement the properties of run-length when expressed as function of the probability q become distribution free, or independent of the underlying distribution, $F(x)$. Therefore, the probability distribution function of run-length becomes

$$P[K=k] = P[X_1 \leq C, X_2 \leq C, \ldots, X_k \leq C, X_{k+1} > C | X_1 \leq C] P[X_1 \leq C] + \sum_{j=1}^{\infty} P[X_i > C, i = 1,\ldots,j; X_{j+i} \leq C, i = 1,\ldots,k; X_{j+k+1} > C | X_1 > c].$$

$$P[X_1 > C],$$

or

$$P[K=k] = pq^{k-1}, \quad (2\text{-}104)$$

with $E[K] = 1/p$, and $Var[K] = q/p^2$.

Llamas (1968) studied the case of the standardized, one-parameter gamma independent random variable, with the probability distribution function

$$F(x) = \int_{-\sqrt{\alpha}}^{x} \frac{\alpha(\alpha+t\sqrt{\alpha})^{\alpha+1}}{\Gamma(\alpha)} e^{-\alpha-t\sqrt{\alpha}}\, dt , \quad (2\text{-}105)$$

which, for the truncation level $C = 0$, gives $F(0) = P(\alpha,\alpha) = p$, where $P(\alpha,\alpha)$ is the incomplete gamma function or

$$I[\Gamma(\alpha)] = \frac{1}{\Gamma(\alpha)} \int_0^{\alpha} e^{-t}\, t^{\alpha-1} dt . \quad (2\text{-}106)$$

Gabriel and Neumann (1957), in studying the distribution of a weather cycle, investigated the distribution of the total run-length (the negative run-length plus the continuing positive run-length) by assuming that the negative run-length is independent of the positive run-length, and that both follow the geometric distribution. For X representing a positive run-length and Y a negative run-length, each having the positive integer events as discrete random variables, their probabilities are

$$P[X = k] = q_1(p_1)^{k-1} = (1-p_1)(p_1)^{k-1} ,$$
$$P[Y = m] = q_2(p_2)^{m-1} = (1-p_2)(p_2)^{m-1} , \quad (2\text{-}107)$$

with $p_1 + q_1 = p_2 + q_2 = 1$, and $p_1 \neq p_2$, $q_1 \neq q_2$, where p_1 is the probability of a positive value to be followed by a positive value, p_2 is the probability of a negative value followed by a negative value, q_1 represents the probability of a positive value to be followed by a negative value, and q_2 the probability of a negative value to be followed by a positive value. Even though the run is not defined in the paper, it is easy to infer that the definition of runs given by Feller (1957) was used. Let Z represent the total run-length, equal to $X + Y$. Then its probability is

$$P[Z = n] = (n-1)(1-p)^2 p^{n-2} \quad (2\text{-}108)$$

Univariate dependent process. For the univariate case with a dependent series, the distribution of the run-length has been obtained in two different ways. First, by approximating the dependent series of the first-order linear autoregressive model by the corresponding Markov chain. Second, by using a truncation on the infinite series of the tetrachoric series expansion of the integral when the underlying process is normal.

The first approach was used by Cox and Miller (1965) giving the distribution of the recurrence time of state (0) in the two-state Markov chain, (0) and (1), with the transition probability matrix

	$X_{i+1} = 0$	$X_{i+1} = 1$
$X_i = 0$	$1 - \alpha$	α
$X_i = 1$	β	$1 - \beta$

This distribution is equal to the run-length of state (1) plus unity, presented as

$$P[K=k] = \alpha\beta(1-\beta)^{k-2}, \text{ for } k = 2,3,\ldots, \quad (2\text{-}109)$$

and

$$P[K=k] = 1 - \alpha, \text{ for } k = 1 . \quad (2\text{-}110)$$

The mean recurrence time of state (0) is

$$E[K] = \frac{\alpha+\beta}{\beta} . \quad (2\text{-}111)$$

Heiny (1970) defines the transition probabilities of two states as

$$P[X_i > C | X_{i-1} > C] = r,$$

and

$$P[X_i \le C | X_{i-1} > C] = s \ ,$$

with $r + s = 1$. With the Markov chain approximation to the first-order linear autoregressive model, probabilities, the expected value and the variance are

$$P[K=k] = sr^{k-1}[1 + 0\ (\rho^2)],\ k = 1,2,\ldots, \quad (2\text{-}112)$$

$$E[K] = \frac{1}{s}\ [1 + 0\ (\rho^2)], \quad (2\text{-}113)$$

and

$$Var\ [K] = \frac{r}{s^2}\ [1 + 0\ (\rho^2)], \quad (2\text{-}114)$$

with $0(\rho^2)$ the error term becoming negligible for small values of ρ.

The second approach in considering the first-order linear autoregressive model is studied by Saldarriaga (1969, 1970). For this type of univariate dependent process, the development of the distribution of run-length requires the joint probability distribution of variables $X_1, X_2,\ldots,$ assumed by Saldarriaga to be multivariate normal. For example, to find the probability of the negative run-length J^- it is necessary to integrate

$$P[J^-] = P[X_1 \le C,\ X_2 \le C,\ldots,X_j \le C] = \int_{-\infty}^{C}\int_{-\infty}^{C}\ldots\int_{-\infty}^{C} dF,$$

and

$$P[K \ge k] = P[J^-] + \sum_{k=1}^{\infty} P[k^+,\ J^-]$$

with dF given by Eq. 2-61. The general solution of the multivariate normal integral is not available, except through the tetrachoric series expansion, given by Kendall (1941), for finding the approximations to exact solutions. The use of the experimental statistical (Monte Carlo) method permitted to check the above approximations, which were presented by Saldarriaga (1969, 1970) in the form of graphs and tables.

2-9 Run-Length Distributions for Infinite Populations for the Bivariate Case

Similar as for the bivariate case of the probability distribution of the longest run-length for a given sampie size, the same four alternatives are investigated for the probability distribution of the run-length of infinite series for the bivariate case: (1) series are serially and mutually independent, (2) series are serially independent but mutually dependent, (3) series are serially dependent but mutually independent, and (4) series are both serially and mutually dependent. Similarly as for the longest run-length, only the negative-negative and the negative-positive run-lengths are treated in this paper. Also the bivariate case is reduced to a univariate case by using the transformed variables.

Two series serially and mutually independent. The case of two series being serially and mutually independent can be treated by transforming the original variables to random variables with 0,1 events, which corresponds to $P(X' = 1) = P(X \le C)$ and $P(X' = 0) = P(X > C)$, and similarly for Y_i. Let consider a sequence of a bivariate process, (X_i,Y_i), $i = 1,2,\ldots$, with the two variables having the same distribution and being serially and mutually independent. For levels of truncation C_1 and C_2, four runs are NN, NP, PN, and PP. The joint probabilities are the product of marginal probabilities.

The distribution of negative-negative run-length can be obtained by means of a transformation to a new random variable $Z = X'Y'$, with $Z = 1$ only when $X' = 1$ and $Y' = 1$; otherwise it is zero. The distribution of Z is:

$$f_Z(z) = (p_1p_2)^z\ (1-p_1p_2)^{1-z},\ \text{for}\ z = 0 \text{ or } 1, \quad (2\text{-}115)$$

which is the Bernoulli distribution with $p_1 = F_X(C_1)$ and $p_2 = F_Y(C_2)$. The probability distribution of the run-length is a geometric distribution

$$f_{NN}(K) = (p_1p_2)^{k-1}(1-p_1p_2),\ \text{for}\ k = 1,2,\ldots, \quad (2\text{-}116)$$

with

$$E[K] = \frac{1}{1-p_1p_2}\ ,\ \text{and}\ \ Var[K] = \frac{p_1p_2}{(1-p_1p_2)^2}\ . \quad (2\text{-}117)$$

The negative-positive run-length distribution can be studied similarly. A new random variable $V = X'(1-Y')$ is such that $V = 1$ only when $X' = 1$ and $Y' = 0$; otherwise it is zero. Its distribution is

$$f_V(v) = [p_1(1-p_2)]^v[1-p_1(1-p_2)]^{1-v},\ \text{for}$$
$$v = 0 \text{ or } 1, \quad (2\text{-}118)$$

which is the Bernoulli distribution. The probability distribution of negative-positive run-length is the probability distribution of the run-length of $V = 1$ and is

$$f_{NP}(K) = [p_1(1-p_2)]^{k-1}\ [1-p_1(1-p_2)],\ \text{for}$$
$$k = 1,2,\ldots, \quad (2\text{-}119)$$

with

$$E[K] = \frac{1}{1-p_1(1-p_2)}\ \ \text{and}\ \ Var[K] = \frac{p_1(1-p_2)}{[1-p_1(1-p_2)]^2}. \quad (2\text{-}120)$$

Two series serially independent but mutually dependent. For two series serially independent but mutually dependent, a similar way may be used as for the independent case. Consider a sequence of a bivariate process (X_i,Y_i), $i = 1,2,\ldots$, with two variables of the same distribution but mutually dependent while serially independent. For the truncation levels C_1 and C_2, the four types of runs can be investigated with the joint probabilities of X and Y given by the underlying bivariate distribution, say the bivariate normal. As for the independent case, the probability of the negative-negative run-length is obtained from the variable $Z = X'Y'$ by obtaining the probability distribution of the run-length of $Z = 1$ of the new random variable Z, which distribution is geometric

$$f_{NN}(K) = [F_{X,Y}(C_1,\ C_2)]^{k-1}[1-F_{X,Y}(C_1,C_2)],\ \text{for}$$
$$k = 1,2,\ldots, \quad (2\text{-}121)$$

with $F(X,Y)$ the standard bivariate normal. Similarly, the probability of the negative-positive run-length is obtained from the variable V, for $V = 1$.

Two series serially dependent but mutually independent. For two series serially dependent but mutually independent, the analysis is similar to the case of series serially and mutually independent. This difference is that the joint probabilities must take into account the serial dependence in X_i and Y_i. Probabilities of the negative-negative run-length and negative-positive run-length are also obtained from the variables Z and V, respectively. These solutions are approximate only, since the Markov chain is also an approximation to the first-order linear autoregressive model.

The approximate integration may be used in this latter case by the tetrachoric series expansion, with this approximate solution less accurate than in the above approach. The negative-negative run-length has probabilities

$$P[x_1 \le C_1,\ y_1 \le C_2;\ x_2 \le C_1,\ y_2 \le C_2;\ldots,x_k \le C_1\ y_k \le C_2] = P[x_1 \le C_1,\ x_2 \le C_1,\ldots,x_k \le C_1]\ P[y_1 \le C_2,\ y_2 \le C_2,\ldots,y_k \le C_2]$$

$$= \int_{-\infty}^{C_1}\int_{-\infty}^{C_1}\cdots\int_{-\infty}^{C_1} dF_1 \cdot \int_{-\infty}^{C_2}\int_{-\infty}^{C_2}\int_{-\infty}^{C_2}\cdots\int_{-\infty}^{C_2} dF_2, \tag{2-122}$$

where dF is the multivariate normal integral of Eq. 2-61. Using the univariate dependent case given by Saldarriaga and Yevjevich (1970), probabilities of negative-negative run-length are obtained by multiplying the marginal probabilities obtained and given as tables for a given parameter of dependence. Probabilities of the negative-positive run-length are obtained by the same procedure, because

$$P[x_1 \le C_1,\ y_1 > C_2;\ x_2 \le C_1,\ y_2 > C_2;\ldots;\ x_k \le C_1,\ y_k > C_2] = \int_{-\infty}^{C_1}\int_{-\infty}^{C_1}\cdots\int_{-\infty}^{C_1} dF_1 \int_{C_2}^{\infty}\int_{C_2}^{\infty}\cdots\int_{C_2}^{\infty} dF_2 . \tag{2-123}$$

Two series serially and mutually dependent. Analytical treatment is much more complex for two series mutually and serially dependent. Approximate analytical solutions are presented in this paper. The degree of approximation can be determined by using the experimental method.

Four different approximations are given: (1) by considering the Markov chain lumpability, (2) by using the Markov chain approximations for the two processes, then determine the Markov chains for the transformed variables; (3) by considering a four-state Markov chain, and (4) by approximate integrations using the tetrachoric series expansion.

For the Markov chain lumpability approach, consider a sequence of a bivariate process (X_i,Y_i), $i = 1,2,\ldots$, with two series of the same normal distribution, mutually and serially dependent. By considering the four-state Markov chain of Table 2-1, first the lumpability for this chain is investigated both for the marginal distributions of X and Y, as well as for the transformed random variables Z and V. If found lumpable into a two-state Markov chain, it becomes feasible to find probability distribution functions of negative-negative and negative-positive run-lengths by using the variables Z and V, respectively.

For the Markov chain approximations approach, the marginal distributions of X and Y are only considered, with their series dependence approximated by Markov chains, and transition probabilities as schematically represented in Tables 2-2 and 2-3, and obtained for the Z variable by Eqs. 2-57 and 2-58, which require the solutions of a quadrivariate normal distribution. By computing the transition probability matrix of Z, probabilities of negative-negative run-length are obtained by using the equations developed for the univariate dependent case, Eqs. 2-109 and 2-112. Similarly, probabilities of negative-positive run-length are obtained by using V, and Eqs. 2-109 and 2-112, with transition probabilities of Table 2-5 computed by Eqs. 2-59 and 2-60.

For the four-state Markov chain approach, let consider the four-state chain as represented by Table 2-1, obtained as approximations using the tetrachoric series expansion.

The matrix of joint probabilities U is obtained either by integration or from tables of a bivariate normal distribution, or from the four-state Markov chain Q, with

$$\underline{U}_{i+1} = \underline{Q}^T\ \underline{U}_i . \tag{2-124}$$

The vector U of joint probabilities gives

$$\begin{aligned} U_i(1,1) &= P[X_i \le C_1,\ Y_i \le C_2] , \\ U_i(0,1) &= P[X_i > C_1\ ,\ Y_i \le C_2] , \\ U_i(1,0) &= P[X_i \le C_1,\ Y_i > C_2], \\ U_i(0,0) &= P[X_i > C_1,\ Y_i > C_2]. \end{aligned} \tag{2-125}$$

Probabilities of negative-negative run-length are schematically represented as

$$P\begin{bmatrix} \cdot & 1 & 1 & \cdots & 1 & \cdot \\ \cdot & 1 & 1 & \cdots & 1 & \cdot \end{bmatrix}$$

(k)

at least one is zero — at least one is zero

They are obtained by considering all possible events by

$$P[NN] = a_1^{k-1}\ [U(0,0)d_1(a_2+a_3+a_4) + U(0,1)\ c_1(a_2+a_3+a_4) + U(1,0)\ b_1(a_2+a_3+a_4)]$$

$$= a_1^{k-1}\ (1-a_1)\ [U(0,0)\ d_1+U(0,1)\ c_1 + U(1,0)b_1]. \tag{2-126}$$

Similarly probabilities of negative-positive run-length are represented by

$$P\begin{bmatrix} \cdot & 0 & 0 & \cdots & 0 & \cdot \\ \cdot & \underbrace{1 \quad 1 \quad \cdots \quad 1}_{k} & & & & \cdot \end{bmatrix}$$

where the pairs shown with arrows can be (0,0), (1,1), or (1,0). Probabilities are obtained by considering all possible events, as

$$P[NP] = c_3^{k-1}(1-c_3)\ [U(0,0)\ d_3 + U(1,0)b_3 + U(1,1)\ a_3]. \qquad (2\text{-}127)$$

For the tetrachoric series expansion approach, based on the fact that no explicit expression exists for the general solution of the multivariate normal integral, the approximated solutions are obtained by using the tetrachoric series expansion. This case is similar to the solution given for the univariate case by Saldarriaga and Yevjevich (1970). Probabilities of negative-negative run-length are obtained as follows. The negative-negative run-length of $k = 1$ is equal to $P(1^-,3^-)$, in the nomenclature used for the quadrivariate normal integration. The negative-negative run-length of $k = 2$ is $P(1^-,2^-,3^-,4^-)$. The negative run-length k is obtained by generalizing Eqs. 2-75 and 2-76 for the multivariate normal case.

Probabilities of negative-positive run-length are obtained in a similar way. For the negative-positive run-length for $k = 1$ it is $P(1^-,3^+)$, for the negative-positive run-length of $k = 2$ it is $P(1^-,2^-,3^+, 4^+)$, as given by Eq. 2-85; for any negative-positive run-length they can be obtained by generalizing Eq. 2-84.

It should be stressed that in this approximation if a truncation in the tetrachoric series expansion after $i = 2$ is made this implies that the terms containing ρ^3 or higher powers of ρ can be neglected. The error introduced, however, is small and can be assessed by using the experimental statistical (Monte Carlo) approach.

2-10 Probability Distributions of Run-Sums of Infinite Series

Univariate case. It was shown in Section 2-7 that finding the distributions of largest run-sums in a given sample is complex even for the simple case of univariate independent normal process. For run-sums of infinite series the same difficulties are encountered as for the largest run-sum of a sample. For the univariate independent normal process, Downer et al. (1967) give the exact properties of run-sums using the cumulants. Few first moments of the distribution of run-sums can also be obtained from the crossing theory. Llamas and Siddiqqi (1969) summarize the essentials of the above paper, some results reported in a strengthened form, and some new results included. A truncated distribution was used for the negative run-sum, namely

$$F_1(x) = \frac{F(C) - F(C-x)}{F(C)}, \quad \text{if } C \le 0$$

and $$F_1(x) = 0, \quad \text{if } C > 0 \qquad (2\text{-}128)$$

with

$$F_1(x) = P[C - x_i \le x | x_i \le C] .$$

The probability density function of run-sum for a more general case is given by Heiny (1968) in an approximate way using the two-parameter gamma probability density function and the associated Laguerre polynomials. The two parameters of this gamma function are estimated by equating the first two moments of run-sum with the first two moments of the gamma function, with

$$Q(z;g;h) = \frac{e^{-z/2g}\ z^{h/2-1}}{(2g)^{h/2}\ \Gamma(\frac{h}{2})}, \quad \text{for } z > 0 ,$$

$$E[Z] = hg , \text{ and } Var[Z] = 2g^2 h , \qquad (2\text{-}129)$$

where

$$g = \frac{p\kappa_1 + q\kappa_2}{2q\kappa_1} \quad \text{and} \quad h = \frac{2\kappa_1}{p\kappa_1 + q\kappa_2} \qquad (2\text{-}130)$$

with $p = F(C)$, $q = 1-F(C)$ and κ_1 and κ_2 the first two cumulates. If a greater accuracy is required the approximation can be improved, Siddiqqi (1960), by using the associated Laguerre polynomials as shown by Heiny; however, no explicit expression is obtained.

For a univariate dependent process, it is more complex to obtain the exact distribution and parameters of the run-sum. Approximate expressions for parameters are obtained by Heiny (1968) as

$$E[S_n] = (m_1 + \frac{r^2}{s} m_2 + r m_3)[1 + O(\rho^2)] , \qquad (2\text{-}131)$$

with r and s the same as in Eq. 2-112, and m_1, m_2, and m_3 the moments of the random variable Y_i whose density function is given by

$$G_{Y_i}(y_i) = P[X_i - C \le y_i | X_1 > C, X_2 > C, \ldots, X_n > C, X_{n+1} \le C], \text{ if}$$

$$y_i > 0 \qquad (2\text{-}132)$$

$$= 0 \text{ elsewhere,}$$

and

$$S_n = \sum_{i=1}^{n} Y_i .$$

The variance of the run-sum is given in approximate form also by Heiny as

$$Var[S_n] = \{\sigma_1^2 + \frac{r^2}{s}\sigma_2^2 + r\sigma_3^2 + 2r r_1 + 2r r_3 + \frac{2r(1-2s)}{s} r_2 + r s m_3^2 + \frac{1-s-2s^2+3s^3-s^4}{s^2} m_2^2 + 2r^2 m_2 m_3\} \{1+O(\rho^2)\} . \qquad (2\text{-}133)$$

Bivariate case. As shown in the preceding text, distributions of the run-sum are not simple to obtain as in the case of run-length, even for simple processes. The bivariate case is expected to be even more complex than the univariate case. Similarly as for the run-length, four cases, NN, NP, PN, and PP, for each of the four bivariate cases should be investigated. Approximate expressions have been found for parameters of run-sum distributions for the serially

and mutually independent, serially dependent but mutually independent, and serially independent but mutually dependent, but have not yet been investigated for mutually and serially dependent processes. Most of the approximations were developed by Heiny (1968) for the univariate case. However, the degrees of approximation are not shown since the experimental method was not used.

The negative-negative run-sums are composed of the negative run-sum of each sequence over the common negative-negative run-length. The run-sums are not for the complete univariate runs.

The case of mutually and serially independent components of bivariates, as analyzed by Llamas (1968) and Llamas and Siddiqqi (1969), has

$$E[S_{11}] = \frac{E[X_1^*]}{1-p_1p_2}, \quad E[S_{21}] = \frac{E[Y_1^*]}{1-p_1p_2}, \tag{2-134}$$

$$Var[S_{11}] = \frac{(1-p_1p_2)Var(X_1^*) + p_1p_2[E(X_1^*)]^2}{(1-p_1p_2)^2}, \tag{2-135}$$

$$Var[S_{21}] = \frac{(1-p_1p_2)\ Var[Y_1^*] + p_1p_2[E(Y_1^*)]^2}{(1-p_1p_2)^2},$$

and

$$Cov[S_{11},S_{21}] = \frac{p_1p_2E[X_1^*]\ E[Y_1^*]}{(1-p_1p_2)^2} \tag{2-136}$$

Llamas (1968) gives parameters of the standard normal variable with the truncation level of the population mean.

The serially independent but mutually dependent bivariate case, studied by Llamas (1968) and Heiny (1968), according to Heiny has the following parameters of the distribution of positive run-sum

$$E[S_L] = \frac{1}{q}\kappa_1, \quad E[T_L] = \frac{1}{q}\lambda_1, \tag{2-137}$$

$$Var[S_L] = \frac{p}{q^2}\kappa_1^2 + \frac{1}{q}\kappa_2 \text{ , and } Var[T_L] = \frac{p}{q^2}\lambda_1^2 + \frac{1}{q}\lambda_2 \tag{2-138}$$

which κ_1 and κ_2, and λ_1 and λ_2, the first and second cumulants of Y_{11} and Y_{12}, respectively, $p = P(X_{11} > C_1, X_{21} > C_2)$, and (Y_{11}, Y_{12}) a sequence of bivariate series with common density function given by

$$G(x,y) = 0, \text{ if } x \le 0 \text{ and/or } y \le 0$$

and (2-139)

$$G(x,y) = \frac{F(x,y)-q}{p}, \text{ if } x > 0,\ y > 0.$$

Llamas (1968) gives an approximation for the case where the truncation level in both components is the median and the underlying distribution is standard normal. The degree of approximation was not determined.

The case of serially independent but mutually dependent components, as discussed by Heiny (1969), is similar to the univariate case with the only modification for the run-sum to take into account the different run-sums of components.

All above studies, however, consider only the cases of both runs being either negative or positive, with no attempt to study the positive-negative or negative-positive runs. The complexity in analytical developments were likely the major reason for the lack of studies in literature related to the serially and mutually dependent bivariate case of run-sums. Therefore, they will be investigated by the experimental method, similarly as it was done for the largest run-sum in a sample of the given size.

Chapter III
EXPERIMENTAL APPROACH FOR STUDYING DROUGHT CHARACTERISTICS OF STATIONARY STOCHASTIC PROCESSES

The data generation or experimental Monte Carlo method derives, in an approximate way, the drought frequencies as the estimates of drought probabilities of large return periods by generating a given number of samples of data of given sizes.

The analytical method derives the probability of any drought parameter by generalizing the properties of the available time series. When the mathematics involved become very complex, the analytical method may help in setting up the data generation approach and in the interpretation of its results. The data generation method requires univariate, bivariate, or multivariate generations of samples, in the latter two cases also for the case of mutually and serially dependent components. The dependence used here is in the form of the first-order linear autoregressive model for all dependent components of the bivariate or multivariate case, with the serial correlation coefficients differing from one component to another.

3-1 A Multivariate Generation Model

Hydrologic variables, such as streamflow at different stations in a region, are both spatially and serially correlated since they are affected by similar climatic and hydrologic factors. The drought in a region depends highly on the level of water demand besides depending on the available water.

Demand levels are not necessarily the same throughout a region. Furthermore, since historical records are short and consequently less reliable it is necessary to study droughts of long return periods on simulated records at each station by preserving both the time structure and the interstation correlation of historical series. This requires the use of multivariate data generation approach.

The parameters that are unbiased and have the lowest sampling variation are ones to be best preserved in the data generation method. Saldarriaga and Yevjevich (1970) show that the run-length properties of stationary processes are independent of their means and the standard deviations while being dependent on the probability q of the truncation level, the series dependence structure and the skewness of distribution. On the contrary, the run-sum properties depend on all above properties and in particular they are directly proportional to standard deviation of the process. Once the run-sum of the standardized variable is known, the run-sum for any other σ is obtained by multiplying the run-sum of the standardized variable by this σ. As a consequence, the generation of long samples of two series will be made each with the mean of zero, standard deviation of one, two truncation levels q_1 and q_2, given sample size n, and the first-order autoregressive time dependence models for their serial correlation coefficients $\rho_1(\varepsilon_x)$ and $\rho_1(\varepsilon_y)$ and their lag-zero cross correlation coefficient $\rho(\varepsilon_x,\varepsilon_y)$. The generated samples are then used for the analysis of probabilities of runs covering the cases most likely to occur in practice. In the bivariate case considered, the two streamflow station series, generally cross correlated, are used.

Multivariate time series analysis has been studied for some time, Quenouille (1957). However, its use in hydrology for the purpose of generating new series may have been initiated by Fiering (1963), who treated both the bivariate and the multivariate model. The bivariate model was

$$\frac{y_{i+1}-\bar{y}}{s_y} = \rho_{x,y}\,\frac{x_{i+1}-\bar{x}}{s_x} + (1-\rho_{x,y}^2)^{1/2}\,u_i \quad , \qquad (3\text{-}1)$$

with $\bar{x}$ and $\bar{y}$ the means, s_x and s_y the standard deviations, $\rho_{x,y}$ the lag-zero cross correlation coefficient, and

$$u_i = \left(\frac{y_i-\bar{y}}{s_y}\right)\pi + v_i\,(1-\pi)^{1/2} \quad , \qquad (3\text{-}2)$$

in which π is the cross correlation coefficient between u_i and y_i, expressed in function of the first serial correlation coefficient $\rho_1(x)$ of X and the first serial correlation coefficient $\rho_1(y)$ of Y, with π given by

$$\pi = \frac{\rho_1(y) - \rho_1(x)\,\rho_{x,y}^2}{\sqrt{1-\rho_{x,y}^2}} \qquad (3\text{-}3)$$

with v_i a random normal deviate with zero mean and unit variance. The means, variances, the respective serial correlation coefficients, and the lag-zero cross correlation coefficient between the two variables are preserved approximately through this model.

Matalas (1967) gives a lag-one multivariate Markov model which preserves the means, variances, the respective first serial correlation coefficients and the lag-zero cross correlation coefficients, and if desired, the lag-one cross correlation coefficients. The presented model is based on a multivariate weakly stationary generating process, defined by

$$\underline{X}_{i+1} = \underline{A}\,\underline{X}_i + \underline{B}\,\underline{\xi}_{i+1} \qquad (3\text{-}4)$$

with $\underline{X}_{i+1}$, $\underline{X}_i$ and $\underline{\xi}_{i+1}$ being (m x 1) matrices, the independent random components $\underline{\xi}$ mutually independent and independent of components $\underline{X}_i$ and $\underline{A}$ and $\underline{B}$ the (m x m) matrices whose elements are defined in such a way as to preserve the desired statistics. In this case $\underline{A}$ is a diagonal matrix. Young and Pisano (1967) in a comment on paper by Matalas, and later on in a more detailed presentation (Young and Pisano, 1968), give an alternative method of solving the $\underline{B}$ matrix by means of an orthogonalization or a recursive scheme, making it a simpler solution. Pegram and James (1973) present an extension of this model to the multi-lag case, specifically to the lag-two case, in order to preserve the means, the variances, the respective lag-one and lag-two serial correlation coefficients. The Young and Pisano model preserves only the first two moments. It means that residuals should be normally distributed, or transformed to become normally distributed (McGinnis and Sammons, 1970). If it is not feasible to use such a transformation for any reason, a model is required for preserving the

third-order moments, such as the one given by Moreau (1970), which preserves the skewness coefficients.

Let consider an ensemble of the trend-free streamflow samples from a region as the $X_{i,j}$ series, with i the station number ($i = 1,2,\ldots,m$) and j the time sequence ($j = 1,2,\ldots,n$). The streamflow time series can be considered as composed of a deterministic component (periodicity in parameters) and a dependent stochastic component, or

$$X_{i,j} = D_1 + D_2 \, \varepsilon_{i,j} \, . \tag{3-5}$$

Periodicities in the mean and standard deviation are removed by

$$\varepsilon_{i,j} = \frac{X_{i,j} - D_1}{D_2} \, . \tag{3-6}$$

The linear models for the stationary time series are studied by using correlograms or spectra. For monthly runoff, with the periodicity in parameters removed, it has been found that a first- or a second-order linear autoregressive model often fit well the time series dependence of the stochastic component (Roesner and Yevjevich, 1967). For annual time series used in this study, the first-order linear autoregressive model was often used (Yevjevich, 1964) as a good approximation to time dependence. Therefore, the first-order model as a first, basic approximation is exclusively used in this study. To simplify the analysis, stochastic components are standardized. The $\varepsilon_{i,j}$ variable is considered normally distributed with the mean zero and variance unity. In case $\varepsilon_{i,j}$ is not normally distributed, transformations such as logarithmic, square root, cubic root, or others are made to approach a normal distribution as closely as feasible. The multivariate case of the model is then

$$\underline{\varepsilon}_{j+1} = \underline{A} \, \underline{\varepsilon}_j + \underline{B} \, \underline{\xi}_{j+1} \, , \tag{3-7}$$

with j the time, $\underline{A}$ and $\underline{B}$ the ($m \times m$) diagonal matrices, $\underline{\xi}_j$ an ($m \times 1$) matrix of independent components following the standard normal distribution, and

$$E(\underline{\varepsilon}) = 0; \quad E(\underline{\xi}) = 0; \quad Var(\underline{\varepsilon}) = 1; \quad Var\,(\underline{\xi}) = 1. \tag{3-8}$$

Calling M_o the lag-zero covariance matrix of $\underline{\varepsilon}_j$, then

$$E(\underline{\varepsilon}_j \, \underline{\varepsilon}_j^T) = M_o \, , \tag{3-9}$$

and M_1 the lag-one covariance matrix, or

$$E(\underline{\varepsilon}_{j+1} \, \underline{\varepsilon}_j^T) = M_1 \, . \tag{3-10}$$

Taking the expectation of Eq. 3-7, the check is made whether the means are preserved. Multiplying the same equation by $\underline{\varepsilon}_j^T$ and taking the expected values then

$$E(\underline{\varepsilon}_{j+1} \, \underline{\varepsilon}_j^T) = \underline{A} \;\; E(\underline{\varepsilon}_j \, \underline{\varepsilon}_j^T) + \underline{B} \, E(\underline{\xi}_j \, \underline{\xi}_j^T) \, . \tag{3-11}$$

By replacing

$$M_1 = \underline{A} \, M_o \, , \text{ or } \; \underline{A} = M_1 \, M_o^{-1} \, , \tag{3-12}$$

and multiplying Eq. 3-7 by $\underline{\varepsilon}_{j+1}^T$, replacing $\underline{\varepsilon}_{j+1}^T$ of the right hand side by its value of Eq. 3-7, and finally taking the expected value, then

$$M_o = \underline{A} \, M_o \, \underline{A}^T + \underline{B} \, (\underline{\xi}_{j+1} \, \underline{\xi}_{j+1}^T) \, \underline{B}^T \, . \tag{3-13}$$

If $\underline{\xi}$ are considered to be mutually independent components as well as serially uncorrelated, then $E(\underline{\xi}_{j+1} \, \underline{\xi}_{j+1}^T) = I$, the identity matrix, which means that the $m \times m$ matrix has each diagonal element equal to unity and all off-diagonal elements equal to zero, so that

$$\underline{B} \, \underline{B}^T = M_o - M_1 \, M_o^{-1} \, M_1^T \, . \tag{3-14}$$

Equations 3-12 and 3-14 define the coefficients of matrices $\underline{A}$ and $\underline{B}$. Equation 3-14 is straightforward to solve. For the bivariate case used in this study with components X and Y,

$$M_1 = \begin{bmatrix} \rho_1(\varepsilon_x) & \rho_{+1}(\varepsilon_x,\varepsilon_y) \\ \rho_{-1}(\varepsilon_x,\varepsilon_y) & \rho_1(\varepsilon_y) \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix} \begin{bmatrix} 1 & \rho(\varepsilon_x,\varepsilon_y) \\ \rho(\varepsilon_x,\varepsilon_y) & 1 \end{bmatrix} , \tag{3-15}$$

so that

$$\rho_1(\varepsilon_x) = a_{11}, \qquad \rho_{-1}(\varepsilon_x,\varepsilon_y) = \rho_1(\varepsilon_x)\rho(\varepsilon_x,\varepsilon_y), \tag{3-16}$$
$$\rho_1(\varepsilon_y) = a_{22}, \text{ and } \; \rho_{+1}(\varepsilon_x,\varepsilon_y) = \rho_1(\varepsilon_y) \, \rho(\varepsilon_x,\varepsilon_y),$$

in which $\rho_1(\varepsilon_x)$ and $\rho_1(\varepsilon_y)$ are the first serial correlation coefficients of the ε_x and ε_y series, and $\rho(\varepsilon_x,\varepsilon_y)$ and $\rho_{+1}(\varepsilon_x,\varepsilon_y)$ are the lag-zero and lag-one cross correlation coefficients between the ε_x and ε_y series, respectively.

The term $\underline{B} \, \underline{\xi}_{j+1}$ consists of independent stochastic components of the model, which are independent of ε_x and ε_y but are mutually dependent. Replacing $\underline{B} \, \underline{\xi}_{j+1}$ by $\underline{\nu}_{j+1}$, with $\underline{\nu}_{j+1}$ a ($m \times 1$) matrix, it becomes an independent stochastic component. Multiplying $\underline{\nu}_{j+1}$ by $\underline{\nu}_{j+1}^T$ and taking the expectation, the covariance matrix C of the stochastic serially independent component is obtained. Since this is a symmetric matrix, one solution is a lower triangular matrix, so that the solution for $\underline{B}$ can be obtained either by orthogonalization or recursive scheme technique, or by principal components technique. Then

$$E(\underline{\nu}_{j+1} \, \underline{\nu}_{j+1}^T) = \underline{B} \, E(\underline{\xi}_{j+1} \, \underline{\xi}_{j+1}^T) \, \underline{B}^T \, , \tag{3-17}$$

with $E(\underline{\xi}_{j+1} \, \underline{\xi}_{j+1}^T) = I$. Since $\underline{\xi}$ are mutually independent components,

$$E(\underline{\nu}_{j+1} \, \underline{\nu}_{j+1}^T) = \underline{B} \, \underline{B}^T \, . \tag{3-18}$$

For the bivariate case, replacing $\underline{B}$ and $\underline{B}^T$ by its matrices, Eq. 3-18 gives

$$\begin{bmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{21} \\ 0 & b_{22} \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{Var}(\nu_{1,j+1}) & \mathrm{Cov}(\nu_{1,j+1}, \nu_{2,j+1}) \\ \mathrm{Cov}(\nu_{1,j+1}, \nu_{2,j+1}) & \mathrm{Var}(\nu_{2,j+1}) \end{bmatrix}, \qquad (3\text{-}19)$$

which when solved, becomes

$$B = \begin{bmatrix} S_{\nu_1} & 0 \\ \rho(0)s_{\nu_2} & S_{\nu_2}\sqrt{1-\rho^2(0)} \end{bmatrix} . \qquad (3\text{-}20)$$

To obtain the relation between the cross correlation coefficient $\rho(\varepsilon_x,\varepsilon_y)$ the stochastic dependent components and the cross correlation coefficients $\rho(0)$, between the stochastic independent components, the corresponding values are replaced in Eq. 3-14, or

$$\begin{bmatrix} 1 - \rho_1^2(\varepsilon_x) & \rho(\varepsilon_x,\varepsilon_y)[1 - \rho_1(\varepsilon_x)\rho_1(\varepsilon_y)] \\ \rho(\varepsilon_x,\varepsilon_y)[1-\rho_1(\varepsilon_x)\rho_1(\varepsilon_y)] & 1 - \rho_1^2(\varepsilon_y) \end{bmatrix}$$

$$= \begin{bmatrix} S_{\nu_1}^2 & \rho(0)S_{\nu_1}S_{\nu_2} \\ \rho(0)S_{\nu_1}S_{\nu_2} & S_{\nu_2}^2 \end{bmatrix} , \qquad (3\text{-}21)$$

with

$$S_{\nu_1} = \sqrt{1 - \rho_1^2(\varepsilon_x)} \quad , \quad S_{\nu_2} = \sqrt{1 - \rho_1^2(\varepsilon_y)} \quad ,$$

$$\rho(0) = \frac{\rho(\varepsilon_x,\varepsilon_y)[1 - \rho_1(\varepsilon_x)\rho_1(\varepsilon_y)]}{\sqrt{1 - \rho_1^2(\varepsilon_x)}\sqrt{1 - \rho_1^2(\varepsilon_y)}} \quad , \qquad (3\text{-}22)$$

$$\rho(\varepsilon_x,\varepsilon_y) = \frac{\rho(0)\sqrt{1 - \rho_1^2(\varepsilon_x)}\sqrt{1 - \rho_1^2(\varepsilon_y)}}{1 - \rho_1(\varepsilon_x)\rho_1(\varepsilon_y)} \quad , \qquad (3\text{-}23)$$

and

$$B = \begin{bmatrix} \sqrt{1 - \rho_1^2(\varepsilon_x)} & 0 \\ \rho(0)\sqrt{1 - \rho_1^2(\varepsilon_x)} & \sqrt{1 - \rho_1^2(\varepsilon_x)}\sqrt{1 - \rho_1^2(\varepsilon_y)} \end{bmatrix} . \qquad (3\text{-}24)$$

The advantage of making use of correlation between the serially independent stochastic components is the statistical inference about the correlation coefficient.

The parameters are now $\rho(0)$, $\rho_j(\varepsilon_x)$, and $\rho_j(\varepsilon_y)$, with the model preserving the means, the variances, the lag-zero cross correlation, and the respective serial correlation coefficients of normal variables. The lag-one cross correlation coefficient if insignificant, as commonly found, need not be preserved.

Another advantage of the model presented is that in cases of no serial dependence, say for annual precipitation series, $\rho_j(\varepsilon_x)$ and $\rho_j(\varepsilon_y)$ become zeros, with the model reduced to a simple form.

3-2 Investigated Drought Characteristics

The drought properties investigated in the case of stationary time series are of two kinds, runs statistics related to a given sample size, and runs of infinite series. The statistics of interest for runs of samples of a given size in this study are the longest run-length and the largest run-sum. The distribution of these random variables vary with the truncation levels C_1 and C_2, the sample size, and the parameters of the underlying process as described in Section 3-1, of which $\rho_1(\varepsilon_x)$, $\rho_1(\varepsilon_y)$, and $\rho(0)$ are the most significant.

Runs of interest for infinite series are the run-length, the run-sum and the run-intensity. These random variables vary with the truncation levels C_1 and C_2, and the parameters of the underlying process, particularly $\rho_1(\varepsilon_x)$, $\rho_1(\varepsilon_y)$ and $\rho(0)$.

For the bivariate case, the run-sum is defined as the sum of the partial run-sums, whether positive or negative. For the negative-negative run-sum it is

$$S_{nn} = \sum_{i=1}^{k} (C_1 - X_i) + \sum_{i=1}^{k} (C_2 - Y_i) . \qquad (3\text{-}25)$$

Similar definitions hold for the negative-positive (S_{np}), positive-negative (S_{pn}) and positive-positive (S_{pp}) run-sums.

For infinite series, a joint distribution of run-length and run-sum may be obtained, and from it the properties of the run-sum may be derived. Changing parameters to consider are five: C_1, C_2, $\rho_1(\varepsilon_x)$, $\rho_1(\varepsilon_y)$, and $\rho(0)$, for standardized normal variables. Several combinations are selected for the use in generation method. For their selected numbers m_1 through m_5, respectively, the total number of cases to be investigated is $m_1m_2m_3m_4m_5$, which will be considered in computing the total number of samples to be generated.

The truncation level of each series can be better expressed in the form of quantiles q, with $q = P(X \leq C)$. Three values are selected for each of the two series in the bivariate case: $q_1 = 0.50, 0.35, 0.20,$ and $q_2 = 0.50, 0.35, 0.20$, respectively. The selected values of serial correlation coefficients are: $\rho_1(\varepsilon_x) = 0.0, 0.2, 0.4,$ and $\rho_1(\varepsilon_y) = 0.0, 0.2, 0.4$. The lag-zero cross correlation coefficients between the serially independent stochastic components are selected as: $\rho(0) = 0.3, 0.5, 0.7$. The total number of combinations for all three correlation coefficients is fifteen with twelve combinations resulting from $\rho(o) = 0.3, 0.5, 0.7$; $\rho_1(\varepsilon_x) = 0.2, 0.4,$ and $\rho_1(\varepsilon_y) = 0.2, 0.4,$ plus three combinations resulting from $\rho(0) = 0.3, 0.5, 0.7$; $\rho_1(\varepsilon_x) = 0$ and $\rho_1(\varepsilon_y) = 0$. The sample sizes selected are $n = 25, 50,$ and 200. The value of 200 was chosen in order to consider an extreme of large historical samples presently available.

No consideration was given to eventual varying the skewness coefficient of the ξ_i components. This would increase the total number of cases and samples. The study by Millan and Yevjevich (1970) showed that the distribution of the longest run-length was only slightly affected by the skewness coefficient of univariate asymmetrical dependent stochastic components, while the distribution of the largest run-sum is much more affected by the skewness coefficient. In total, 135 combinations of five parameters are selected for the study by the experimental (data generation) method, for deriving the distributions of runs for an infinite series, and 405 combinations for the distributions of runs for the selected sample sizes.

The selection of the number of samples to be generated was studied by Millan and Yevjevich (1970) for the longest run-length, considering the distribution of the sample mean run, m_r, which was said to be asymptotically normal based on the central limit theorem. In this study, the number N of samples of a given size n, to be generated in such a way that the probability is at least 0.95 for the estimate m_r to be within the tolerance limits $\mu_r \pm 0.2\sigma_r$, is computed to be 200. A total size of generated numbers is then $Nn = 40,000$. For the case of runs of a given sample size n the number of bivariate samples is then $m = 40,000/n$, or

n	25	50	200
N	1600	800	200

Once Nn random numbers are generated for $n = 200$, all numbers are used for the smaller values of n in order to allow an increase in the accuracy of estimating distributions of runs.

3-3 Algorithms Used for Computing Relative Frequency Distributions of Runs

The procedure followed in the experimental method is divided in three parts: (1) generation of bivariate samples; (2) determination of frequency distributions of selected runs for the bivariate case and infinite series; and (3) determination of frequency distributions of selected runs for the bivariate case and given sample sizes. The distribution of runs of both kinds are obtained for all the combinations of selected parameters: C_1, C_2, $\rho_1(\varepsilon_x)$, $\rho_1(\varepsilon_y)$, $\rho(0)$, and n.

For generating the bivariate samples the model presented in Section 3-1 is used, represented by Eq. 3-7, expressed as

$$\varepsilon_{1,j+1} = \rho_1(\varepsilon_x)\varepsilon_{1,j} + \sqrt{1 - \rho_1^2(\varepsilon_x)}\ \xi_{1,j+1}$$

$$\varepsilon_{2,j+1} = \rho_1(\varepsilon_y)\varepsilon_{2,j} + \rho(0)\sqrt{1 - \rho_1^2(\varepsilon_y)}\ \xi_{1,j+1} + \sqrt{1 - \rho_1^2(\varepsilon_x)}\sqrt{1 - \rho_1^2(\varepsilon_y)}\ \xi_{2,j+1}\ , \quad (3\text{-}26)$$

with $\rho_1(\varepsilon_x)$, $\rho_1(\varepsilon_y)$, and $\rho(0)$ as defined earlier, and $\xi_{1,j+1}$ and $\xi_{2,j+1}$ the two independent series of random numbers. The series of standard normal random numbers are generated directly, namely by transforming the uniform numbers into the normal random numbers given in Box and Muller (1958), whose equations for a pair of standard normal random numbers ξ_1 and ξ_2 are

$$\xi_1 = (-2 \ln \lambda_1)^{\frac{1}{2}} \cos 2\pi\lambda_2\ ,$$

and (3-27)

$$\xi_2 = (-2 \ln \lambda_1)^{\frac{1}{2}} \sin 2\pi\lambda_2\ ,$$

where λ_1 and λ_2 are two consecutive independent random numbers which are uniformly distributed in the interval (0,1).

The 80,000 random numbers required were generated by means of Eq. 3-27 and used in Eq. 3-26 for each of the 15 combinations of $\rho_1(\varepsilon_x)$, $\rho_1(\varepsilon_y)$, and $\rho(0)$. They were stored on magnetic tape. A tape of 250,000 standard normal random numbers was used for the purpose. To obtain the distribution for selected runs of infinite series, a computer flow chart was prepared, Fig. 3-1. To obtain the distribution for selected runs for the given sample sizes, a computer flow charge was also prepared, Fig. 3-2.

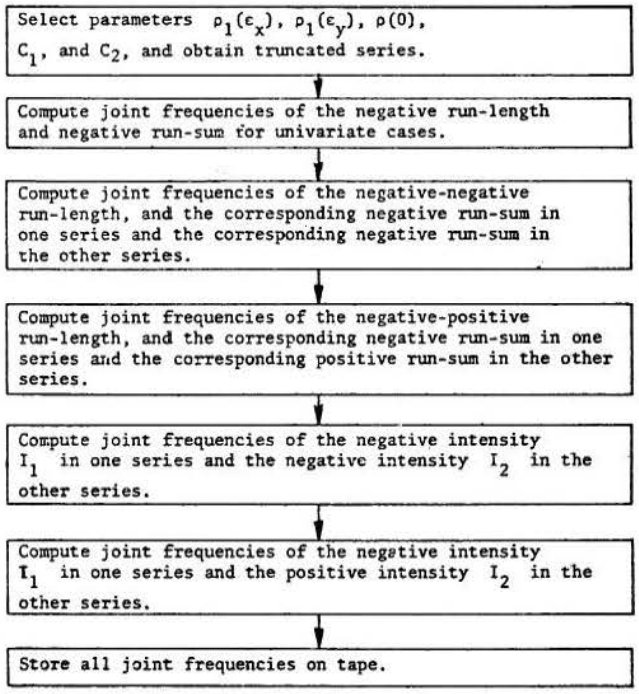

Fig. 3-1 Flow Chart of the Algorithm for the Analysis of Runs of Infinite Populations.

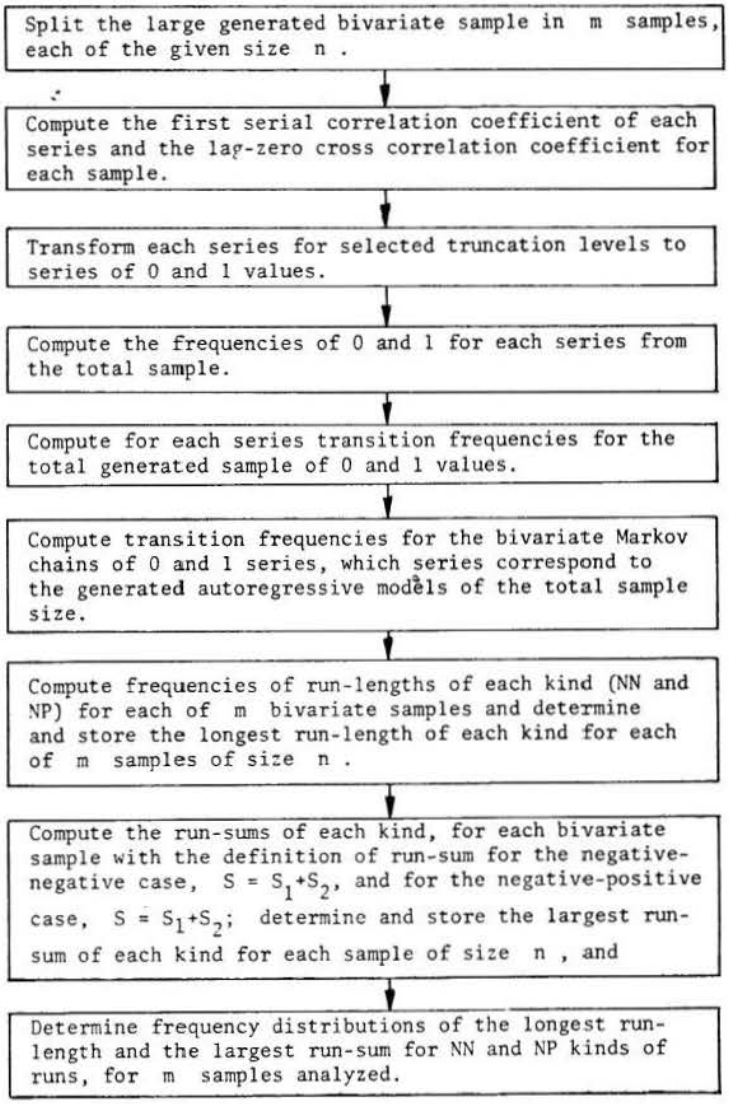

Fig. 3-2 Flow Chart of the Algorithm for the Analysis of Runs for Given Sample Sizes.

Chapter IV
ANALYSIS OF RESULTS OBTAINED BY THE EXPERIMENTAL METHOD

Frequency distributions of various runs, for both infinite series and samples of given sizes, obtained by the experimental method in generating a very large sample for drought variables of given characteristics, are fitted by the selected probability distribution functions.

The results can be presented in two ways: (a) as graphs and/or tables of frequencies distributions, and (b) as estimated parameters of fitted probability distribution functions. The latter approach has the advantage of condensing the information, because two to four parameters are sufficient to define the probability distribution functions. Furthermore, the estimated parameters of probability distributions of runs can be expressed in terms of parameters of underlying time series and their truncation levels. Because of these two particular advantages, the second approach is used only.

Since distributions of run-lengths are discrete, discrete probability functions are fitted to frequency distributions of run-lengths, while continuous probability distribution functions are fitted to frequency distributions of run-sums.

4-1 Fitting Discrete Probability Distribution Functions to Frequency Distributions of Run-Lengths

Ord (1972) and Johnson and Kotz (1969a) present a detailed analysis of discrete distributions, with systems of discrete distributions defined by difference equations. This is analogous to the Pearson system of continuous distribution functions, defined by differential equations. The discrete system is based on the fact that for the hypergeometric distribution the ratio of the probability functions $(P_{j+1}-P_j)/(P_{j+1}+P_j)$ is of the form: linear function of j divided by quadratic function of j, Ord (1967). The difference equation is

$$\Delta P_{r-1} = \frac{(a-r)P_{r-1}}{b_o+b_1 r+b_2 r(r-1)}, \qquad (4\text{-}1)$$

and the criterion is defined by

$$\kappa = \frac{(b_1-b_2-1)^2}{4b_2(b_o+a)}. \qquad (4\text{-}2)$$

The alternative form of Eq. 4-1 is

$$\Delta P_{r-1} = \frac{(a-r)P_r}{(a+b_o)+(b_1-1)r+b_2 r(r-1)}, \qquad (4\text{-}3)$$

with values of parameters expressed in terms of the first four moments, as given in Table 4-1, with

$$D_P = 2(5\beta_2-6\beta_1-9),\quad D_G = 4\mu\mu_3+2\mu_2(\mu+\mu^2-3\mu_2),$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3},\ \text{and}\ \beta_2 = \frac{\mu_4}{\mu_2^2}.$$

For distributions with lower boundary at the origin, the index of dispersion I is defined by Ord as $I = \mu_2/\mu$; another similar index is $S = \mu_3/\mu$. The system of distributions presented by Ord uses the criterion of Eq. 4.2 for selecting distributions. Another method of distinguishing amongst distributions over the range $(0,N)$ and $(0,\infty)$ is the (I,S)-plane, as shown in Fig. 4-1.

Table 4-1 Expressions for the Parameters of the Difference Eq. 4-1.

Parameter	Pearson	Discrete with Range (0,N)
a	$-(\mu_3/\mu_2)(\beta_2+3)D_p$	$(1-2b_2)\mu+1-b_1$
b_0	$\mu_2(4\beta_2-3\beta_1)/D_p$	0
b_1	$-a$	$\{\mu_2-b_2(3\mu_2+\mu^2-\mu)\}/\mu$
b_2	$(2\beta_2-3\beta_1-6)/D_p$	$\{\mu(\mu_3+\mu_2)-2\mu_2^2\}D_G$

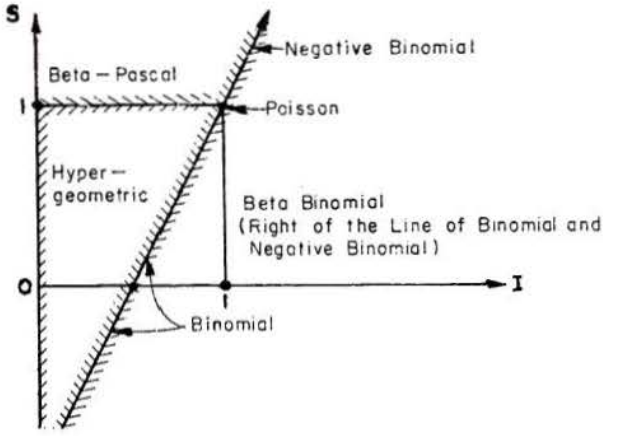

Fig. 4-1 (I-S) - Diagram for Selecting Discrete Probability Distribution Functions for Fitting Discrete Frequency Distributions of Run-Lengths.

If none of the above discrete probability distribution functions fits well a frequency distribution, measured by the chi-square statistic, it is possible to use the list of discrete distributions given by Patil, Joshi and Rao (1968) for selecting the other discrete probability distributions functions.

Other alternatives are: the polynomial expansion transformation (say, the Charlier Type-B series), or the use of mixtures of distributions. The chi-square test of goodness of fit is used in testing all fits by various probability distribution functions.

4-2 Distributions of Run-Length of Infinite Series

The case of distributions of run-length of infinite series for univariates were studied, and checked by the experimental method, by Downer, Siddiqqi and Yevjevich (1967) for the independent case, and by Saldarriaga and Yevjevich (1970) for the dependent case. The case of the bivariate independent case was studied by Llamas (1969) and Heiny (1968), with no experimental procedure used for the check. The bivariate case for mutually and serially correlated components is of the main concern in this paper. This section presents the general forms of run-length frequency distributions, with the fitted discrete probability distribution functions, and the regression relations of estimated parameters of fitted probability distributions to the parameters of the two component series.

To assess how good are the results obtained by the experimental method, the cases of the fit of known exact probability distribution functions to computed frequency distributions of runs are used, also. This gives the level of confidence in the method applied, even for cases for which either the exact or approximate analytical results cannot be obtained. Since the exact probability distributions of run-lengths are known for simple cases of underlying processes, say for the bivariate case of serially independent but mutually dependent components, the comparison of results of the experimental method with the exact distribution gives measures of the deviates of fitted probability distributions from exact probability distributions.

Exact distributions of negative-negative and negative-positive run-lengths are given in Section 2-6. A selected case is presented in Fig. 4-2 for comparison of probabilities of negative-negative and negative-positive run-lengths of the bivariate case: serially independent but mutually dependent components, with $\rho(0) = 0.7$ and truncation levels $C_1 = 0.0000$ and $C_2 = -0.38535$. The experimental frequency distributions are obtained by using the algorithms given in Section 3-3. Probability distributions, selected by criteria given in Section 4-1 for discrete distributions, are both negative binomial with the two parameters, p and r. The parameters are estimated by the method of moments. The exact distributions are obtained by using the function given in Section 2-6, with joint bivariate normal probabilities obtained from the normal distribution table. Visual inspection shows that the above three methods of computing or estimating probabilities of run-lengths are essentially identical for practical purposes. The chi-square test of goodness of fit, applied to compare the fitted probability distribution function to frequency distribution, gives the chi-square value of 2.59 for the case of negative-negative run-length and the value of 5.71 for the case of negative-positive run-length. Both are smaller than the critical value for two degrees of freedom at the 95 percent probability level of significance.

Since the fitting of this probability distribution is acceptable, the same function is fitted in all 135 cases of combinations of five parameters. All the cases are analyzed in using the Ord's approach, namely by finding whether the negative binomial distribution is acceptable,

$$f_{\chi}(x) = \binom{r+x-1}{x} p^r (1-p)^x, \qquad (4\text{-}4)$$

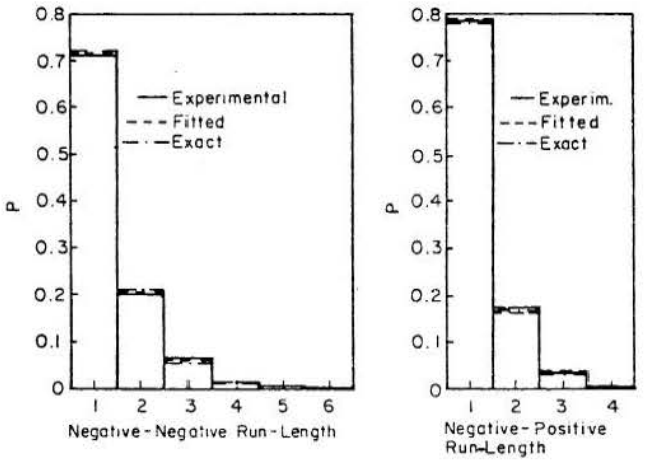

Fig. 4-2 Comparison of Experimental Frequencies, Fitted Negative Binomial Distribution to the Experimental Frequencies, and the Exact Distribution for Negative-Negative and Negative-Positive Run-Lengths for the Bivariate Process of Serially Independent but Mutually Dependent Series with $\rho(0) = 0.7$ and Truncation Levels $C_1 = 0.0000$ and $C_2 = -0.38535$.

with $\hat{p} = x/s^2$ and $\hat{r} = \bar{x}p/(1-p)$. With parameters estimated, probabilities of equal class intervals of chi-square statistics, as expected frequencies, are then computed.

The chi-squares are transformed into the corresponding probabilities by using the chi-square cumulative distribution function

$$F(\chi^2) = \frac{1}{2^{1/2}\Gamma(\frac{1}{2}\nu)} \int_0^{\chi^2} (x^2)^{1/2(\nu-2)} e^{-1/2x^2} dx^2, \qquad (4\text{-}5)$$

with ν stands the number of degrees of freedom and χ^2, the upper integral limit, the computed chi-square. Probabilities of chi-squares instead of chi-squares themselves are used as comparable measures of goodness of fit of probability functions to observed frequency distributions. Values of $F[P(\chi^2)]$ for $P(\chi^2) = 95$ greater than 50 percent were considered acceptable as approximation to the distribution desired. Probabilities of observed chi-squares are classified into ten equal class intervals in this and subsequent sections of the paper, the class frequencies of results of experimental method are determined, and the cumulative relative class frequencies computed. Results for the negative-negative and the negative-positive run-length distributions are given in Figs. 4-3 and 4-4. For the 95 percent level, 82.5 percent and 90.2 percent of computed chi-squares for the negative-negative and the negative-positive run-length distributions, respectively, were smaller than the critical chi-squares. It is concluded that the negative binomial distribution is adequate and an acceptable approximation of distributions of negative-negative and negative-positive run-lengths for the serially and mutually dependent components of a normal bivariate process, for the range of parameters and truncation levels investigated.

Instead of presenting the two estimated parameters $\hat{p}$ and $\bar{x}$ of the fitted negative binomial

distribution in tables, the multiple regression analysis is used to express these estimates in terms of parameters of the underlying bivariate process and truncation levels.

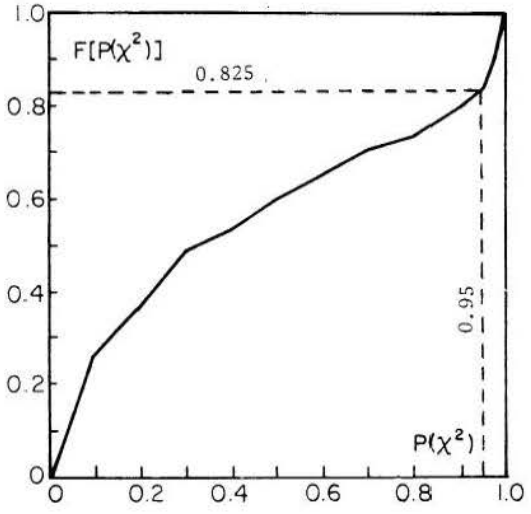

Fig. 4-3 Cumulative Distribution Curve $F[P(\chi^2)]$ of Probabilities of Chi-Squares of the Negative-Negative Run-Length, $P[\chi^2]$.

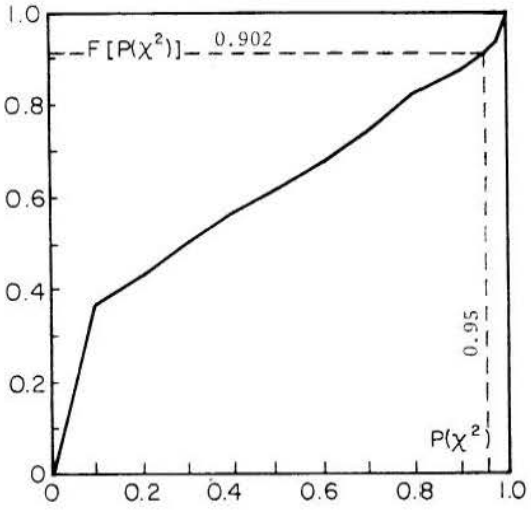

Fig. 4-4 Cumulative Distribution Curve $F[P(\chi^2)]$ of Probabilities of Chi-Squares of the Negative-Positive Run-Length, $P(\chi^2)$.

The approach used in this analysis is to express the estimated parameters as two functions

$$p = f_1[C_1, C_2, \rho_1(\varepsilon_x), \rho_1(\varepsilon_y), \rho(0)], \qquad (4\text{-}6)$$

and

$$\bar{x} = f_2[C_1, C_2, \rho_1(\varepsilon_x), \rho_1(\varepsilon_y), \rho(0)]. \qquad (4\text{-}7)$$

Stepwise multiple linear regression analyses were performed, based on Eqs. 4-6 and 4-7. The independent variables in these regression equations are the selective parameters: $C_1, C_2, \rho_1(\varepsilon_x), \rho_1(\varepsilon_y)$, and $\rho(0)$. The dependent variables are the two parameters of the negative binomial distribution, symbolized here by u, expressed in the linear form, as the first alternative:

$$u = a + b\,\rho_1(\varepsilon_x) + c\rho_1(\varepsilon_y) + d\,\rho(0) + e\,C_1 + f\,C_2. \qquad (4\text{-}8)$$

Table 4-2 gives the estimated regression coefficients, as a condensation of information on estimated probability distributions of negative-negative and negative-positive run-lengths, for the range of parameters studied. The two parameters in the negative binomial distribution of Eq. 4-4 are p and r. However, the explained variances by the multiple regressions obtained for the parameter r were smaller than for $\bar{x}$.

Table 4-2 Estimated Regression Coefficients of Eq. 4-8 for the Negative-Negative and Negative-Positive Run-Lengths.

u	a	b	c	d	e	f	R^2
p_{nn}	.79046	-.21027	-.21798	-.21702	-.16930	-.13896	.9244
$\bar{x}_{nn}$	.33064	.38954	.39343	.32733	.32756	.31954	.9127
p_{np}	.70729	-.30926	-.05499	.23828	-.20907	.16244	.9451
$\bar{x}_{np}$	.40699	.40253	.03957	-.35300	.38565	-.30129	.9137

Similarly, the stepwise multiple linear regression analysis, based on Eqs. 4-6 and 4-7, to determine whether the use of probabilities of quantiles q_1 and q_2 as independent variables, and corresponding to the two truncation levels C_1 and C_2 together with $\rho_1(\varepsilon_x)$, $\rho_1(\varepsilon_y)$, and $\rho(0)$ would give a larger explained variance R^2 than in the case of using C_1 and C_2. The dependent variables were again p and $\bar{x}$, expressed in the form

$$u = a + b\,\rho_1(\varepsilon_x) + c\rho_1(\varepsilon_y) + d\,\rho(0) + e\,q_1 + f\,q_2. \qquad (4\text{-}9)$$

Table 4-3 gives the estimated regression coefficients. Since the results are similar to those obtained using the truncation levels, only the quantiles will be used in the remaining parts of this paper. It should be stressed that the multiple nonlinear regression analysis was also investigated, resulting in the lower values of R^2 than obtained for the selected multiple linear regression.

Table 4-3 Estimated Regression Coefficients of Eq. 4-9 for the Negative-Negative and Negative-Positive Run-Lengths.

u	a	b	c	d	e	f	R^2
p_{nn}	1.21820	-.21027	-.21798	-.21702	-.47400	-.38791	.9191
$\bar{x}_{nn}$	-.57004	.38954	.39343	.32733	.91963	.89756	.9194
p_{np}	.77257	-.30926	-.05499	.23828	-.58807	.45602	.9459
$\bar{x}_{np}$	.28646	.40253	.08957	-.35300	1.08919	-.84341	.9168

4-3 Distributions of Longest Run-Length in Samples of Given Sizes

The case of the longest run-length in the sample of size n for the independent univariate process has been studied at length in the past. The univariate dependent process was investigated experimentally by Millan and Yevjevich (1971) and by Millan (1972). The probability distribution functions of the longest negative run-length in n observations, obtained by the experimental method, were fitted by a lognormal distribution function even though it was recognized that this statistic is a discrete random variable (only positive integers are random events). Distributions of the longest run-length of a given type in samples of a bivariate process have not been yet studied. Its analytical treatment either in an exact or in an approximate way, as stated in Chapter II, consists of four combinations of serial and mutual dependence of the two components. This section gives the general forms of frequency distributions of the negative-negative and negative-positive longest run-length in samples of size n, fitted but the approximate probability distributions, and the multiple regression equations of estimated parameters of fitted distribution functions in terms of parameters of assumed underlying bivariate processes.

Similarly as for the case of run-lengths for infinite series, the results of the experimental method were checked by using the case of the known probability distribution function of the longest run-length. The case used in the bivariate process of serially independent but mutually correlated series, with the exact results given in Section 2-3 for both the longest negative-negative and the longest negative-positive run-lengths in samples of size n, $\rho(0) = 0.7$, $\rho_1(\varepsilon_x) = \rho_1(\varepsilon_y) = 0.0$, and $C_1 = C_2 = 0.0$. The frequency distributions are obtained by using the algorithm given in Section 3-3. Figure 4-5 gives the comparison of the experimental frequencies, probabilities of fitted function (mixture of geometric distributions), and exact probabilities.

The fit of discrete probability distributions is more complex for the negative-negative and negative-positive longest run-lengths in case of samples than in case of infinite series. The analysis by using the family of discrete distributions, as given by Ord, inferred that the function should be of the Beta-Pascal type. When its parameters were estimated by the method of moments, square roots of negative numbers were obtained, making the fit impossible. The reason for this was that the values of S and I were near the boundary with the hypergeometic distribution. The fit of hypergeometric distribution [Patil and Joshi (1968)], produced similar results. The attempt to use the series expansion approach of the Charlier Type B series, as given by Kendall (1943), gave similar results as the binomial distribution which were tried initially, with the chi-squares much greater than the critical chi-squares. The fit of discrete distributions of the Hyper-Poisson family, given by Bardwell and Crow (1964), gave similar results as the use of Charlier Type B series with no reduction in probability of chi-squares. The method of moments estimation of parameters was used for all the above distributions. A continuous distribution was also used with the understanding that it would be only an approximation to discrete distributions, and that probability densities multiplied by the unit-time interval around the integer values would represent the probability mass at the integer value. Also, this

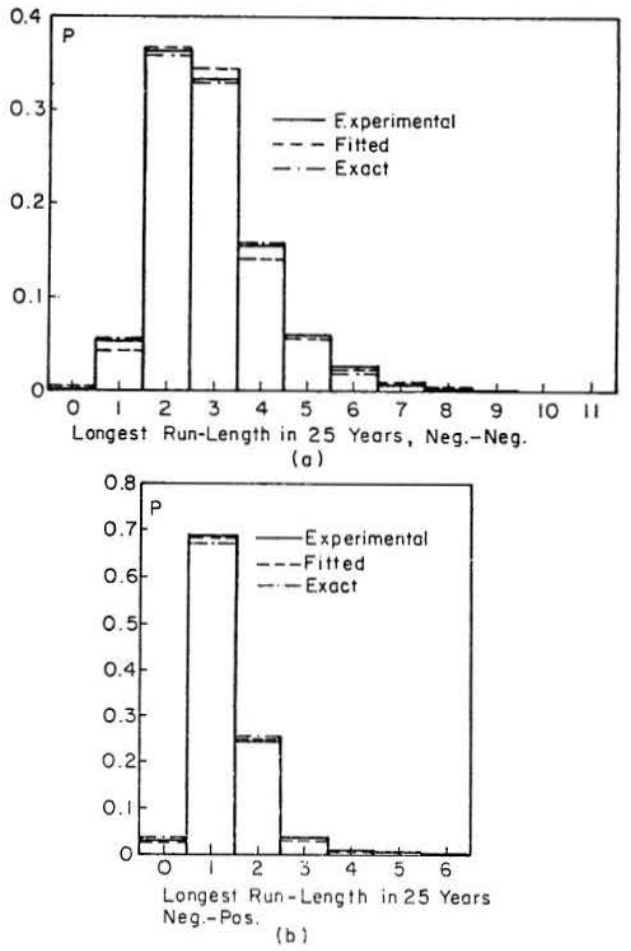

Fig. 4-5 Comparison of Experimental Frequencies, Probabilities of the Fitted Distribution Function (a mixture of two geometric distributions) to Experimental Frequencies and Probabilities of Exact Distribution of the Negative-Negative and Negative-Positive Longest Run-Length in 25 Years for the Bivariate Process with Serially Independent but Mutually Dependent Component Series with $\rho(0) = 0.7$ and Truncation Probabilities, $q_1 = q_2 = 0.0$.

approach failed to pass the chi-square test. A mixture of discrete distributions was then used.

A visual inspection of experimentally obtained frequency distributions suggested the use of a mixture of two geometric distributions: left side with a truncated geometric distribution and right side with a standard geometric distribution. The discrete distribution of this mixture is suggested to the writer by D. Boes in 1973, as

$$f_X(x) = \alpha \frac{(1-\theta_1)\,\theta_1^{\gamma-x}}{1-\theta_1^{\gamma+1}} I_{\{0,1\ldots,\gamma\}} + (1-\alpha) \frac{\theta_2(1-\theta_2)^x}{(1-\theta_2)^{\gamma+1}} I_{\{\gamma+1,\ldots,\}}, \qquad (4\text{-}10)$$

with θ_1 and θ_2 parameters of each part, respectively, γ a location parameter and α a partition parameter. The estimation of these parameters is made by the maximum likelihood method. The location γ is estimated either by the mode $\hat{\gamma} = m$ or by $\hat{\gamma} = m-1$, α is estimated by $\hat{\alpha} = \sum_{i=1}^{\gamma} p_i$, θ_2 by $(\bar{x} - \hat{\gamma})^{-1}$ with $\bar{x}$ the mean of observations greater than $\hat{\gamma} + 1$, and θ_1 by an iterative solution of

$$\frac{1}{1 - \hat{\theta}_1} = \frac{\hat{\gamma} - \bar{x}_1}{\hat{\theta}_1} + \frac{(\hat{\gamma}+1)\hat{\theta}_1^{\gamma}}{1-\hat{\theta}_1^{\gamma+1}}, \qquad (4\text{-}11)$$

with $\bar{x}_1$ the mean of observations less than $\hat{\gamma}$.

Exact distributions of Fig. 4-5 for the negative-negative and negative-positive longest run-length, respectively, were obtained by using the distribution given in Section 2-3, with joint probabilities of the bivariate obtained from tables of normal distribution and Eq. 2-15.

Visual inspection of Fig. 4-5 shows the three compared distributions to be close for practical purposes. The chi-square test of the goodness of fit of selected mixed distribution function to fit the experimental frequency distribution was used. For a sample size of 25, the chi-square was 6.34 for the longest negative-negative run-length, and 0.24 for the longest negative-positive run-length, both being smaller than the corresponding critical chi-squares of 7.815 and 3.841, respectively, for three and one degrees of freedom at the 95 percent significance level.

Similarly as for the case of run-length of infinite series, the estimated parameters of fitted probability distributions are related to parameters of the underlying bivariate processes, instead of presenting the graphs of experimental frequency distributions. For the 405 different combinations of basic parameters the samples were generated, the frequency distributions obtained and the mixture probability distribution functions fitted.

The parameter γ was estimated either by the mode m, as $\hat{\gamma} = m$, or by $\hat{\gamma} = m-1$, whichever gave the smallest chi-square value. The other parameters were estimated as described previously, and probabilities as the expected frequencies are computed for the chi-square test.

Similarly as for distributions of run-length of an infinite series, the computed chi-square values were transformed to their corresponding probabilities by using the chi-square cumulative distribution function of Eq. 4-5. Probabilities of observed chi-squares are classified into ten equal class intervals, the corresponding observed class frequencies determined and the cumulative relative class frequencies computed. Results of probabilities of chi-square for distributions of the negative-negative and negative-positive longest run-length are shown in Figs. 4-6 and 4-7. At the 95 percent level, 68.3 percent and 50.2 percent of the computed chi-squares for the longest negative-negative and negative-positive run length were smaller than the critical values. The fitted distributions are accepted as adequate approximations for the experimentally derived frequency distributions of the negative-negative and negative-positive longest run-length in n years for a normal bivariate process, in the range of parameters and truncations investigated.

Following the method of previous analyses for runs of infinite series, parameters of the fitted probability distribution functions for the longest

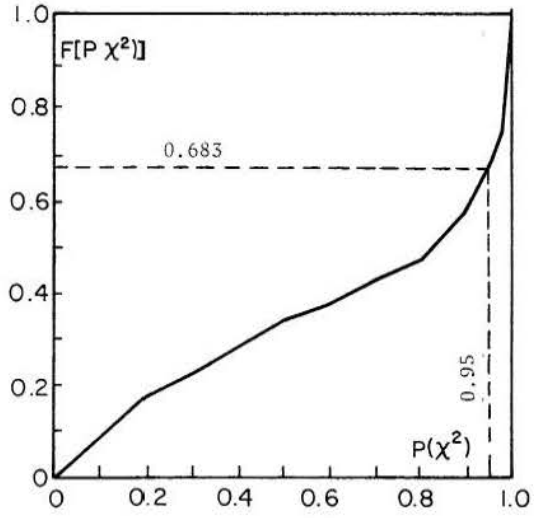

Fig. 4-6 Cumulative Distribution Curve $F[P(\chi^2)]$ of Probabilities $P(\chi^2)$ of Chi-Squares for the Longest Negative-Negative Run-Length.

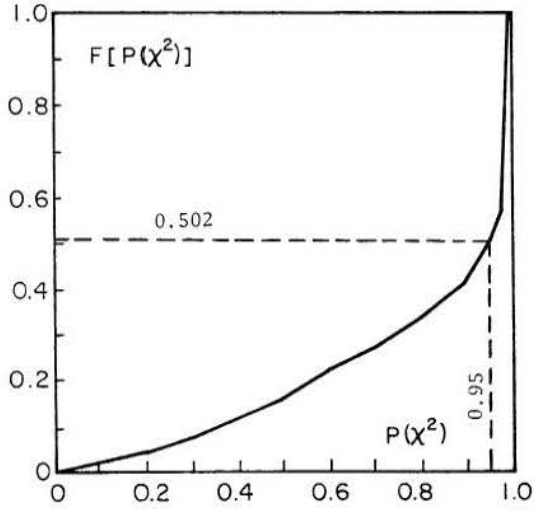

Fig. 4-7 Cumulative Distribution Curve $F[P(\chi^2)]$ of Probabilities $P(\chi^2)$ of Chi-Squares for the Longest Negative-Positive Run-Length.

run-length of given sample sizes are related by the multiple linear regression to parameters of underlying normal bivariate process, the truncation levels and sample size, expressing the estimated parameters in the form of Eqs. 4-6 and 4-7.

Stepwise multiple linear regression analyses were used, with independent variables being q_1 and q_2, the quantile probabilities for truncation levels, $\rho(0)$, the lag-zero cross correlation coefficient between the serially independent stochastic components, and the serial correlation coefficients $\rho_1(\varepsilon_x) = \rho_1(\varepsilon_y) = 0.0$. Dependent variables were the estimated parameters of fitted distribution functions, symbolized by u, in the form

$$u = a+b\rho_1(\varepsilon_x)+c\rho_1(\varepsilon_y)+d\rho(0)+eq_1+fq_2+g\ \log_{10}n. \qquad (4\text{-}12)$$

Tables 4-4 and 4-5 give the estimated regression coefficients. The form of Eq. 4-12 was obtained after different trials with the multiple nonlinear regressions.

Table 4-4 Estimated Regression Coefficients of Eq. 4-12 for the Negative-Negative Longest Run-Length in a Sample of Size n.

u	a	b	c	d	e	f	g	R^2
γ	-4.03751	.67364	.61192	.98148	2.96296	2.83951	1.75236	.8233
α/γ	.69835	-.07551	-.08227	-.12509	-.31244	-.31332	-.11409	.7906
θ_2	1.28471	-.23525	-.22875	-.21906	.52795	-.47152	-.02546	.9036
θ_1/γ	.08726	.04108	.04209	-.01139	-.00463	-.01392	-.02972	.2169

Table 4-5 Estimated Regression Coefficients of Eq. 4-12 for the Negative-Positive Longest Run-Length in a Sample of Size n.

u	a	b	c	d	e	f	g	R^2
γ	-.74931	.18519	.67901	-1.11111	-3.08642	3.72840	1.42368	.8042
α/γ	.50198	-.02325	-.08476	.13836	.34381	-.54054	-.14312	.7985
θ_2	.32634	-.08529	-.27450	.20982	.56094	-.74094	-.03070	.9318
θ_1/γ	.10310	.00701	.05416	.00925	-.01177	-.04440	-.03424	.1596

4-4 Fitting Continuous Probability Distribution Functions to Frequency Distributions of Run-Sums and Run-Intensities

Run sums and run-intensities are continuous random variables, so that continuous distribution functions must be fitted to their experimental frequency distributions. In fitting probability functions to experimental curves, two approaches are used in this study: (1) Pearson family of distribution functions (Pearson, 1895), and (2) probability functions transformed by polynomials. The Pearson family of functions has been discussed by many authors, notably by Elderton (1953), Elderton and Johnson (1969), Kendall and Stuart (1969), Johnson and Kotz (1969b), etc., with detailed analyses available.

The series expansion approach assumes that an arbitrary density function, h(x), can be represented by a series based on a known density function, say the normal density function, in the form

$$h(x) = f(x) \sum_{j=0}^{\infty} C_j H_j(x) , \qquad (4\text{-}13)$$

with $H_j(x)$ polynomials of the order j in x, and C_j the coefficients which depend on the type of polynomial in Eq. 4-13. This approach is used for the joint distribution of run-sums of infinite series.

4-5 Distributions of Run-Sums and Run-Intensities of Infinite Series

Distributions of run-sums of infinite series of independent univariate normal processes were studied analytically by Downer, Siddiqqi and Yevjevich (1967) and the results checked experimentally. This case was also studied by Llamas (1968) for the independent gamma variables. Distributions of run-sums of univariate dependent processes are obtained by Heiny (1970) in an approximate form. The bivariate process, with components mutually and serially independent, was analyzed by Llamas (1968) and Llamas and Siddiqqi (1969) only for the negative-negative run-sum. The serially independent but mutually dependent components of the bivariate process were analyzed by Llamas (1968) and Heiny (1970) only for the negative-negative run-sum.

In this paper, the runs of bivariate normal processes of mutually and serially dependent components are investigated. Because the analytical treatment is complex, experimental method is used to obtain frequency distributions for selected cases. Experimental results are checked by using the properties of run-length, as explained in Section 4-3. Joint frequencies of the corresponding run-lengths and run-sums of series 1 and 2 are obtained for the negative-negative and negative-positive cases. Distributions of the joint frequencies of run-sums and run-intensities, obtained experimentally, were fitted by gamma functions with Laguerre polynomials.

Distribution functions of gamma type with Laguerre polynomials are used for the negative-negative and negative-positive run-sums. The reason is that the bivariate normal distribution was not giving an adequate fit, and the fits of the bivariate gamma distributions, given by Ord and Mardia (1970), were not acceptable. Marginal distributions fitted by the Pearson Type III probability function did not pass the chi-square test of goodness of fit; however, the obtained chi-square values were close to critical values. After these attempts, the two-parameter gamma distribution function was used for the marginal distributions and the product of two gamma functions with Laguerre polynomials for joint distributions. This approximation to joint probability density function with series expansion is

$$f_{X,Y}(x,y) = f_1(x) f_2(y) \sum_j \sum_k a_{jk} L_j^{(\beta_1 - 1)}(\lambda_1 x) L_k^{(\beta_2 - 1)}(\lambda_2 y), \qquad (4\text{-}14)$$

with $L_j^{(\beta_1 x)}(\lambda_1 x)$ and $L_k^{(\beta_2 y)}(\lambda_2 y)$ the Laguerre polynomials of degrees j and k, respectively. A Laguerre polynomial of degree m, $L_m^{(c)}(z)$, is expressed by expansion in power series of z as

$$L_m^{(c)}(z) = z^m - \frac{m}{1!}(m+c) z^{m-1} + \frac{m(m-1)}{2!}(m+c)(m+z-1) z^{m-2} - \dots . \qquad (4\text{-}15)$$

For $c > -1$, the polynomials $L_m^{(c)}(z)$, $(m = 0,1,2,\dots,)$, form an orthogonal system on the semi-axis $(0,\infty)$, with the weight function $f_1(z)$,

$$\int_0^{\infty} L_j^{(c_1)}(z) L_m^{(c)}(z) f_1(z)\, dz = 0, \quad \text{if } j \neq m, \qquad (4\text{-}16)$$

$$= d_j^2, \quad \text{if } j = m. \qquad (4\text{-}17)$$

For

$$f_1(z) = z^s e^{-z}, \text{ then}$$

$$d_m^2 = m!\,\Gamma(c+m+1) . \qquad (4\text{-}18)$$

Coefficients a_{jk} can be estimated by taking the expected value of $L_j^{(\beta_1-1)}(\lambda_1 x)\, L_k^{(\beta_2-1)}(\lambda_2 y)$, namely as

$$E[L_j^{(\beta_1-1)}(\lambda_1 x)\, L_k^{(\beta_2-1)}(\lambda_2 y)]$$

$$= \int_0^\infty \int_0^\infty L_j^{(\beta_1-1)}(\lambda_1 x) L_k^{(\beta_2-1)}(\lambda_2 y)\, f(x,y)\,dx\,dy \ . \qquad (4\text{-}19)$$

Replacing $f_{X,Y}(x,y)$ by its value in Eq. 4-17, and considering that the postulated marginal distributions are

$$f_1(x) = \frac{e^{-\lambda_1 x}(\lambda_1 x)^{\beta_1-1}\lambda_1}{\Gamma(\beta_1)} \ , \qquad (4\text{-}20)$$

and

$$f_2(y) = \frac{e^{-\lambda_2 y}(\lambda_2 y)^{\beta_2-1}\lambda_2}{\Gamma(\beta_2)} \ , \qquad (4\text{-}21)$$

then Eq. 4-19 becomes

$$E[L_j^{(\beta_1-1)}(\lambda_1 x)\, L_k^{(\beta_2-1)}(\lambda_2 y)]$$

$$= \sum_j \sum_k a_{jk} \int_0^\infty L_j^{(\beta_1-1)}(\lambda_1 x) L_m^{(\beta_1-1)}(\lambda_1 x) \frac{e^{-\lambda_1 x}(\lambda_1 x)^{\beta_1-1}\lambda_1}{\Gamma(\beta_1)} dx$$

$$\cdot \int_0^\infty L_j^{(\beta_2-1)}(\lambda_2 y) L_n^{(\beta_2-1)}(\lambda_2 y) \frac{e^{-\lambda_2 y}(\lambda_2 y)^{\beta_2-1}\lambda_2}{\Gamma(\beta_2)} dy \ . \qquad (4\text{-}22)$$

Taking into account Eqs. 4-16 and 4-17, the expected value is

$$E[L_j^{(\beta_1-1)}(\lambda_1 x)\, L_k^{(\beta_2-1)}(\lambda_2 y)] = \frac{a_{jk}}{\Gamma(\beta_1)\Gamma(\beta_2)}\, c_j^2\, d_k^2 \ , \qquad (4\text{-}23)$$

or

$$a_{jk} = \frac{\Gamma(\beta_1)\,\Gamma(\beta_2)}{C_j^2\, d_k^2} E[L_j^{(\beta_1-1)}(\lambda_1 x)\, L_k^{(\beta_2-1)}(\lambda_2 y)]. \qquad (4\text{-}24)$$

Coefficients a_{jk} are obtained for the selected values of j and k in Eq. 4-24. Using values up to $j = k = 3$, and simplifying Eq. 4-24, then

$$a_{00} = 1; \quad a_{10} = a_{01} = a_{20} = a_{02} = 0 \ ,$$

$$a_{11} = \frac{\lambda_1\lambda_2}{\beta_1\beta_2} E(XY) - 1, \text{ with } \lambda_1 = \frac{E(X)}{Var(X)} \ , \ \beta_1 = \lambda_1 E(X),$$

$$\lambda_2 = \frac{E(Y)}{Var(Y)} \ , \quad \beta_2 = \lambda_2 E(Y) \ ,$$

$$a_{21} = \frac{\lambda_1^2\lambda_2}{2(\beta_1+1)\beta_1\beta_2} E(X^2Y) - \frac{\lambda_1\lambda_2}{\beta_1\beta_2} E(XY) + \tfrac{1}{2} \ ,$$

$$a_{12} = \frac{\lambda_1\lambda_2^2}{2(\beta_2+1)\beta_1\beta_2} E(XY^2) - \frac{\lambda_1\lambda_2}{\beta_1\beta_2} E(XY) + \tfrac{1}{2}$$

$$a_{03} = \frac{1}{6\beta_2(\beta_2+1)(\beta_2+2)} \lambda_2^3 E[Y^3] - \frac{1}{2\beta_2(\beta_2+1)} \lambda_2^2 E(Y^2) + \frac{1}{3},$$

$$a_{30} = \frac{1}{6\beta_1(\beta_1+1)(\beta_1+2)} \lambda_1^3 E[X^3] - \frac{1}{2\beta_1(\beta_1+1)} \lambda_1^2 E(X^2) - \frac{1}{3} \ . \qquad (4\text{-}25)$$

Nine parameters of probability distributions to fit the joint frequency distributions of run-sums and run-intensities are $E(X)$, $E(Y)$, $E(X^2)$, $E(Y^2)$, $E(XY)$, $E(X^2Y)$, $E(XY^2)$, $E(X^3)$, $E(Y^3)$.

Nine parameters of joint distributions of negative-negative and negative-positive run-sums were computed for the 405 cases of experimentally generated large samples. They gave the expected frequencies for the chi-square tests of goodness of fit of these functions. Computed chi-square values are transformed into their corresponding probabilities by using Eq. 4-5. They are classified into ten equal class intervals with their class frequencies determined and the cumulative frequencies computed. Results for the negative-negative and negative-positive joint distributions of run-sums are presented in Figs. 4-8 and 4-9. At the 95 percent level, 71.5 and 72.3 percent of computed chi-squares were smaller than the critical values. The fits of gamma functions with Laguerre polynomials are accepted as satisfactory approximations.

Stepwise multiple linear regression analyses were made to express the estimated parameters of the joint probability distribution functions, as dependent variables, in terms of exact parameters of the two series and the corresponding truncation levels, of the type of Eq. 4-8. Independent variables were the same as in Section 4-2. Tables 4-6 and 4-7 give the estimated regression coefficients, which represent a condensation of information on sampling distributions of the joint negative-negative and negative-positive run-sums, respectively. In these tables, D_1 and D_2 are the deficit in series 1 and 2, respectively, and S_2 is the surplus in series 2. Also the R^2 values are given.

For the case of joint distributions of run-intensities the same approach is used as for joint distributions of run-sums. Figures 4-8 and 4-9 show the cumulative relative class frequencies of probabilities of obtained chi-squares. At the 95 percent level, 77.3 and 79.4 percent of the computed chi-squares were smaller than the critical chi-squares. The fits of gamma functions with Laguerre polynomials are accepted as satisfactory approximations. Stepwise multiple linear regression analyses were performed to obtain equations of the type of Eq. 4-8. Independent variables are the same as in Section 4-2, with dependent variables being the estimated parameters of these joint probability distribution functions. Tables 4-8 and 4-9 give the estimated regression coefficients and

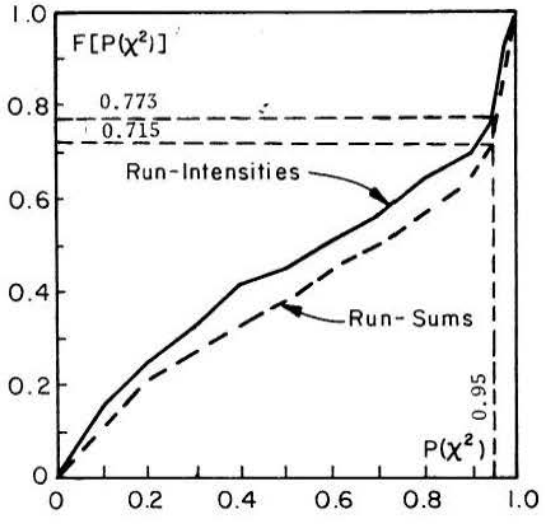

Fig. 4-8 Cumulative Distribution Curve $F[P(\chi^2)]$ of Probabilities $P(\chi^2)$ of Chi-Squares for the Joint Distribution of Negative-Negative Run-Sums and Negative-Negative Run-Intensities for an Infinite Population.

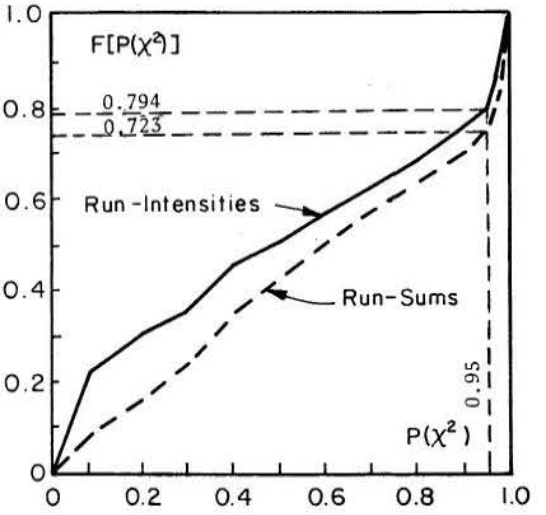

Fig. 4-9 Cumulative Distribution Curve $F[P(\chi^2)]$ of Probabilities $P(\chi^2)$ of Chi-Squares for the Joint Distribution of Negative-Positive Run-Sums and Negative-Positive Run-Intensities for an Infinite Population.

the R^2 values. A comparison of results of the regression analysis of parameters of the distribution of the joint negative-negative and negative-positive run-sums presented in Tables 4-6 and 4-7 with the results of the regression analysis of parameters of the distribution of the negative-negative and negative-positive run-intensities presented in Tables 4-8 and 4-9 shows the higher values of R^2 for the intensities than for the sums. This can be explained by the fact that the existing correlation between the run-length and its corresponding run-sum produces a small sampling variation in their ratios.

4-6 Distributions of Largest Run-Sum in Samples of Given Sizes

The case of distributions of run-sums for given sample sizes, particularly the largest run-sum in n years, was studied by Millan and Yevjevich (1970) for the univariate processes by the experimental method. The analytical treatment in simple form seems not feasible. Distribution functions of the largest run-sum of the independent and serially dependent univariate normal processes, as obtained by Millan and Yevjevich, are fitted by the lognormal distribution function with the use of the Smirnov-Kolmogorov goodness-of-fit test.

Table 4-6 Estimated Regression Coefficients of Equation 4-9 for the Joint Distribution of the Negative-Negative Run-Sum.

Parameter	a	b	c	d	e	f	R^2
$E(D_1)$	- 0.13828	.29217	.32565	.47747	2.27276	.24389	.9541
$E(D_1^2)$	- 3.45871	2.24692	1.50811	2.08177	9.39387	2.13814	.8811
$E(D_2)$	- 0.14848	.33977	.28768	.49921	.26572	2.28574	.9560
$E(D_2^2)$	- 3.66554	1.58701	2.23931	2.30118	2.36588	9.52111	.8834
$E(D_1D_2)$	- 3.55407	1.65397	1.60771	2.21931	5.31621	5.23639	.8519
$E(D_1^2D_2)$	-21.47830	10.48223	8.00554	10.93330	30.38095	22.21412	.7396
$E(D_1D_2^2)$	-21.91244	8.44969	10.19924	11.33341	23.26341	30.01339	.7389
$E(D_1^3)$	-23.77062	11.43511	15.57942	7.52094	46.67559	14.87846	.7796
$E(D_2^3)$	-25.46707	12.99252	8.21255	15.52531	16.93461	47.35485	.7781

Table 4-7 Estimated Regression Coefficients of Equation 4-9 for the Joint Distribution of the Negative-Positive Run-Sum.

Parameter	a	b	c	d	e	f	R^2
$E(D_1)$	.80961	.21366	.06127	- .75001	1.51613	- .79645	.9428
$E(D_1^2)$	1.35259	1.16286	.15491	-2.28541	4.57479	- 2.66804	.8313
$E(S_2)$	1.66053	.29840	.06292	-1.15574	1.30632	- 1.92101	.9387
$E(S_2^2)$	3.98495	1.51274	.47539	-4.02996	5.12212	- 6.88044	.8130
$E(D_1S_2)$	1.78577	1.14991	.22799	-2.39263	3.98309	- 3.60603	.7970
$E(D_1^2S_2)$	4.18571	5.20360	.61882	-7.70233	13.97333	-11.73363	.6358
$E(D_1S_2^2)$	6.24975	5.71521	1.27931	-9.64958	15.92044	-16.02572	.6313
$E(D_1^3)$	3.77980	-8.66343	6.19092	.32688	17.05712	-10.96483	.6804
$E(S_2^3)$	13.82225	-17.25806	8.25806	3.02832	24.16221	-30.38923	.6642

Distributions of the largest run-sum of a given type for given sample sizes of an independent bivariate normal process have not been studied either analytically or experimentally, because they are much more complex cases than the cases of the univariate normal process. The bivariate normal dependent processes have not been studied either. This section shows only the general form of sampling distributions of the largest negative-negative run-sum and the largest negative-positive run-sum in samples of size n. The fitted probability distributions are obtained as gross approximations, and the multiple linear regressions are given between the estimated parameters for the bivariate case and the parameters of the underlying processes, similarly as it was done in previous sections.

Since no analytical exact distributions of the largest run-sum is available for checking purposes,

Table 4-8 Estimated Regression Coefficients of Equation 4-9 for the Joint Distribution of the Negative-Negative Run-Intensities.

Parameter	a	b	c	d	e	f	R^2
$E(I_1)$	.47401	.01053	.10115	.15554	- .36964	1.01882	.9527
$E(I_1^2)$	.38111	-.02097	-.19865	.24595	- .75152	1.83578	.9370
$E(I_2)$	.46436	-.09920	.00297	.15288	1.01316	- .36285	.9547
$E(I_2^2)$	.38478	-.19402	-.03291	.21926	1.79156	- .73459	.9368
$E(I_1 I_2)$	.10585	-.09072	-.10057	.35195	.52130	.52115	.9609
$E(I_1^2 I_2)$	-.02215	-.15602	-.20465	.54794	.30227	1.18212	.9653
$E(I_1 I_2^2)$	-.01154	-.19210	-.17135	.52624	1.16466	.30407	.9631
$E(I_1^3)$	.42313	.37252	-.08313	-.35952	-1.36124	3.06647	.9219
$E(I_2^3)$	.45223	.29816	-.34514	-.10170	2.96276	-1.32683	.9210

Table 4-9 Estimated Regression Coefficients of Equation 4-9 for the Joint Distribution of the Negative-Positive Run-Intensities.

Parameter	a	b	c	d	e	f	R^2
$E(I_1)$	1.19141	- .02739	-.02169	- .63241	.34307	- .91721	.9731
$E(I_1^2)$	1.61523	- .08231	-.03027	-1.09912	.49313	-1.53274	.9574
$E(I_2)$	.57319	- .06822	.00614	- .37257	.60713	- .18799	.9690
$E(I_2^2)$	.52111	- .12128	------	- .55067	.83410	- .21583	.9611
$E(I_1 I_2)$	.64442	- .05133	-.00514	- .57154	.53718	- .56824	.9712
$E(I_1^2 I_2)$	.84367	- .07991	-.00784	- .79648	.58710	- .85141	.9428
$E(I_1 I_2^2)$	.55261	- .07996	-.01116	- .61613	.62530	- .48747	.9516
$E(I_1^3)$	2.42237	-1.80594	-.17756	- .03915	.70077	-2.45258	.9389
$E(I_2^3)$	.60376	- .79708	-.20022	- .01025	1.14461	- .25024	.9492

the results must rely on the check with the longest run-length of the same series.

The Pearson family of distribution functions was used, with the available criteria for identifying its type of the best fit to frequency distributions obtained by the experimental method. The parameters of these functions were estimated by the method of moments, as a characteristic of the Pearson approach. The chi-square test of goodness-of-fit was used with the 95 percent significance level. Tables 4-10 and 4-11 show the number of cases fitted by Pearson Type VI, I, and IV distribution functions. The total number of simulated cases is 405.

Table 4-10 Pearson Type VI, I and IV Probability Distribution Functions Fitted to the Experimental Frequency Distributions of Negative-Negative Largest Run-Sums.

Function type \ Sample Size n	25	50	200	Total
VI	80	45	29	154
I	50	88	103	241
IV	5	2	3	10
				405

Table 4-11 Number of Cases Fitted by Pearson Type VI, I, and IV Probability Distribution Functions Fitted to the Empirical Frequency Distributions of Negative-Positive Largest Run-Sums.

Function type \ Sample Size n	25	50	200	Total
VI	1	23	59	83
I	-	2	39	41
IV	134	110	37	281
				405

To have the same distribution for different sample sizes, Type IV was chosen for the negative-positive largest run-sum, by having 281 cases of good fit out of 405 cases, or 69.38 percent. Similarly, Type I was chosen for the negative-negative largest run-sum, by having 241 cases of good fit out of 405 cases or 59.50 percent.

The Pearson Type IV distribution function, with the origin at the mean, is

$$y = y_o\{1 + (\frac{x}{a} - \frac{\nu}{r})^2\}^{-\frac{1}{2}(r+2)} e^{-\nu \tan^{-1}(\frac{x}{a} - \frac{\nu}{r})}, \qquad (4\text{-}26)$$

with y_o, a, ν and r the parameters. These parameters were estimated by the method of moments. The parameters β_1 and β_2 are defined by the second, third and fourth moments about the mean, as

$$\beta_1 = \frac{\mu_2^2}{\mu_3^3}, \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}, \qquad (4\text{-}27)$$

and estimated by the corresponding sample moments. The distribution parameters are then

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6},$$

$$\nu = \frac{r(r-2)\sqrt{\beta_1}}{\sqrt{16(r-1) - \beta_1(r-2)^2}}, \qquad (4\text{-}28)$$

and $$a = \sqrt{\frac{\mu_2}{16}}\sqrt{16(r-1) - \beta_1(r-2)^2}.$$

The μ_3 and ν have the opposite signs. The parameter y_o can be obtained by using the function $H(r,\nu)$, by integrating the values y/y_o and weighting the density function accordingly. The integration was performed numerically using the Simpson rule, obtaining the area under the curve y/y_o, and from it the corresponding weighting factor.

Results of probabilities of chi-square obtained for the distribution of the negative-positive largest run-sum are presented in Fig. 4-10. At the 95 percent level, only 39.1 percent of the computed chi-squares were smaller than the critical chi-squares. Pearson Type IV distribution function was found to be closest approximation to the frequency distributions of the negative-positive largest run-sum in samples of n

years for serially and mutually dependent components of a bivariate normal process.

Estimated parameters of the Pearson Type IV distribution function are related by the multiple regression equation to parameters of the bivariate process and the two truncation levels. Stepwise multiple regression analysis was used in the form of Eq. 4-12. Independent variables were same as in Section 4-3. Dependent variables were the estimated parameters r, ν, a, and y_o. Table 4-12 gives the estimated regression coefficients.

Multiple regression analysis was performed also for β_1 and β_2 parameters as dependent variables, with the same independent variables as above, by using Eq. 4-12. Table 4-13 gives the obtained regression coefficients. The multiple correlation coefficients are very low in this case.

Table 4-12 Estimated Regression Coefficients of Eq. 4-12 for Parameters of Pearson Type IV Distribution Function Fitted to Frequency Distributions of the Negative-Positive Largest Run-Sum in n Years.

Parameter	a	b	c	d	e	f	g	R^2
a	2.60315	- .05298	.22059	-1.10793	- .61857	1.01427	- .63704	.7250
$\nu^{1/3}$	2.70117	.21650	- .74583	2.00844	3.30648	-7.05950	2.17734	.7860
r	4.70243	-2.52153	-1.25596	-4.75224	-4.45692	11.27875	1.85438	.5048
y_o	1.08718	.15227	- .35452	.80765	1.02957	-1.95815	- .38512	.5434

Table 4-13 Estimated Regression Coefficients of Eq. 4-12 for Parameters β_1 and β_2 of Frequency Distribution of the Negative-Positive Largest Run-Sum.

Parameter	a	b	c	d	e	f	g	R^2
β_1	-2.73505	.15755	.70058	.40378	-1.09867	2.47199	1.67406	.3155
β_2	- .51374	1.84089	1.98578	5.94192	1.38068	-2.25997	2.39163	.3468

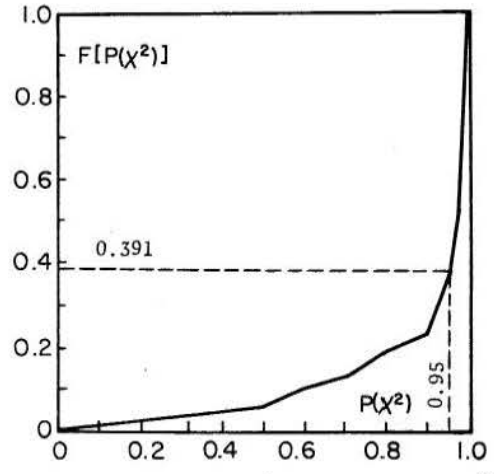

Fig. 4-10 Cumulative Distribution Curve $F[P(\chi^2)]$ of Probabilities $P(\chi^2)$ of Chi-Squares of the Largest Negative-Positive Run-Sum in Using the Pearson Type IV Function.

Figures 4-11 and 4-12 show comparisons of experimental cumulative frequency distributions of the largest negative-positive run-sum in the samples of

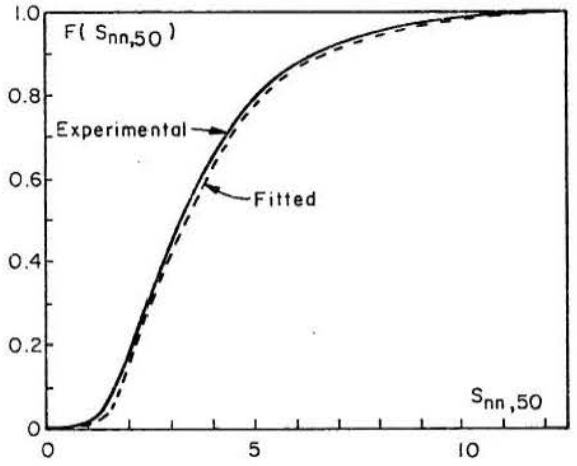

Fig. 4-11 Comparison of the Cumulative Frequency Distribution of Negative-Positive Largest Run-Sum in a Sample of 50 Years with a Fitted Cumulative Pearson Type IV Distribution Function, with Parameters a = 1.4354, ν = 2.5801, r = 5.8329, and y_o = 4.3655.

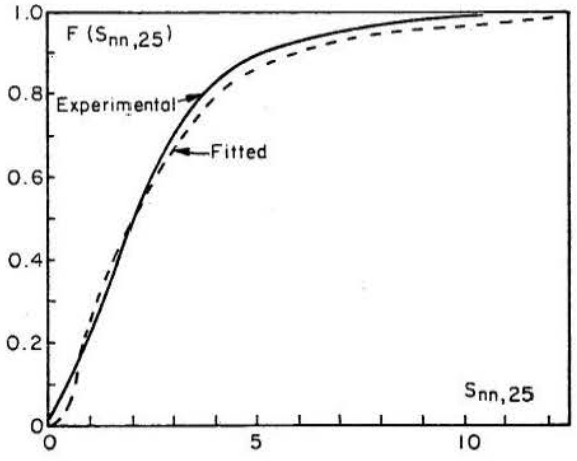

Fig. 4-12 Comparison of the Cumulative Frequency Distribution of Negative-Positive Largest Run-Sum in a Sample of 25 Years with a Fitted Cumulative Pearson Type IV Distribution Function, with a = 0.6505, ν = 0.4248, r = 3.5160 and y_o = 1.2110.

50 and 25 years, with the fitted cumulative Pearson Type IV distribution functions. For the case of a bivariate process, serially and mutually dependent, with $\rho_1(\varepsilon_x) = \rho_1(\varepsilon_y) = 0.4$, $\rho(0) = 0.5$, $q_1 = 0.50$, $q_2 = 0.35$, and n = 50 years, the computed chi-square is 3.24 with three degrees of freedom, which is smaller than the critical value of 7.815. For the case of a bivariate process, serially and mutually dependent, with $\rho_1(\varepsilon_x) = \rho_1(\varepsilon_y) = 0.4$, $\rho(0) = 0.7$, $q_1 = 0.35$, $q_2 = 0.20$, and n = 25 years, the computed chi-square is 159, which is the largest obtained. It can be observed that even for large values of chi-square, the fit looks good, at least for the upper extreme

which is of interest in most applications. It should be stressed that Fig. 4-12 corresponds to the case of the Pearson Type IV distribution adjusted to the distribution of the largest negative-positive run-sum, which gave the smallest value of $F(P(\chi^2)]$ for $P(\chi^2) = 0.95$ among the other drought parameters studied. The case presented in Fig. 4-12 is the one with largest computed chi-square (159). The other cases will have much better fits.

The Pearson Type I distribution function, with origin at the mean, is

$$y = y_e \left(1 + \frac{x}{A_1}\right)^{m_1} \left(1 - \frac{x}{A_2}\right)^{m_2} , \qquad (4\text{-}29)$$

with

$$\frac{m_1+1}{A_1} = \frac{m_2+1}{A_2} .$$

Parameters A_1, A_2, m_1, and m_2 are estimated by the method of moments. Similarly, the parameters β_1 and β_2 are defined by the population moments and estimated by the second, third, and fourth sample moments about the mean. Then the distribution parameters are estimated by

$$r = \frac{6(\beta_2-\beta_1-1)}{6+3\beta_1-2\beta_2} , \qquad (4\text{-}30)$$

$$m_1 = \tfrac{1}{2}\{(r-2)+r(r+2)\sqrt{\frac{\beta_1}{\beta_1(r+2)^2+16(r+1)}}\} , \qquad (4\text{-}31)$$

$$m_2 = \tfrac{1}{2}\{(r+2)-r(r+2)\sqrt{\frac{\beta_1}{\beta_1(r+2)^2+16(r+1)}}\} , \qquad (4\text{-}32)$$

$$A_1 = \tfrac{1}{2}\sqrt{\mu_2}\sqrt{\{\beta_1(r+2)^2+16(r+1)\}}\ \frac{m_1+1}{m_1+m_2+2} , \qquad (4\text{-}33)$$

$$A_2 = \tfrac{1}{2}\sqrt{\mu_2}\sqrt{\{\beta_1(r+2)^2+16(r+1)\}} - A_1 , \qquad (4\text{-}34)$$

with the probability density function at the origin

$$y_e = \frac{1}{A_1+A_2}\ \frac{(m_1+1)^{m_1}(m_2+1)^{m_2}}{(m_1+m_2+2)^{m_1+m_2}}\ \frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)} . \qquad (4\text{-}35)$$

When μ_5 is positive, m_2 is the positive root. The results of probabilities of chi-square for distributions of the negative-negative largest run-sum are given in Fig. 4-13. At the 95 percent level, 58.3 percent of the computed chi-squares are smaller than the critical chi-square values. The Pearson Type I distribution was found to be the closest approximation to frequency distributions of the negative-negative largest run-sum in n years of serially and mutually dependent components of a normal bivariate process.

Estimated parameters of this Pearson Type I function are related by a stepwise multiple linear regression, of the type of Eq. 4-12, to the same independent variables as in Section 4-3. Dependent

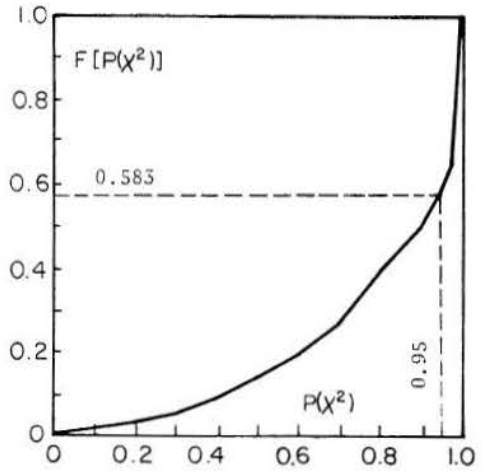

Fig. 4-13 Cumulative Distribution Curve $F[P(\chi^2)]$ of Probabilities $P(\chi^2)$ of Chi-Squares of the Largest Negative-Negative Run-Sum in Using the Pearson Type I Function.

variables were A_1, $\log_{10} A_2$, m_1, and $\log_{10} m_2$. Table 4-14 gives the estimated regression coefficients.

Table 4-14 Estimated Regression Coefficients of Eq. 4-12 for Parameters of Pearson Type I Distribution Function Fitted to Frequency Distributions of the Negative-Negative Largest Run-Sum in Samples of Size n.

Parameter	a	b	c	d	e	f	g	R^2
A_2	-4.29068	2.56351	2.48282	1.17297	6.74935	5.05869	1.02279	.8890
$\log_{10}A_2$	2.37500	.32989	.13936	-.19258	1.43499	.02959	-.67084	.4756
m_1	.92809	-.53316	-.73548	-1.32690	1.76559	-.02879	.29227	.3609
$\log_{10}m_2$	3.19720	-.16236	-.30246	-.67012	1.10093	-.64333	-.80335	.4023

Multiple linear regression equations of parameters β_1 and β_2 are also obtained for the same independent variables, with regression coefficients of the type of Eq. 4-12 given in Table 4-15.

Table 4-15 Estimated Regression Coefficients of Eq. 4-12 for Parameters β_1 and β_2 of Frequency Distribution of the Negative-Negative Largest Run-Sum in Samples of Size n.

Parameter	a	b	c	d	e	f	g	R^2
β_1	3.85483	.37074	.17916	.19429	-.28429	-1.04728	-1.22607	.7397
β_2	9.24014	.40164	.12781	---	.12728	-1.82012	-2.19183	.7125

Figure 4-14 shows a comparison of the experimentally obtained frequency distribution of the negative-negative largest run-sum in samples 50 years long, with the fitted cumulative Pearson Type I distribution function, for the bivariate process with serially and mutually dependent components, and with $\rho_1(\varepsilon_x) = 0.4$, $\rho_1(\varepsilon_y) = 0.2$, $\rho(0) = 0.5$. $q_1 = 0.2$, $q_2 = 0.35$, and $n = 50$. The computed chi-square is 6.72 with five degrees of freedom, which is smaller than the critical chi-square of 11.07. Figure 4-15 shows a similar comparison in case the largest computed

chi-square is 431.0, for the bivariate process, with serially and mutually dependent components, and $\rho_1(\varepsilon_x) = \rho_1(\varepsilon_y) = 0.4$, $\rho(0) = 0.7$, $q_1 = q_2 = 0.2$, and $n = 25$.

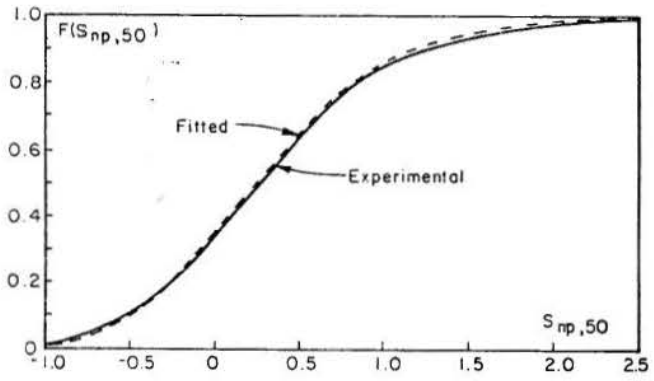

Fig. 4-14 Comparison of the Cumulative Frequency Distribution of the Negative-Negative Largest Run-Sum in a Sample of 50, and the Fitted Cumulative Pearson Type I Distribution Function, with Parameters $A_1 = 2.64655$, $A_2 = 21.28228$, $m_1 = 0.48980$, and $m_2 = 10.9802$.

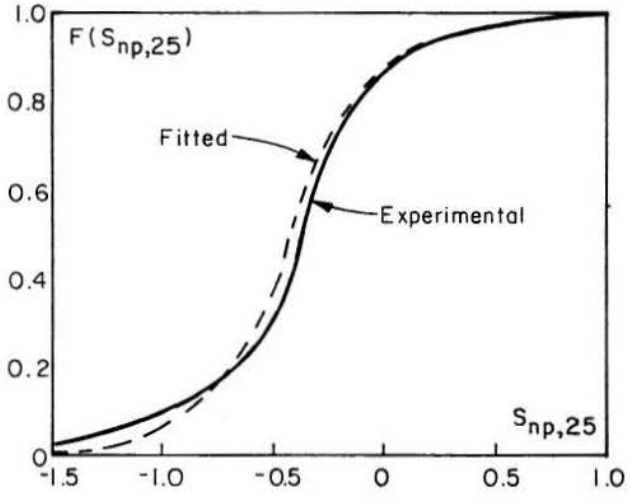

Fig. 4-15 Comparison of the Cumulative Frequency Distribution of the Negative-Negative Largest Run-Sum in a Sample of 25, and the Fitted Cumulative Pearson Type I Distribution Function, with Parameters $A_1 = 2.0510$, $A_2 = 43.5548$, $m_1 = 0.0329$, and $m_2 = 20.9340$.

Chapter V
DROUGHT ANALYSIS OF PERIODIC-STOCHASTIC PROCESSES

5-1 Statement of the Problem

The drought analysis of periodic-stochastic processes is more complex than the drought analysis of stationary stochastic processes. The use of the theory of runs for periodic-stochastic processes may no longer be the best approach unless some decomposition of series is performed for these processes. However, any such transformation may affect the objectives and results of the analysis.

Monthly series will exemplify the periodic-stochastic processes in the following analysis. Series with a shorter time interval than a month may also serve the same purpose. A review of presently available techniques, and of some potential techniques for drought analysis of these processes, are presented in this chapter only. A case study is given in which the drought parameters used are specific drought magnitude criteria.

5-2 A Review of Presently Available Techniques

Drought analysis depends whether the water flow is regulated or not. The instantaneous extremes are used in case of no regulation, while deficits during the critical periods are used when flow regulation is involved.

In the theory of extremes, droughts are defined as instantaneous or interval smallest annual values, with every year giving one lowest value or the drought (Gumbel, 1963). The problem was to find probabilities of these lowest values, called the minimum drought values, either positive or zero. Using the symbol X for random variables defining droughts, the return period $T(X)$, as the expected number of years between the exceedences of a deficit, is

$$T(X) = \frac{1}{P[X \leq x]} = \frac{1}{F(X)}. \qquad (5\text{-}1)$$

Since the exact distribution of drought variables as defined is not available, the asymptotic, bounded exponential distribution of the smallest value of a positive variable was used by Gumbel. This approach to droughts of periodic-stochastic processes may be well used in pollution control problems, attaching probabilities to levels of critical concentrations of pollutants. Similarly, probabilities of minimum consecutive n-days values, with n often 7, 15 and 30 days (Gannon, 1963), are determined. This particular definition of drought as a couple-of-days lowest values in a year may be acceptable for perennial but not for intermittent streams.

For studying droughts in case of flow regulations, Askew et al. (1969) defined the critical period as the time duration during which the hydrologic record would give the most critical deficit with respect to demand. The maximum permissible water extraction rate is used as a variable of this critical period. This permissible rate is based on the active storage available in a hypothetical system of reservoirs. The demand can be smaller, equal or greater than the maximum permissible extraction rate. Generally, the extraction is assumed to be constant during the critical period. Whenever the rate of demand is greater than the maximum permissible extraction rate, the deficit may be conceived as a drought.

Another parameter used for definition of droughts in case of flow regulation is the firm yield criterion (Beard, 1963). The number of shortage periods per year, and the amount of annual firm yield, are defined as drought parameters. Firm yield should be well defined for a reservoir system, with the characteristics of this system specified how it produces the firm yield in terms of monthly and total annual use of water. A single index of the economic effect of shortages was suggested by Beard (1963), in form of the sum of squares of annual shortages in a 100-year period, beginning with an initial or representative amount of water in all storage capacities. The yield needed to be met by the system is the total water requirements of all water users and all losses.

Beard and Kubik (1972), in studying the operation rules of a reservoir system, stated that many theoretical studies of potential yield are based on providing a uniform yield throughout the year, whereas virtually all water uses vary seasonally. As a consequence of it, and in order to consider a more realistic situation, they suggested a detailed sequential analysis of the process of runoff storage use, both for making a reliable estimate of required storage and for deriving operation rules of the system.

The water supply in form of runoff time series have been studied extensively. Their description by mathematical models of periodic-stochastic nature of monthly, weekly or daily series has been extensively investigated. The water use time series have not been studied in such details as the water supply time series. Salas and Yevjevich (1972), in studying the actual water demand or water use time series, concluded that the demand series are basically trend-periodic-stochastic series are to be considered. A need exists for a development of methodology of estimating these parameters and producing the realistic realizations of future samples of water use time series. The lack of these sequences is a likely reason for considering only trend and periodic components in water demand time series. Only the periodic water demand series are used in this paper.

5-3 Potential Techniques for Drought Analysis of Periodic-Stochastic Processes

One alternative in treating the drought of trend-periodic-stochastic series is to remove trends and periodicities in parameters, using either the parametric or nonparametric method of their removal. The procedure in this approach is relatively simple, namely it is assumed that water demand series have both trends and periodicities in basic parameters, with these periodicities being in phase with periodicities in parameters of water supply series. An additional simplification is that they all have the same amplitudes. Llamas and Siddiqui (1969) used this approach for the analysis of a univariate monthly precipitation periodic-stochastic series. The nonparametric method of removing periodicities in parameters was used, and the theory of runs was applied in the drought analysis in case of a dependent stationary time series. It can be shown that the stochastic component of monthly precipitation could be approximated by an independent series for all practical purposes. This fact simplifies the study of droughts for the stochastic component of monthly precipitation in a univariate case. For the bivariate case and removed

periodicities in parameters, in this approach the exact expressions for distributions of different runs can be used, as shown in Chapter II. However, the run-sums may not have a clear meaning if the general but different standard deviations of the two series are not retained while removing periodicities. Run-lengths can be investigated on the standardized stochastic components without too many problems. For dependent second-order stationary univariate or bivariate series either the exact or approximate expressions of the theory of runs, as presented in Chapter II, will produce the properties of droughts. The analysis of droughts for trend-periodic-stochastic processes, by removing trends and periodicities in parameters, depends on the characteristics of demand series.

Another alternative is to use the supply-less-demand series. Since the supply series is periodic-stochastic and the demand series is assumed to be only periodic, in phase with and of the same amplitude as the periodicity in supply series, differences between supply and demand represent a first-order stationary process of deficits and excesses. In case of high variability between the low and high flows, the excess-deficit series still can be periodic, in which case the theory of runs of stationary processes may not be meaningfully applied. This approach has the disadvantage of not being adequate when periodicities in demand are out of phase with and of different amplitude than the periodicity in supply.

The third alternative, used in this paper, is the "drought-magnitude and drought-duration criteria." The magnitude of a drought depends on the demand imposed on the water system. During the planning stage of a water resource development scheme, for example, the choice of droughts for analysis is related to contemplated demand series. As shown by Texas Water Development Board (1971), the severity of the most critical drought affects the selection of the ultimate plan, by influencing decisions on the size and the number of facilities required for optimal performance of a system. The more severe this most critical drought, the larger or more numerous are facilities that are needed to insure an adequate performance of the entire system. Of great interest in the planning process are droughts which require new storage capacity to insure uninterrupted deliveries, or which require importation of water from other sources.

When severity of a drought is studied, special considerations must be given to relations between the drought duration and all the physical storage and other capacities of the system, which are required to meet the demand during the drought period. A drought of a given duration, equal to or longer than the time required to use the storage system from a full to an empty state, will have quite a different effect on the system than a short drought not requiring more than the total water storage.

The magnitude of a drought can be defined as the maximum absolute value of monthly differences between supply and demand over the drought duration. In mathematical terms, this magnitude is

$$M_t = \min_t \left[\min_k \sum_{i=k+1}^{k+t} \frac{X_i - D_i}{t} \right], \qquad (5\text{-}2)$$

with X_i the monthly supply (in the case of a system of reservoirs, it is the sum of the monthly inflows to all the reservoirs), D_i the monthly demand (in case of complex systems, it is the sum of monthly demands at all system demand points), k any starting time point for studying droughts, and t the duration of the critical drought period in samples used for analysis. This concept is analogous to studying the negative run-intensity for univariates or the joint negative-negative run-intensity for bivariates.

Another parameter, proposed by the Texas Water Development Board (1971), is the drought time position, defined by the time of the drought mid-point. For a drought with duration t and the absolute starting time k, the position is

$$L_t = k + \frac{t}{2} \qquad (5\text{-}3)$$

There may be individual months during drought periods when supply exceeds demand. However, effects of these months may not be sufficiently large to overcome the general drought consequences, since all the other months would have significant deficits. Because some sort of flow regulation may always be involved, the storage easily takes care of individual months with small surplus and distributes it over the months of significant deficit. If no regulation is involved, the surplus of these couple of months is simply lost.

This alternative for drought measuring parameters has the basic disadvantage that the theory of runs cannot be easily applied, since the periodicities are involved. It is a somewhat different approach to drought definition. The main advantage over the other approaches to drought definitions is that it can easily treat the cases of demand being out of phase and of different amplitude in comparison with those of water supply.

A fourth alternative is based on a simultaneous generation of annual and monthly series, by jointly preserving their parameters such as the mean, variance, serial correlation coefficients, among the others. Harms and Campbell (1967) used this type of generation, claiming to preserve the normal distribution of annual flows, the lognormal distribution of monthly flows, and the serial correlation of annual and monthly flows. The technique is based on the assumption that a first-order linear autoregressive model is adequate to represent the dependence of annual flow series, and that the Thomas-Fiering model is adequate to represent the structure of monthly flow series, with an adjustment being sufficient to take care of the linkage between the annual and monthly flow series. Their expressions are

$$\frac{Q_{j+1} - \overline{Q}'}{S'} = R \frac{Q_j - \overline{Q}'}{S'} + t_i (1 - R^2)^{1/2}, \qquad (5\text{-}4)$$

and

$$\frac{\log q_{i+1,j} - \overline{\log q'_{i+1}}}{S'_{i+1}} = r_{i+1} \frac{\log q_{i,j} - \overline{\log q'_i}}{S'_i} + t_i (1 - r_{i+1}^2)^{1/2}, \qquad (5\text{-}5)$$

with the adjustment

$$q_{i,j} = \frac{365\, q_{i,j} Q_j}{\sum a_i q_{i,j}}, \qquad (5\text{-}6)$$

where a_i is the number of days in the j-th month, $q_{i,j}$ and Q_j are the generated monthly and annual

flows, respectively, $q'_{i,j}$ and Q'_j are the historical monthly and annual flows respectively, R_i and r_i are the first serial correlation coefficients of annual and monthly series, respectively. Another technique available for this type of generation is the disaggregation process, outlined by Valencia and Schaake (1973). For the cases considered in this paper, namely the first-order linear autoregressive model for both the annual and monthly series, the technique which considers a sequential generation of annual events with a disaggregation model for generating seasonal, monthly, weekly, or daily events within the year, can be adjusted and used. Due to computer storage requirements, these authors suggest first to generate seasonal values and then to repeat the process on a season by season basis to generate monthly values in a second disaggregation step.

For this fourth alternative of simultaneous generation of annual and monthly time series, once samples are generated, the theoretical analysis or approximations in case of dependent processes can be applied to annual series to determine the probabilities of drought runs. For example, if the annual process is inferred to be stationary process having the first serial correlation coefficient ρ_1, then the probabilities of a long drought or probabilities of the longest run-length, say for a project of economic life of 50 years, can be determined. For simultaneous generation, a k-year or the longest drought in annual series may be singled out, and the monthly series of this period can be investigated. The annual series permit the identification of critical drought periods to design the system, with the sequential patterns of monthly series studied for these periods. The main advantage of this alternative is the use of a more reliable estimation procedure for probabilities of droughts rather than obtaining these probabilities from less reliable frequencies of historical records.

Further advantage of the fourth approach relates to the use of optimization techniques in design and operation of water resources systems, because, after the critical droughts of given probabilities are determined, the optimization procedures can be applied to parts of monthly series during these critical periods instead of optimization extended throughout the total generated monthly series.

The approach of drought magnitude and drought duration criteria, as outlined herein, has the potential to be developed in a technique of drought analysis of periodic-stochastic processes. To demonstrate this potential, a case study has been worked out and presented in the next section.

5-4 A Case Study

The drought analysis of periodic-stochastic processes is complex not only due to periodicities, but also because the number of parameters for both the supply and demand series is much greater than in the case of stationary stochastic processes. The large number of parameters requires a large number of cases to be studied in the general form, and this number can be excessive. As a consequence, no attempt is made here to generalize all cases or to cover some or most of them in this paper. A case study is given only in order to show the use of drought parameters, as presented in Section 5-3 and therefore, the case study covers a small number of parameters. In spite of simplifications it is thought that the case has a practical significance.

The monthly supply series is assumed to be a periodic-stochastic process, with periodic mean and periodic standard deviation composed only of the 12-month harmonic. The resulting stationary stochastic component follows the first-order linear autoregressive model, as given by

$$X_i = \bar{x}+C_1(\mu)\cos[\tfrac{\pi}{6}\tau+\theta_1(\mu)] + \{\bar{s}_\tau+C_1(\sigma)\cos[\tfrac{\pi}{6}\tau+\theta_1(\sigma)]\}(\rho_1\varepsilon_{i-1}+\xi_i), \quad (5\text{-}7)$$

with $\bar{x}$ the overall mean (2.885), $C_1(\mu)$ the amplitude of the 12-month harmonic in the mean (1.889), $\theta_1(\mu)$ the phase of the first harmonic in the means (0), $\bar{s}_\tau$ the overall mean standard deviation (1.848), $C_1(\sigma)$ the amplitude of the first harmonic in the standard deviation (0.946), $\theta_1(\sigma)$ the phase of the first harmonic in the standard deviation (0), and ρ_1 the first serial correlation coefficient (0.5). The independent stochastic component ξ_i is assumed to follow the three-parameter lognormal distribution

$$f(x) = \frac{1}{(x-\beta)s_n\sqrt{2\pi}}\exp\{-\frac{(\ln(x-\beta)-\mu_n)^2}{2s_n^2}\}, \quad (5\text{-}8)$$

with the lower bound $\beta = -1.5$, $\mu_n = 0.2216$, and $s_n^2 = 0.3677$.

The monthly demand series is assumed to be a periodic process, with periodicity composed only of a 12-month harmonic. This is a simplification in comparison to reality. Nevertheless, it is a common practice in water resources planning to simplify the complex nature, because quite often the lack of data may not justify the more complex models. The demand model then is

$$D_i = \bar{D} + C_1^*\cos(\tfrac{\pi}{6}\tau + \theta_1^*), \quad (5\text{-}9)$$

with $\bar{D}$ the overall mean (2.50), C_1^* the amplitude of the first harmonic (1.00), and θ_1^* the phase angle of the first harmonic (0). For supply and demand series in phase, $\theta_1(\mu) = \theta_1^*$; otherwise they are different. In this case study three alternatives are used for phase differences $[\theta_1(\mu) - \theta_1^*] = 0.0$, $\pi/2$ and π.

Since the monthly demand has been shown by Salas and Yevjevich (1971) to be periodic-stochastic processes, the demand could have been modeled as

$$D_i = \bar{D} + C_1^*(\mu)\cos[\tfrac{\pi}{6}\tau + \theta_1^*(\mu)] + \{\bar{S}_D + C_1^*(\sigma)\cos[\tfrac{\pi}{6}\tau + \theta_1^*(\sigma)]\}(\rho_1^*\varepsilon_{i-1}+\xi_i^*) \quad (5\text{-}10)$$

with parameters $\bar{D}$, $C_1^*(\mu)$, $\theta_1^*(\mu)$, $\bar{S}_D$, $C_1^*(\sigma)$, $\theta_1^*(\sigma)$, ρ_1^*, and the independent stochastic component of demand series ξ_i^* as the counterpart of that of the supply series. Since results are expected to be similar to those when only the periodic demand is

considered, as far as computing the drought characteristics, the case study treats only the periodic demand.

A program was prepared to compute drought characteristics as defined in Section 5-3, namely the drought magnitude, Eq. 5-2, the drought durations for its given magnitude and the corresponding deficit. Figures 5-1, 5-2, and 5-3 shown supply and demand series, for the three cases of phase differences. Figures 5-4, 5-5, and 5-6 show the drought magnitude computed for a set of samples each of 30 years of monthly flows, as well as the cumulative deficit during the drought of a given duration for the three phase differences between the supply and the demand.

Values of drought characteristics as shown in Figs. 5-4, 5-5, and 5-6, are presented in Table 5-1. Figures 5-4, 5-5, and 5-6, give an idea of the range of values of the drought magnitude criteria and its duration. The selected value corresponds to the maximum deficit. It should be noted that generally the criteria decrease with an increase in duration. The analysis of the values obtained show that phase differences (up to half a cycle) do not influence very much the drought duration or the volume of deficit for the case study. However, it should be recognized that these results apply only to the selected values of the case study and no generalization could be made.

Table 5-1 Values for Drought Characteristics of the Case Study with Three Phase Differences between Supply and Demand Series.

Phase difference	Drought magnitude	Duration in months	Volume deficit
0	.6376	43	27.4169
$\pi/2$	.6969	42	29.2717
π	.6734	42	28.2812

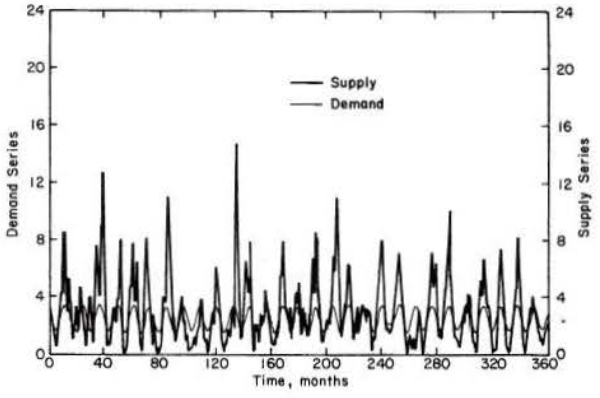

Fig. 5-1 Supply Series (Periodic-Stochastic) and Demand Series (Periodic) for the Case Study with No Phase Difference $[\theta_1(\mu)-\theta_1^* = 0.0]$.

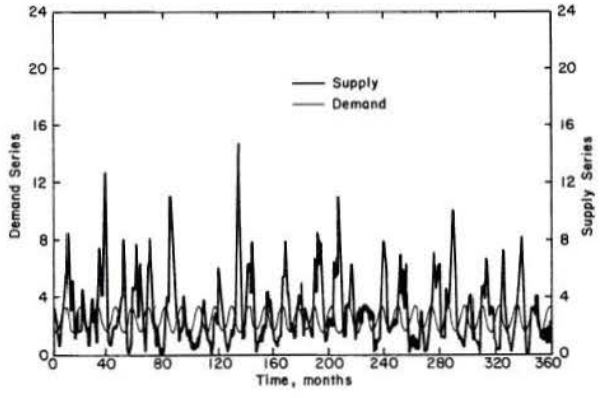

Fig. 5-2 Supply Series (Periodic-Stochastic) and Demand Series (Periodic) for the Case Study with Phase Difference of $[\theta_1(\mu) - \theta_1^*] = \pi/2$.

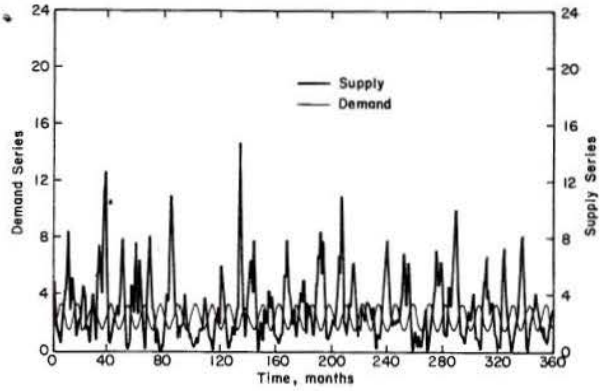

Fig. 5-3 Supply Series (Periodic-Stochastic) and Demand Series (Periodic) for the Case Study with Phase Difference of π, $[\theta_1(\mu) - \theta_1^* = \pi]$.

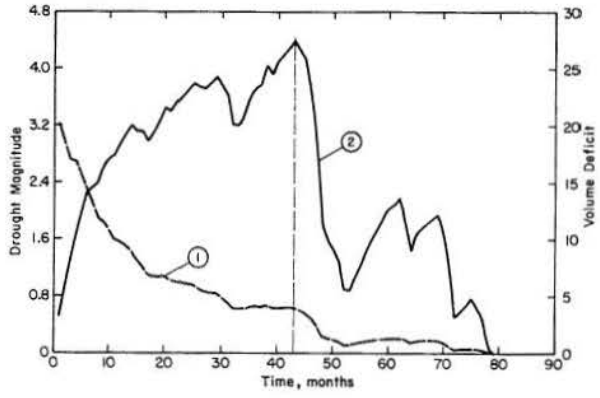

Fig. 5-4 Drought Magnitude (1) and Corresponding Volume Deficit (2) for Given Drought Duration, Which Correspond to the Case Study of No Phase Difference.

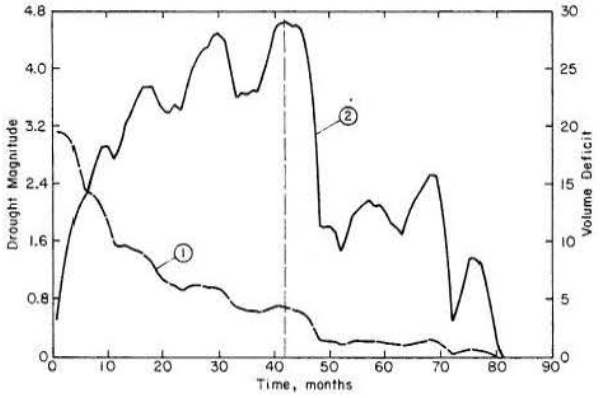

Fig. 5-5 Drought Magnitude (1) and Corresponding Volume Deficit (2) for Given Drought Duration, Which Correspond to the Case Study of Phase Difference Equal to $\pi/2$.

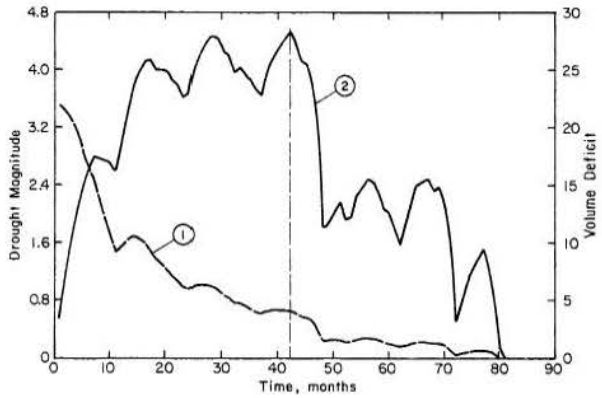

Fig. 5-6 Drought Magnitude (1) and Corresponding Volume Deficit (2) for Given Drought Duration, Which Correspond to the Case Study of Phase Difference Equal to π.

Chapter VI
CONCLUSIONS

The main contributions and conclusions of investigations in this paper are;

(1) For a unidimensional case, an exact expression for the longest run-length of a given kind for a Markov chain has been developed, based on Bateman's work as presented by Eq. 2-37. The analysis provides an adequate approximation to distributions of the longest run-length of unidimensional dependent series following the first-order linear autoregressive model for values of the first serial correlation coefficient not greater than $\rho_1 = 0.4$.

(2) For a bivariate case, with the two components serially and mutually dependent, serially dependent but mutually independent, and serially independent but mutually dependent, transformation to the univariate case is accomplished by defining the new random variables. To obtain the four-state Markov chain approximation to the serially and mutually dependent components of a bivariate series, of the first-order linear autoregressive dependence of each component, the quadrivariate normal distribution and its integration is performed by using the tetrachoric series expansion. The approximation seems to be satisfactory for values of the first serial and cross correlation coefficients up to 0.4. For better results, more terms should be included in the series expansion. Since lumpability requirements are too restrictive, the use of transformations of bivariate to univariate cases and of the Markov chains as approximations to autoregressive models for the new univariate variables, gives good approximations.

(3) For the distribution of the longest run-length of a given kind in a sample of size n for serially and mutually independent, and serially independent but mutually dependent components of bivariate processes, an exact expression is developed, based on the analysis of the four possible outcomes and the work by David and Barton.

(4) The experimental Monte Carlo method was used to find the distribution of the run-sums in bivariate processes with serially and mutually dependent components.

(5) Frequency distributions of selected runs for the study of serially and mutually dependent components of bivariate processes were obtained by using a bivariate linear autoregressive model. Results are presented in the form of estimated parameters of fitted probability distribution functions to experimentally obtained frequency distributions. Discrete probability distribution functions are fitted to frequency distributions of run-length, and continuous distribution functions for the run-sums.

(6) For the run-lengths of infinite populations, the negative-binomial distribution is found adequate to approximate the frequency distributions of negative-negative and negative-positive run-lengths of infinite series. In finite samples of data a mixture of two geometric distributions is found adequate to approximate the frequency distributions of negative-negative and negative-positive longest run-length.

(7) Distribution functions of gamma type, transformed by Laguerre polynomials, are used to approximate joint distributions of negative-negative or negative-positive run-sums and run-intensities, respectively, for infinite series. Expressions for coefficients of Laguerre polynomials are obtained and coefficients in multiple regression equations determined for parameters of joint distributions. For the negative-negative largest run-sum in samples of size n, the Pearson Type I distribution function is selected as an approximation, and the Pearson Type IV distribution function is selected as an approximation for distributions of the negative-positive largest run-sum in samples of size n, with parameters of the Pearson Type I and IV distribution functions. Multiple linear regression equations are determined to express the estimated parameters of fitted probability distribution functions in terms of parameters of the underlying bivariate processes and the two truncation levels in conclusions (5), (6) and (7).

(8) Explained variances by the multiple regression equations for the parameters of distributions fitted to frequency distributions of run variables of infinite series are much higher, on the average, than the corresponding explained variances for run variables in case of samples of a given size.

(9) The present theory of runs is not adequate to treat the periodic-stochastic processes. At the present stage an alternative type of drought analysis has to be used. A case study based on a particular monthly data series and drought parameters shows that the parameters are not affected significantly by the differences in phases up to a half cycle between supply and demand.

REFERENCES

1. Abramowitz, M. and Stegun, I., 1965, Handbook of mathematical functions, Dover Pub., New York.
2. Askew, A.J., Yeh, W.W.G., and Hall, W.A., 1970, Streamflow generating techniques: a comparison of their abilities to simulate critical periods of drought, University of California, Water Resources Center, Contribution 131, January, 1970.
3. Bardwell, G.E. and Crow, E.L., 1964, A two parameter family of hyper-Poisson distributions, Journal of the American Statistical Association, 59, pp. 133-141.
4. Bateman, G., 1948, On the power function of the longest run-length as a test for randomness in a sequence of alternatives, Biometrika, 35, pp. 97-112.
5. Baticle, M.E., 1946, Le problème des stocks, Comptes Rendus, Academie des Sciences, Paris, 222, Pt 1, Jan-Mar, pp. 355-357.
6. Beard, L.R., 1963, Estimating long term storage requirements and firm yield of rivers, IASH Berkeley Symposium, August, 1963, Pub. 63, pp. 151-166.
7. Beard, L.R. and Kubik, H.E., 1972, Drought severity and water supply dependability, Journal of the Irrigation and Drainage Division, IR-3, September, 1972, pp. 433-442.
8. Box, G.E.P. and Muller, M.E., 1958, A note on the generation of normal deviates, Annals of Math. Stat., Vol. 28, pp. 610-611.
9. Burr, E.J. and Cane, G., 1961, Longest run of consecutive observations having a specified attribute, Biometrika, 48, pp. 461-465.
10. Cramer, H., 1946, Mathematical methods of statistics, Princeton University Press, Princeton.
11. Cox, D.R. and Miller, H.R., 1968, The theory of stochastic processes, New York, J. Wiley and Sons, Inc., 398 p.
12. David, F.N. and Barton, D.E., 1962, Combinatorial chance, Griffin Pub. Co., London.
13. Downer, R.N., Siddiqqi, M.M., and Yevjevich, V., 1967, Applications of runs to hydrologic droughts, Proceedings of the International Hydrology Symposium, Fort Collins, Colorado.
14. Elderton, Sir W.P., 1953, Frequency curves and correlation, Harren Press, Washington, D.C., Fourth ed., 272 p.
15. Elderton, Sir W.P. and Johnson, N.L., 1969, Systems of frequency curves, London: Cambridge University Press.
16. Feller, W., 1957, An introduction to probability theory and its applications, Vol. 1, Wiley and Sons, New York.
17. Fiering, M.B., 1964, Multivariate technique for synthetic hydrology, Proceedings of the American Society of Civil Engineers, Vol. 90, HY5, September, pp. 43-60.
18. Gabriel, K.R. and Neumann, J., 1957, On a distribution of weather cycles by length. Quarterly Journal of the Roy. Met. Society 83, p. 375.
19. Gabriel, K.R., 1959, The distribution of the number of successes in a sequence of dependent trials, Biometrika, 46, pp. 454-460.
20. Gannon, J.J., 1963, Definition of river drought flow characteristics, IASH Berkeley Symposium, August, 1963, Pub. 63, pp. 137-150.
21. Gumbel, E.J., 1963, Statistical forecast of droughts, Bulletin IASH, Vol. VIII, No. 1, pp. 5-23.
22. Harms, A.A. and Campbell, T.H., 1967, An extension of the Thomas Fiering model for the sequential generation of streamflows, Water Resources Research, Vol. 3.
23. Heiny, R., 1968, Stochastic variables of surplus and deficit, Unpublished Ph.D. Dissertation, Colorado State University.
24. Johnson, N.L. and Kotz, S., 1969a, Distributions in statistics: discrete distributions, Houghton Mifflin Company, Boston.
25. Johnson, N.L. and Kotz, S., 1969b, Distributions in statistics: continuous distributions, Vol. I and II, Houghton Mifflin Company, Boston.
26. Kendall, M.G., 1941, Proof of relations connected with tetrachoric series and its generalization, Biometrika, 32, pp. 196-198.
27. Kendall, M.G. and Stuart, A., 1969, The advanced theory of statistics, Vol. I, II and III, Hafner Publishing Company, New York.
28. Kemeny, J.G. and Snell, J.L., 1960, Finite Markov chains, D. Van Nostrand Co. Inc.
29. Llamas, J., 1968, Deficit and surplus in precipitation series, Ph.D. Dissertation, Colorado State University.
30. Llamas, J. and Siddiqqi, M.M., 1969, Runs of precipitation series, Hydrology Paper No. 33, Colorado State University.
31. Mardia, K.V., 1970, Families of bivariate distributions, London: Griffin.
32. Matalas, N.C., 1967, Mathematical assessment of synthetic hydrology, Water Resources Research, Vol. 3, No. 4, pp. 937-945.
33. McGinnis, D.F. and Sammons, W.H., 1970, Discussion on, Daily streamflow simulation, by K. Payne, W.R. Neumann and K.D. Perry, Proc. Amer. Soc. Civil Eng., 96(HY5), 1201-1206, 1970.
34. Millan, J. and Yevjevich, V., 1971, Probabilities of observed droughts, Hydrology Paper No. 50, Colorado State University.
35. Millan, J., 1972, Drought impact on regional economy, Hydrology Paper No. 55, Colorado State University.
36. Moivre, A. De, 1738, The doctrine of chances, or a method of calculating the probabilities of events in play, New impression of second edition, 1967, Frank Cass, London.
37. Mood, A.M., 1940, The distribution theory of runs, Annals of Mathematical Statistics, 11, pp. 367-392.
38. Moreau, D.H., 1971, The synthesis of flows in North Carolina streams for computer simulation experiments, Water Resources Research Institute of the Universities of North Carolina, Report No. 50, April, 1971.

39. Mosteller, F., 1941, Note on an application of runs to quality control charts, Annals of Mathematical Statistics, Vol. 11, pp. 228-232.

40. Ord, J.K., 1967, On a system of discrete distributions, Biometrika, 54, pp. 649-656.

41. Ord, J.K., 1972, Families of frequency distributions, Griffin's statistical monographs and courses, Griffin, London.

42. Patil, G.P., Joshi, S.W., and Rao, C.R., 1968, A dictionary and bibliography of discrete distributions, Hafner Publishing Co., New York.

43. Pegram, G.G.S. and James, W., Multilag multivariate autoregressive model for the generation of operational hydrology, Water Resources Research, Vol. 8, No. 4, August, 1972, pp. 1074-1076.

44. Pearson, K., 1895, Contributions to the mathematical theory of evolution II, Skew variations in homogeneous material, Philosophical Transactions of the Royal Society of London, Ser. A. 186, pp. 343-414.

45. Quenouille, M.H., 1957, The analysis of multiple time series, Griffin statistical monographs and courses, Griffin, London.

46. Roesner, L.A. and Yevjevich, V., 1966, Mathematical models for time series of monthly precipitation and monthly runoff, Hydrology Paper No. 15, Colorado State University.

47. Salas-La Cruz, J.D. and Yevjevich, V., 1972, Stochastic structure of water use time series, Hydrology Paper No. 52, Colorado State University, June, 1972.

48. Saldarriaga, J., 1969, Investigation of wet and dry years by runs, Ph.D. Dissertation, Colorado State University.

49. Saldarriaga, J. and Yevjevich, V., 1970, Application of run lengths to hydrologic series, Hydrology Paper No. 40, Colorado State University.

50. Siddiqqi, M.M., 1960, Test for regression coefficients when errors are correlated, Annals of Mathematical Statistics, Vol. 31, pp. 931-932.

51. Subrahmanyam, V.P., 1967, Incidence and spread of continental drought, Report No. 2, Reports on WMO/IHD Projects, W.M.O.

52. Texas Water Development Board, 1971, Stochastic optimization and simulations techniques for management of regional water resource systems, Report 131, July, 1971, 129 p.

53. Thornthwaite, C.W., 1948, An approach toward a rational classification of climate, Geogr. Review, Vol. 38, No. 1, pp. 55-94.

54. Uspensky, J.V., 1937, Introduction to mathematical probability, McGraw Hill, New York.

55. Valencia, D.R. and Schaake, J.C. Jr., 1973, Disaggregation processes in stochastic hydrology, Water Resources Research, Vol. 9, No. 3, June, 1973, pp. 580-585.

56. Whitworth, W.A., 1896, Choice and chance, Deighton Beel and Co., New impression by Hafner, Pub. Company.

57. Wolfowitz, J., 1943, On the theory of runs with some applications to quality control, Annals of Math. Stat., Vol. 14, 1943, pp. 280-288.

58. Yevjevich, V., 1964, Fluctuations of wet and dry years, part II, analysis by serial correlation, Hydrology Paper No. 4, Colorado State University.

59. Yevjevich, V., 1967, An objective approach to definitions and investigations of continental hydrologic droughts, Hydrology Paper No. 23, Colorado State University.

60. Yevjevich, V., 1972a, Probability and statistics in hydrology, Water Resources Publications, Fort Collins, Colorado.

61. Yevjevich, V., 1972b, Stochastic processes in hydrology, Water Resources Publications, Fort Collins, Colorado.

62. Young, G.K. and Pisano, W.C., 1967, Discussion of mathematical assessment of synthetic hydrology by N.C. Matalas, Water Resources Research, Vol. 4, No. 3, pp. 681-682.

63. Young, G.K. and Pisano, W.C., 1968, Operational hydrology using residuals, Journal of Hydraulics Division, ASCE, HY4, 94, pp. 909-923.

Key Words: Droughts, Water Deficits, Water Shortages, Runs, Theory of Runs, Bivariate Runs.

Abstract: Methodologies for analysis of droughts are presented for stationary and periodic-stochastic processes. Droughts are studied by means of the theory of runs. Distributions of the longest run-length and the largest run-sum in a series of a given length, and distributions of the run-length and the run-sum of infinite series for various cases of univariate and bivariate series are investigated. Exact, approximate or experimentally obtained expressions are presented for univariate and bivariate independent and dependent series. For bivariate series combinations of serially independent and dependent, and mutually independent and dependent series are studied. When exact analytical solutions could not be obtained, the data generation method is used. Frequency distributions of various drought characteristics associated with the

Key Words: Droughts, Water Deficits, Water Shortages, Runs, Theory of Runs, Bivariate Runs.

Abstract: Methodologies for analysis of droughts are presented for stationary and periodic-stochastic processes. Droughts are studied by means of the theory of runs. Distributions of the longest run-length and the largest run-sum in a series of a given length, and distributions of the run-length and the run-sum of infinite series for various cases of univariate and bivariate series are investigated. Exact, approximate or experimentally obtained expressions are presented for univariate and bivariate independent and dependent series. For bivariate series combinations of serially independent and dependent, and mutually independent and dependent series are studied. When exact analytical solutions could not be obtained, the data generation method is used. Frequency distributions of various drought characteristics associated with the

Key Words: Droughts, Water Deficits, Water Shortages, Runs, Theory of Runs, Bivariate Runs.

Abstract: Methodologies for analysis of droughts are presented for stationary and periodic-stochastic processes. Droughts are studied by means of the theory of runs. Distributions of the longest run-length and the largest run-sum in a series of a given length, and distributions of the run-length and the run-sum of infinite series for various cases of univariate and bivariate series are investigated. Exact, approximate or experimentally obtained expressions are presented for univariate and bivariate independent and dependent series. For bivariate series combinations of serially independent and dependent, and mutually independent and dependent series are studied. When exact analytical solutions could not be obtained, the data generation method is used. Frequency distributions of various drought characteristics associated with the

Key Words: Droughts, Water Deficits, Water Shortages, Runs, Theory of Runs, Bivariate Runs.

Abstract: Methodologies for analysis of droughts are presented for stationary and periodic-stochastic processes. Droughts are studied by means of the theory of runs. Distributions of the longest run-length and the largest run-sum in a series of a given length, and distributions of the run-length and the run-sum of infinite series for various cases of univariate and bivariate series are investigated. Exact, approximate or experimentally obtained expressions are presented for univariate and bivariate independent and dependent series. For bivariate series combinations of serially independent and dependent, and mutually independent and dependent series are studied. When exact analytical solutions could not be obtained, the data generation method is used. Frequency distributions of various drought characteristics associated with the

runs, obtianed by the generation method for the bivariate case, are fitted by discrete or continuous probability distribution functions. Multiple regression analysis is used to obtain useful relationships between the parameters of fitted distribution functions and the parameters of time series dependence, cross dependence and the truncation levels. Periodic-stochastic series are studied by defining drought and its parameters for this particular type of hydrologic processes.

Reference: Guerrero-Salazar, Pedro, and Yevjevich, Vijica, Colorado State University, Hydrology Paper No. 80, (September 1975), Analysis of Drought Characteristics by the Theory of Runs.

runs, obtianed by the generation method for the bivariate case, are fitted by discrete or continuous probability distribution functions. Multiple regression analysis is used to obtain useful relationships between the parameters of fitted distribution functions and the parameters of time series dependence, cross dependence and the truncation levels. Periodic-stochastic series are studied by defining drought and its parameters for this particular type of hydrologic processes.

Reference: Guerrero-Salazar, Pedro, and Yevjevich, Vijica, Colorado State University, Hydrology Paper No. 80, (September 1975), Analysis of Drought Characteristics by the Theory of Runs.

runs, obtianed by the generation method for the bivariate case, are fitted by discrete or continuous probability distribution functions. Multiple regression analysis is used to obtain useful relationships between the parameters of fitted distribution functions and the parameters of time series dependence, cross dependence and the truncation levels. Periodic-stochastic series are studied by defining drought and its parameters for this particular type of hydrologic processes.

Reference: Guerrero-Salazar, Pedro, and Yevjevich, Vijica, Colorado State University, Hydrology Paper No. 80, (September 1975), Analysis of Drought Characteristics by the Theory of Runs.

runs, obtianed by the generation method for the bivariate case, are fitted by discrete or continuous probability distribution functions. Multiple regression analysis is used to obtain useful relationships between the parameters of fitted distribution functions and the parameters of time series dependence, cross dependence and the truncation levels. Periodic-stochastic series are studied by defining drought and its parameters for this particular type of hydrologic processes.

Reference: Guerrero-Salazar, Pedro, and Yevjevich, Vijica, Colorado State University, Hydrology Paper No. 80, (September 1975), Analysis of Drought Characteristics by the Theory of Runs.