THE APPLICATION OF SURPLUS, DEFICIT AND RANGE IN HYDROLOGY

By

Vujica M. Yevdjevich

September 1965

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

10

THE APPLICATION OF SURPLUS, DEFICIT AND RANGE IN HYDROLOGY

by

Vujica M. Yevdjevich

HYDROLOGY PAPERS

COLORADO STATE UNIVERSITY

FORT COLLINS, COLORADO

September 1965 No. 10

ACKNOWLEDGMENT

The writer wishes to acknowledge support by the U. S. National Science Foundation for providing the grant which made this study possible, and to the National Center for Atmospheric Research, Boulder, Colorado, for their support in allowing free computer time on the CDC 3600 computer for this study. The writer also acknowledges the assistance of Mrs. Lois Nieman, computer program adviser, in carrying the computational phase of this study. Appreciation is also extended to the graduate and undergraduate students who helped the writer in processing the results supplied by the computer. Appreciation is also extended to Dr. M. M. Siddiqui, Professor of Mathematics and Statistics at Colorado State University, who reviewed parts of this paper. His suggestions were most valuable in the further improvement of the text.

TABLE OF CONTENTS

TABLE OF CONTENTS - continued

LIST OF FIGURES AND TABLES

ABSTRACT

Surplus is defined as the maximum positive sum, deficit as the minimum negative sum, and range as their difference (or the sum of their absolute values) on a curve of cumulative deviations for a given subseries of length n. Several types of surplus, deficit and range are defined depending on the base variable from which the cumulative deviations of a variable x are obtained, especially for the base value $\bar{x}$, and the changing value $\bar{x}_n$ for subseries (adjusted surplus, adjusted deficit and adjusted range). An attempt is made to systematize the types of storage equations. The application of surplus, deficit and range in hydrology is discussed. Storage problems and the use of surplus, deficit and range in analyzing these problems are viewed from the three approaches: empirical method, data generation method and analytical method. Properties of these three methods, as applied to the surplus, deficit and range, are investigated in detail. Smoothness in results of the latter two methods in comparison with the first method should not be mistaken for increased information. The three methods are compared on the bases of the Rhine River's annual flows.

The distributions and the parameters of distributions for the surplus, range, adjusted surplus and adjusted range of independent and dependent normal variables are investigated by: the analytically derived expressions or by exact distributions; and by the data generation method in obtaining samples generated of 100,000 independent and/or dependent normally distributed numbers.

The effect of dependence in time series on distributions of surplus, range, adjusted surplus and adjusted range is studied for the Markov first order linear dependence model of a normal variable, with both the independent and dependent variable having means zero and variances unities. The statistical parameters of distributions of surplus, range, adjusted surplus and adjusted range change significantly with an increase of the dependence parameter of this model.

The effect of skewness of basic variable on the statistical parameters (mean, variance, and skewness coefficient) or surplus, deficit, range, adjusted surplus, adjusted deficit and adjusted range are investigated for independent gamma variables with skewness coefficients ranging from zero to $\sqrt{8}$. The effect of skewness is larger on surplus, deficit, adjusted surplus and adjusted deficit than on the range and adjusted range. The effect increases with an increase of the order of statistical moment used in the computation of these parameters.

THE APPLICATION OF SURPLUS, DEFICIT AND RANGE IN HYDROLOGY

By: Vujica M. Yevdjevich*

CHAPTER I

INTRODUCTION

1. Time series. A sequence of observations on a quantity in time is a time series. If the quantity under observation is symbolized by X, its value at time t is designated by X_t. In the probability theory of time series, each X_t is considered as a stochastic variable. In this case, the time series is also called a stochastic process. If X_t is defined for all t in an interval $a \leq t \leq b$, it is called a continuous-time process. On the other hand, if X_t is defined only at discrete times $t_1, t_2, \ldots$, it is called a discrete-time process. In many practical situations X_t may be a continuous-time process but is observed at equally spaced intervals of time, giving a sequence $X_\Delta, X_{2\Delta}, \ldots$. Or, the average of X_t over a period Δ is calculated giving the sequence of means: $\overline{X}_\Delta, \overline{X}_{2\Delta}, \ldots$, where $\Delta, 2\Delta, 3\Delta, \ldots$, denote the successive equal intervals of time.

A great many hydrologic variables are observed or derived as time series. Properties of these series are of ever-increasing significance in planning, designing and operating water resource projects. Hydrology places emphasis on techniques available for the analysis of time series, and potential techniques which can be developed for general or special problems. This paper deals only with a particular definition of discrete-time series or with those continuous-time series which are made discrete. Discrete-time series will be defined later in this text.

The maximum surplus, the maximum deficit, and the maximum range for a time series from time 0 to time t may be defined in various ways. In this paper these factors will be defined as the maximum value, the minimum value and the difference between the maximum and minimum value on the cumulative curve. This curve reflects the cumulative sums of deviations of a variable from a defined value, from a changing parameter or a function of time, for a given length of a time series. Detailed definitions of maximum surplus, maximum deficit and maximum range are given in Chapter II. Even though the study is limited to the analysis of surplus, deficit and range as they apply to a hydrologic time series, the results and techniques given here apply to fields other than hydrology.

2. Techniques for analysis of time series. Theory of probability, mathematical statistics, stochastic processes and other fields of mathematics are among the many techniques used for the analysis of stationary time series. Of these techniques the most commonly used are: harmonic analysis (based on Fourier series); serial correlation analysis; power spectrum analysis; analysis by surplus, deficit and range; analysis by runs; and others. Describing time-dependent stochastic processes by developing mathematical models (linear or non-linear) is presently the best method of analyzing hydrologic time series. These mathematical models are developed with statistical inference of parameters estimated from available samples.

This paper is concerned with the properties of maximum surplus, maximum deficit and maximum range. Specifically, it deals with their distributions, starting from the probability distribution of a variable and from the mathematical model of dependence in the corresponding stationary time series.

3. Terminology related to the analysis of surplus, deficit and range. A time series may only have a deterministic component, which consists of either cyclic (or combination of cycles) or of trends (and jumps), or it may have only a stochastic component. Or, it may be a combination of deterministic and stochastic components. If a hydrologic magnitude is cumulative in nature, and has a substantial stochastic component, the theory of stochastic processes may be applied to determine the probabilities of water surplus, deficit or range (or storage). Determination of these probabilities must be based on a given inflow regime into the storage space and a given outflow regime. When the stochastic theory is applied to waiting lines (especially human lines), the methods developed for the probability distribution of the accumulated line are encompassed by the theory of queues (often called the queueing theory, or queueing process and bulk service). When the stochastic term is applied to inventory or production problems, methods of computing the surplus, deficit or storage are called the inventory problem, theory of provisioning or probability theory of storage systems. Techniques developed when applying the stochastic term to water surplus, deficit and storage in lakes and reservoirs are usually called the probability theory of reservoir storage, storage problem, probability theory of storage system, dam theory (where the word "dam" replaces the word "reservoir storage"), or theory of dams.

Generally when studying accumulated deviations as part of the stochastic process theory, the following terms are used: partial sums of a finite number of variables (independent normal or any other), sums of independent or dependent random variables, and maximum ranges. As this problem of accumulated deviations has many applications, the terms "maximum surplus," "maximum deficit" and "maximum range" or simply, "surplus," "deficit"

* Professor of Civil Engineering, Colorado State University, Fort Collins, Colorado.

and "range" are used here exclusively.* It covers most of the techniques which are encompassed in the probability theory of storage systems.

4. Short historical review. Contributions to the analysis of maximum surplus, maximum deficit and maximum range by various authors as applied to water resources problems are not summarized in this introduction. However, the basic ideas and mathematical expressions developed by some authors are given and discussed in the following chapters of this paper.

W. Rippl [1], in 1883, first used cumulative curves (mass-curves) of river flow to determine the capacity of storage reservoirs for water supply. From that time until the present, mass-curves have been used extensively when designing storage reservoirs, and many particular variations of the method have been developed. The following is an example of the application of mass-curve: Assume that the river flow for each year should be regulated to the mean flow of that particular year. The mass-curve for that year will produce the necessary storage or range.

* The definition of this range should not be confused with the concept of the range as the difference $X_{max} - X_{min}$ in a sample of size N of a variable X.

For N years of observations there are N values of range. These values then represent a new sample that supports the study of the probability of range.

5. Subjects of this paper. The various and detailed definitions of maximum sum (surplus), minimum sum (deficit) and maximum range are elaborated on in Chapter II. Chapter III deals briefly with the applications of surplus, deficit and range as techniques of the probability theory and mathematical statistics for the analysis of hydrologic problems. This study probes general and particular cases of the distributions of surplus, deficit and range, for given properties of a variable (the probability density function and the mathematical model of dependence in time for a stationary time series). These cases are outlined in Chapter IV and treated in subsequent chapters.

In this study the analysis of surplus, deficit and range refer only to the population (universe) of a variable. This study does not deal with the statistical inference about the properties of the population starting from the available sample. However, in many cases distributions of statistical parameters, as summarized from available literature, or developed in this paper, enable the statistical inference to be carried out.

CHAPTER II

DEFINITIONS OF MAXIMUM SURPLUS, MAXIMUM DEFICIT AND MAXIMUM RANGE

1. Cumulative series of a variable. Let X_1, X_2, ... , be a sequence of non-negative random variables. Let for n = 1, 2, ...

$$C_n = X_1 + X_2 + \ldots + X_n \qquad 2.1$$

with $C_n \leq C_{n+1}$, and with the understanding that $C_o = 0$ for $n = 0$. For river flows, X_i may represent the total flow for the i-th year, and C_n the cumulative flow for all the years 1, 2, ... , n. Let the sample size consist of N values, while n is a variable number, and let

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i = \frac{C_N}{N} \qquad 2.2$$

If in eq. 2.1 each X_i is replaced by $\overline{X}$, then $C_n(\overline{X}) = n\overline{X}$, for n = 1, 2, ... , N. If $\overline{X}$ is the average annual outflow, $C_n(\overline{X})$ represents the situation of a constant outflow for a period of n years, equal to the average outflow.

Figure 2.1, (1), shows an example of cumulative sums C_n as it changes with n for a sample of size N. The straight line $C_n(\overline{X})$ is also plotted on this graph, (2). For $n = 0$, $C_o = 0$.

In this study the cumulative series of a variable, and the discrete-time series are defined in a particular manner. A hydrologic process of flow or precipitation is a continuous-time series (zero values included). By selecting a unit period, Δt, (day, month, year), the sequence of the total or average flow or precipitation for this unit period forms a discrete-time series. Authors approach this case several ways in literature. Some authors replace the continuous process by point values. For example, Moran [8] considers the annual inflows into reservoirs and outflows from them as concentrated values at points, or as instantaneous values at given time intervals (end of years). Similarly, Anis [6] considers that the cumulative series C_n of a variable does not start at zero but as X_1. The definition of cumulative-time series in this text is based on the assumption that the flow or precipitation within a selected unit period (day, month, year) is uniform. This uniform value produces the same total value at the end of a unit period as the actual non-uniform flow or precipitation. In other words, if an annual value of non-uniform river flow or precipitation is X_i, it is assumed when defining the cumulative series of X that X_i is obtained from a uniform flow or precipitation inside the unit period. By this definition, $C_n = 0$ at $n = 0$ (or $t = 0$), and $C_n = X_1$ at $n = 1$. Practical application of this assumption makes C_n a continuous series in the form of a polygone with breaking points at 0, X_1, ... , X_n, and not as pure discrete ordinates. The selection of $n = 0$ or $t = 0$ (initial time) is necessary to any regulation problem, and from that point the accumulation of input and output is usually counted.

In this study the values n and N do not represent the number of ordinates in a sample of discrete time series. These values do represent the number of unit periods Δt for which the variable values are computed, either as total sums or as mean values. When considering a time unit of one year, Δt, river runoff is the mean or the total annual flow representing the variable values, and n or N are numbers of years. In this way (n + 1) ordinates have n unit periods. This fact should be remembered whenever comparing the results and formulas of this study with those which consider n as the number of discrete ordinates.

The difference between X, and a given constant, X_o, given as $\Delta X_i = X_i - X_o$, is the deviation or departure of X_i from X_o. It is to be noted that

$$\sum_{i=1}^{N} \Delta X_i = \sum_{i=1}^{N} X_i - NX_o,$$

and this sum is zero if and only if $X_o = \overline{X}$. The cumulative sum of deviations from X_o is defined as

$$S_n(X_o) = \sum_{i=1}^{n} \Delta X_i = \sum_{i=1}^{n} X_i - nX_o = C_n - nX_o, \qquad 2.3$$

for n = 1, 2, ... , N, and for completeness also $S_o(X_o) = 0$ for any X_o. If a reservoir has a constant outflow, X_o, and random inflows X_1, X_2, ... , then $S_n(X_o)$ denotes the total water storage after n years, with surplus of storage if $S_n(X_o) > o$, and deficit of storage if $S_n(X_o) < o$.

Two methods are used when plotting cumulative curves: (1) Cumulative sum of the variable, C_n, as in fig. 2.1; and, (2) Cumulative sum of deviations, $S_n(X_o)$, from a selected constant value X_o. Usually, this value X_o is the mean for the total period of observations as shown in fig. 2.2, upper graph, or it is a variable parameter. The second method of representation is preferable from the standpoint of accuracy and ease of graph manipulation. Even though this fact is known, this study employs both methods of plotting (as in figs. 2.1 and 2.2) for the purpose of defining various types of surplus, deficit and range.

The basic value X_o from which the deviations are calculated can be considered either as independent or dependent on the sample values. If the release of water X_o, from a reservoir, is

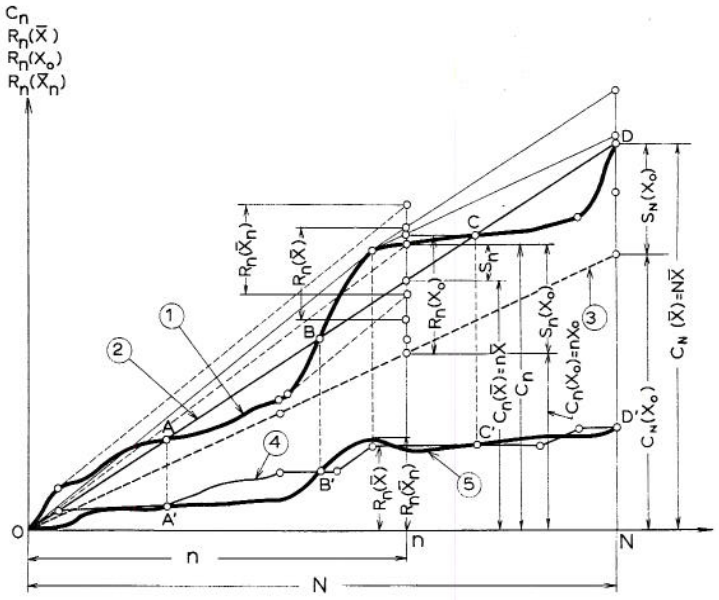


Fig. 2.1 Definitions of surplus, deficit, range, adjusted surplus, adjusted deficit and adjusted range, as well as of surplus, deficit and range for any base value X_o and any variate value n: (1) Cumulative sum C_n of the variable X; (2) Cumulative sum of constant value $\bar{X}$, given as $C_n(\bar{X}) = n\bar{X}$; (3) Cumulative sum of constant value X_o, given as $C_n(X_o) = nX_o$; (4) The change of the range, R_n, as n increases from 0 to N; and, (5) The change of the adjusted range, $R_n(\bar{X}_n)$, as n increases from 0 to N.

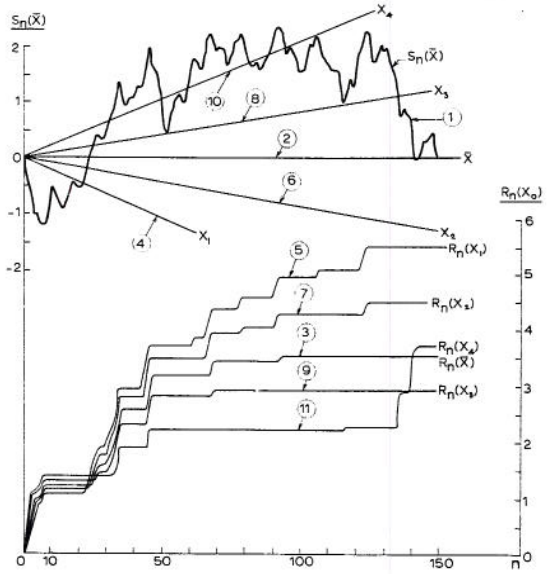


Fig. 2.2 Cumulative sums, $S_i(X_o)$, of deviations $\Delta X_i = X_i - X_o$ for five values of X_o (upper graph) and the sequence of range, $R_n(X_o)$, as n increases from 0 to N, for five values of X_o (lower graph); (1) Cumulative sum, $S_n(\bar{X})$; (2) Cumulative sum of $\bar{X}$; (3) Range, R_n, as n increases from 0 to N; (4) and (6) Cumulative sums of $(X_i - X_o)$ with $X_o < \bar{X}$; (5) and (7) The change of $R_n(X_o)$ as n increases from 0 to N for $X_o < \bar{X}$; (8) and (10) Cumulative sums of $(X_i - X_o)$ for $X_o > \bar{X}$; (9) and (11) The change of $R_n(X_o)$ as n increases from 0 to N for $X_o > \bar{X}$. The graphs refer to the relative values $X_i = V_i/\bar{V}$ (V = annual flows and $\bar{V}$ = mean annual flow) of the Göta River in Sweden for N = 150.

prescribed in advance as a constant, then X_o is independent of $S_n(X_o)$. Furthermore, X_o may be a function of time, but still independent of $S_n(X_o)$. In other words, the outflow regime is independent of the inflow regime and the storage in the reservoir (volume or elevation of stored water). If X_o is a function of the inflow regime, or of the water stored in the resevoir, then X_o is dependent on $S_n(X_o)$. In practice, the outflow is a function of the water stored in the reservoir, the predicted future inflow and the water demand. Thus, the outflow varies either continuously or discontinuously with time.

This study probes the simple case of a constant X_o either for the length N or subsamples n. The two cases: (a) X_o changes with time inside a given n, and is independent of $S_n(X_o)$; and (b) X_o changes with time and is a function either of inflows or of $S_n(X_o)$, as further generalizations, are not considered in this paper.

In this study a time series of sample size N is used for various definitions. Definitions also refer to an infinite stationary time series of a variable X, with the mean μ. In this case $\bar{X}$ should be replaced in definitions by the population mean μ.

2. Definition of maximum and minimum sums of deviations. The sequence of the sums of the deviations of X_i from X_o: $S_o(X_o)$, $S_1(X_o)$, ... , $S_n(X_o)$, for each n, has a maximum and a minimum value.

Let

$$S_n^+(X_o) = \max[S_o(X_o) = 0, S_1(X_o), \ldots, S_n(X_o)] \qquad 2.4$$

as the maximum of the sums of the deviations, and

$$S_n^-(X_o) = \min[S_o(X_o) = 0, S_1(X_o), \ldots, S_n(X_o)] \qquad 2.5$$

as the minimum of the sums of deviations. The probability distributions of these two parameters depend on the joint distribution of $(X_1, X_2, \ldots, X_n)$, or in the case of stationary time series on the distribution of the variable X and its patterns in time series sequence.

It is obvious from the above definitions and $S_o(X_o) = 0$ that $S_n^+(X_o) \geq 0$, and $S_n^-(X_o) \leq 0$.

If $X_o = N^{-1} \sum_{i=1}^{N} X_i = \bar{X}$, the sums $S_n^+(\bar{X})$ and $S_n^-(\bar{X})$ will be simply denoted as S_n^+ and S_n^-. The variable $S_n^+(X_o)$ will be called here the maximum surplus, and the variable $S_n^-(X_o)$ the maximum deficit, for a given X_o and n.

Another method of defining and calculating the maximum and minimum sum of deviations, for each n, is to take deviations from the mean of the first n values. Thus, let

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i, \quad n = 1, 2, \ldots, \qquad 2.6$$

and let

$$S_j(\bar{X}_n) = \sum_{i=1}^{j} (X_i - \bar{X}_n) \qquad 2.7$$

with $j = 0, 1, 2, \ldots, n$. For example, if $n = 3$, then $S_j(\bar{X}_n)$ is

$$S_o(\bar{X}_3) = 0$$

$$S_1(\bar{X}_3) = X_1 - \bar{X}_3$$

$$S_2(\bar{X}_3) = X_1 + X_2 - 2\bar{X}_3, \text{ and}$$

$$S_3(\bar{X}_3) = X_1 + X_2 + X_3 - 3\bar{X}_3 = 0.$$

It is obvious from the definition of $\bar{X}_n$ in eq. 2.6 that $S_n(\bar{X}_n) = 0$. From the double sequence $S_j(\bar{X}_n)$, $j = 1, 2, \ldots, n$; and $n = 1, 2, \ldots$, the maximum sum of deviation is

$$S_n^+(\bar{X}_n) = \max[0, S_1(\bar{X}_n), S_2(\bar{X}_n), \ldots, S_{n-1}(\bar{X}_n), 0] \qquad 2.8$$

and the minimum sum of deviation $S_n^-(\bar{X}_n)$ is similarly defined. This maximum sum is called the adjusted maximum sum or the adjusted maximum surplus. The minimum sum is called the adjusted minimum sum or the adjusted maximum deficit. W. Feller [4] called the difference of these two sums the adjusted maximum range which is defined later in this text.

3. Definition of maximum range for a constant value X_o. The maximum range for a given constant value X_o is defined here as the difference between

$$S_n^+(X_o) \text{ and } S_n^-(X_o), \text{ or}$$

$$R_n(X_o) = S_n^+(X_o) - S_n^-(X_o), \qquad 2.9$$

with $R_n(X_o)$ as a non-decreasing function of n, for a given sample N, or

$$0 \leq R_n(X_o) \leq R_{n+1}(X_o), \text{ for all } n. \qquad 2.10$$

By definition $R_o(X_o) = 0$. For $n = 1, 2, \ldots$ then

$$S_{n+1}^+(X_o) = \max[0, S_1(X_o), \ldots, S_{n+1}^+(X_o)] \geq$$

$$\geq \max[0, S_1(X_o), \ldots, S_n(X_o)] = S_n^+(X_o), \text{ and}$$

similarly $S_{n+1}^-(X_o) \leq S_n^-(X_o)$. Thus,

$$R_{n+1}(X_o) = S_{n+1}^+(X_o) - S_{n+1}^-(X_o) \geq S_n^+(X_o) - S_n^-(X_o) = R_n(X_o).$$

The properties of $R_n(X_o)$ for a variable X, therefore, depend on n and X_o. There must be a distinction between n and X_o. This distinction is necessary because both factors can be considered as changing parameters or variables (X_o can take any value from X_{min} to X_{max} while n can take on only discrete values of integers). To fulfill this

requirement n will be referred to as a variate.

Figure 2.1 shows the sums of the variable X, as well as the increase of range $R_n(X_o)$ with n, (4). It does not show the distribution of $R_n(X_o)$.

4. Definition of maximum range for the special case $X_o = \bar{X}$. Taking $\bar{X}$ = mean of the available sample size N, as a special value of X_o, then

$$R_n(\bar{X}) = S_n^+ - S_n^- \qquad 2.11$$

The values S_n^+ and S_n^- are the maximum and minimum values of the sums $S_n(\bar{X})$ in a subsample of size n, where $S_n(\bar{X})$ is determined by

$$S_n(\bar{X}) = C_n - C_n(\bar{X}) = C_n - n\bar{X} \qquad 2.12$$

as shown in fig. 2.1, or as

$$S_n(\bar{X}) = S_n(X_o) - \frac{n}{N} S_N(X_o) \qquad 2.13$$

Figure 2.1, (4), shows the maximum range $R_n(\bar{X})$ as it changes with an increase of n. Figures 2.1 and 2.2 show only how numerous variables in the form of sums, maximum surplus, maximum deficit and maximum range change with an increase of n. If a series of sample size N is divided into m parts or subsamples, each with the length n, and if for each subsample the corresponding statistics are determined for a given X_o, $\bar{X}$ or $\bar{X}_n$, then m values for each of these variables are obtained. This enables the determination of distributions and patterns in sequence of these statistical parameters.

Figure 2.2, (1), shows the cumulative sum of deviations

$$S_n(\bar{X}) = \sum_{i=1}^{n} (X_i - \bar{X})$$

for annual flow of the Göta River in Sweden for the period 1807 - 1957 (150 years). It is given as $S_n(X)/\bar{X}$, where $\bar{X} = \bar{V}$ is the average annual flow. The computed values $S_n(\bar{X})$, or any other $S_n(X_o)$, as well as the statistics $S_n^+(X_o)$, $S_n^-(X_o)$, and $R_n(X_o)$ must be multiplied by $\bar{V}$ (in this case $\bar{V} = 16.2$ in 10^9 m^3) in order to obtain their values in cubic meters.

In this study the maximum surplus, maximum deficit and maximum range, which correspond to $X_o = \bar{X}$, are called surplus, deficit and range, respectively. When these terms refer to range for X_o an understanding is that the terms always mean maximum surplus, maximum deficit and maximum range, respectively, for a given X_o.

The range $R_n(\bar{X})$ represents the storage capacity necessary in a reservoir, if the fluctuations of flows could be suppressed for a period of n time units. The expected value of $R_n(\bar{X})$ increases with an increase of n. Also, the range according to H. E. Hurst [1], [2] and [3] can be conceived as: (a) the maximum accumulated storage when there is no deficit in outflow (for the outflow equal to the mean), with $R_n = S_n^+$, as the range is equal to the surplus; (b) the maximum deficit, when there is never any surplus with $R_n = S_n^-$ or the range is equal to the deficit; and, (c) the sum of accumulated surplus and accumulated deficit, when both surplus and deficit exist, or $R_n = S_n^+ - S_n^-$. The same concept is valid for any value of X_o with a constant outflow X_o which creates either a maximum surplus, a maximum deficit or both. It should be pointed out that in case of a deficit $S_n^-(X_o)$ the constant outflow X_o can be supplied downstream during n unit periods only if there is an equal or greater surplus stored from the previous unit periods.

5. Definition of maximum adjusted range. The range for a given n is defined as

$$R_n(\bar{X}_n) = S_n^+(\bar{X}_n) - S_n^-(\bar{X}_n) \qquad 2.14$$

where $\bar{X}_n$ is the mean for the particular length of n unit periods. W. Feller [4] entitled this range the maximum adjusted range or simply the adjusted range. For any subsample of length n with $n < N$, the mean is $\bar{X}_n$, and considered a sampling statistic. The sums of deviations, when $\bar{X}_n$ is determined for any period of n time units, may be obtained by

$$S_i(\bar{X}_n) = C_i - C_i(\bar{X}_n) \qquad 2.15$$

The last value in fig. 2.1, (4), point D', is $R_N(\bar{X})$ of eq. 2.11, and is also the adjusted range for $n = N$. As $\bar{X}$ is also the mean for the points A, B, and C, the values of range from line (4) are at the same time the values of adjusted range, points A', B' and C' of line (5).

6. Comparison of the three types of surplus, deficit and range. The expected values of surplus, deficit and range increase with an increase in n for a given constant value of X_o. However, the expected change of these values for a given n with a changing X_o is somewhat different. For instance, if $X_o = 0$ the ranges are equivalent to the values of C_n of cumulative sums, fig. 2.1, (1). This equality is due to the fact that $S_n^-(X_o = 0)$ is always zero and $S_n^+(X_o = 0)$ increases steadily as n increases and is equal to C_n. When X_o increases toward the population value μ (estimated by $\bar{X}$) the expected value of the range decreases as the difference $\mu - X_o$ decreases for a given n. This difference results primarily from an accumulated surplus, because the deviations, $X_i - X_o$, are more positive than negative for $X_o < \mu$. When X_o is very close to the population mean μ, the expected value of the range for a given n is a minimum. When X_o increases beyond the value of μ, the expected value of the range for a given n increases in comparison with the corresponding expected value of the range for μ. The negative deviations, $X_i - X_o$, appear more fre-

quently and are of greater absolute value than positive deviations. They are responsible for an accumulated deficit. If X_o still increases and approaches infinity, the expected value of the range also increases toward infinity for a given n.

Ranges are given in fig. 2.2 for five values of X_o: X_1 and X_2 (smaller than $\bar{X}$), $\bar{X}$, and X_3 and X_4 (greater than $\bar{X}$). The cumulative sums of the deviations of X from these five values are given in fig. 2.2, lines (4), (6), (2), (8) and (10). Ranges as they increase with an increase of n for each of these five values of X_o are also seen in fig. 2.2, lines (5), (7), (3), (9) and (11). The range for a small n in a particular sample may be greater than for a larger n because of sampling fluctuations. This difference for small and large n results from a large variation of range about its mean. Variations may be such that for a short time series the range for $X_o + \bar{X}$ can be smaller than that for $X_o = \bar{X}$, fig. 2.2, line (11).

Figure 2.1, line (5), shows the adjusted range as n increases from 0 to N. There is one difference between the adjusted range $R_n(\bar{X}_n)$ and the range R_n as n increases from 0 to N. This difference is that for a given sample the former is without sharp steps and can either increase or decrease with an increase of n, whereas, the latter can only increase. Figure 2.1, lines (4) and (5), give this comparison of range and adjusted range.

Figure 2.3 gives in a simple way the definitions of the following variables: (1) the sums of deviations $X_i - \bar{X}$, line (1); (2) the $\bar{X}$ - sum represented by the line (2); (3) the $\bar{X}_n$ - sum, line (3); and (4) the X_o - sum, from 0 to N, line (4). The nine values are shown in the figure: surplus, $S_n^+ = S_n^+(\bar{X})$; deficit, $S_n^- = S_n^-(\bar{X})$; range, $R_n = R_n(\bar{X})$; adjusted surplus, $S_n^+(\bar{X}_n)$; adjusted deficit, $S_n^-(\bar{X}_n)$; adjusted range, $R_n(\bar{X}_n)$; surplus for X_o, $S_n^+(X_o)$; deficit for X_o, $S_n^-(X_o)$; and range for X_o, $R_n(X_o)$.

If the range and the adjusted ranges are divided by any value X_i, they become the relative ranges. If these values are especially selected to be $\bar{X}$, or X_o, or $\bar{X}_n$, then the ratios $R_n/\bar{X}$; $R_n(X_o)/\bar{X}$; $R_n(\bar{X}_n)/\bar{X}$; $R_n(\bar{X}_n)/\bar{X}_n$ or similar are called relative ranges. The relative ranges $R_n/\bar{X}$ for the annual flows of the Göta River are given as line (3), fig. 2.2. The other lines, (5), (7), (9) and (11), fig. 2.2, are given as relative values $R_n(X_o)/\bar{X}$. If the variable X is standardized with the new variable $x = (X - \bar{X})/s$, then the range refers to a sample with a mean of zero and a standard deviation unity. In the above expression s = standard deviation of X for the sample of size N.

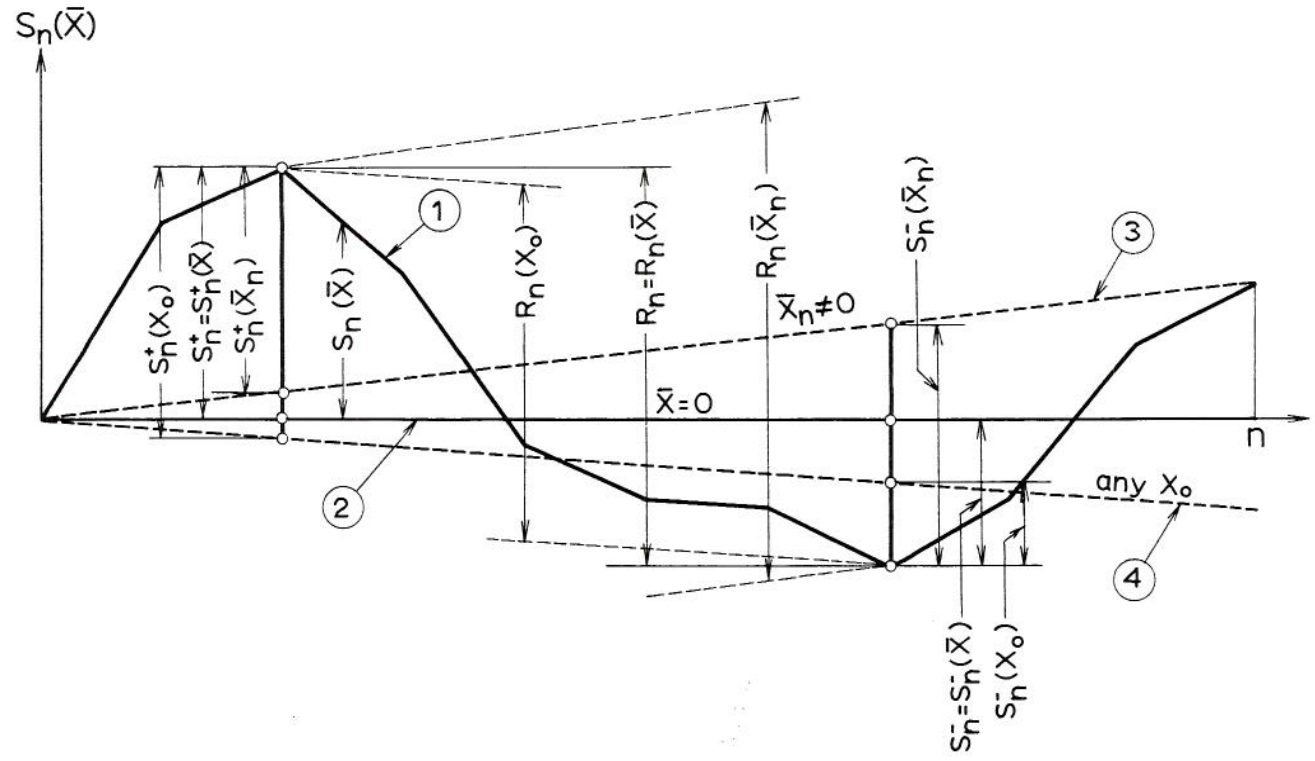


Fig. 2.3 Definitions of surplus, deficit and range: (1) The cumulative sum of deviation, $\Sigma(X_i - \bar{X})$; (2) The sum zero of $\bar{X} = 0$; (3) The sum of $\bar{X}_n$; (4) The sum of X_o. The nine values: S_n^+, $S_n^+(\bar{X}_n)$, $S_n^+(X_o)$; S_n^-, $S_n^-(\bar{X}_n)$, $S_n^-(X_o)$; and R_n, $R_n(\bar{X}_n)$, $R_n(X_o)$ are shown in the figure.

CHAPTER III

APPLICATIONS IN HYDROLOGY

1. Cumulative magnitudes. Generally it is feasible to apply statistical parameters in the form of surplus, deficit and range to any physical magnitude which can be accumulated in a given space, such as: heat, kinetic energy, water vapor, water, water moisture, sediment, mineral content in water, oxygen content in water, pollutents in water, biological matters in water, etc. Thus, any hydrologic magnitude of a cumulative nature may be analyzed by surplus, deficit and range. It is feasible to investigate storage problems with this type of analysis when the following three factors are involved: (a) the characteristics of the storage space (storage response to inflow and outflow); (b) input or inflow into the storage space; and (c) output or outflow from the storage space.

Flow regulation by storage volumes is one of the basic hydrologic problems. The importance of this problem warrants the following discussion on stochastic problems in design and operation of reservoirs. However, the surplus, deficit and range approach can also be used for the analysis of hydrologic time series without referring to storage problems.

2. Independent and dependent reservoirs. A storage reservoir which is operated independently of any other reservoir is called an independent reservoir. If its design and operation are dependent on other reservoirs, it is called a dependent reservoir. Dependent reservoirs are of these three general types: (a) Inflow depends partly or wholly on the regulated outflow of upstream reservoirs; (b) Outflow is governed by joint operation with upstream and downstream reservoirs; and (c) Outflow is affected by reservoirs in adjacent or distant river basins; or combinations of these three types.

Surplus, deficit and range may be used to analyze stochastic design problems of independent reservoirs or of those dependent reservoirs whose characteristics of eventual dependent inflows and/or imposed outflows by the other reservoirs are known or prescribed in advance. Complex stochastic problems in design and/or operation of a system of dependent reservoirs and their solution represent a further generalization in the application of surplus, deficit and range. However, solutions of stochastic problems of individual reservoirs give the basic elements in design of a system of reservoirs.

3. Basic storage equation. The basic classical continuity equation in the design of reservoirs is

$$I - O = S \qquad 3.1$$

with I = inflow, O = outflow, and S = change in reservoir storage, in a given time interval T. Neglecting both the groundwater portion of a predominantly surface storage reservoir and the seepage out of the reservoir; but including the evaporation from the reservoir and the sedimentation of it and passing to the rates of inflow, outflow, evaporation and storage then,

$$P_t - Q_t - E_t = \frac{dS}{dt} \qquad 3.2$$

with P_t = inflow rate, which is a stochastic variable; Q_t = outflow rate, which is also a stochastic variable; E_t = evaporation rate from the reservoir, which is also a stochastic variable because it is dependent on the climatic stochastic movement, and reservoir surface. The last term in eq. 3.2 is the rate of change in stored water. Storage volume of a reservoir, S, is a function of both the reservoir elevation, H, and the time, t, and it can often be approximated by

$$S = a H^m \qquad 3.3$$

with $a = \psi(t)$ and $m = f(t)$ as functions of time. The inflow of sediments into a reservoir is a stochastic variable. Thus, $\underline{a}$ and $\underline{m}$ are stochastic variables. The basic input-storage-output relationship of flow regulation by reservoirs, presented by eq. 3.2, is an ordinary differential equation of stochastic variables. By introducing the functions $a = \psi(t)$ and $m = f(t)$ into eqs. 3.2 and 3.3, eq. 3.2 becomes a partial differential equation of stochastic variables.

Storage capacity, S_f, of a reservoir is a finite value. It is a stochastic variable because $S_f = a\,(H_{max}^{\ m} - H_{min}^{\ m})$, where H_{max} and H_{min} are the maximum and the minimum reservoir heights, with $\underline{a}$ and $\underline{m}$ stochastic variables. Practical applications allow the above variables to be neglected under the following conditions: (a) If the average annual evaporation E_t from a reservoir is small in comparison with the average annual inflow and outflow; and (b) If the sediment inflow is small in comparison with the finite storage capacity. In this case, the stochastic variables in eqs. 3.1 and 3.2 are the inflow and the outflow and storage volume. The water storage problem of the reservoir can be described by stochastic variables and their parameters.

4. Change of characteristics of inflow and outflow with time. The inflow changes with time because of natural fluctuations. However, its mean, variance, skewness coefficient and time dependence may change with time because of various changes and developments in the river basin. These changes can be assessed, but usually with a small amount of accuracy. This fact limits the insistence for exceptional accuracy in determining parameters of inflow as a stochastic variable.

The outflow changes with time because of unavoidable changes in objectives of storage use and because of influences by various river basin developments. Personnel that design and operate reservoirs must solve an ordinary differential equation with stochastic variables which are nonstationary. These variables are not stationary because of evolving conditions in the environment. This complexity explains why there are so many approaches to solving stochastic problems in the design of reservoirs.

5. Methods of solving stochastic problems in design of reservoirs. Approaches currently used in solving stochastic problems in design and operation of reservoirs may be classified in three large groups:

(1) Empirical method. This method uses mass curves of available flow time series to derive various variables associated with storage.

(2) Data generation method. This method solves stochastic storage problems by generating large samples of data. Statisticians call it the Monte Carlo Method. Hydrologists denote it as synthetic hydrology, simulation, data generation, or operational hydrology. The data generation method uses random numbers of one or several variables (normal, log-normal, gamma, or other theoretical distribution functions; or empirical distributions), with the stochastic dependence process or cyclic movement superimposed. Final treatment of generated samples is similar to the empirical method.

(3) Analytical method. This method consists of mathematical derivations of exact properties for various variables related to storage problems. Difficulties in integrating exact distribution equations and sequence patterns in a time series usually lead to the application of a numerical finite differences method.

This paper deals with the application of these three methods in analyzing storage problems by the properties of surplus, deficit and range. Potentials and limitations of these methods are of significance when applied to the water resources field in general, and storage problems in particular.

6. Variables which describe natural flows. The instantaneous discharge is the basic stochastic variable in describing river flows. However, the daily, monthly and annual flows are used as variables in practical problems. Properties of instantaneous inflow may be considered as approximated by properties of daily flows.

Annual flow, as a stochastic variable, removes the cycle of a year and any of its harmonics. Recent investigations by the writer [9, 10] on a large number of river gaging stations resulted in the conclusion that there is no evidence of cycles greater than a year in the sequence of river flows. However, the change in water carryover in river basins from year to year creates a dependence in time series of annual flow. This dependence can be described mathematically mostly by the first or second order Markov linear models (autoregressive schemes), or moving average schemes of various types.

Annual flows of several hundred rivers investigated show two extremes of time dependence as encountered in their series: (a) Independent variables; and (b) Dependent variables with the first order linear Markov dependence model. In some cases, the second order linear Markov model fits the correlograms of annual flows. Whenever a large storage capacity for overyear flow regulations is being designed or operated, the inflows on annual basis may be described by corresponding stochastic mathematical models.

If river flows are not affected by some important accident in nature, and if the inconsistency (man-made systematic errors in data) and non-homogeneity in data (man-made changes in river basin) are negligible, the series of annual flow are usually second order stationary (the expected mean, the variance and the autocovariance are independent of the position in the series, and ergodicity requirement is satisfied). If not, the non-stationarity (linear or non-linear trends) must be removed and the new stationary series as expected to be experienced in the future should be used in design and operation of reservoirs.

The sequence in time of monthly flows shows a cyclic movement of 12-month or its harmonics (usually 6-months), and a stochastic movement. Mathematical description of monthly flow time series becomes feasible in the light of sampling errors which are inherent in the limited period of observation of monthly flows. This description is usually composed of three parts: (a) Cyclic movement; (b) An independent stochastic component; and (c) A stochastic process, usually of the first or second order Markov linear models.

7. Variables which describe reservoir outflows. The reservoir outflows are usually expressed as the same variable (instantaneous, daily, monthly or annual flow) as the inflow. A similar mathematical approach may be used in describing reservoir outflows. In the case of lakes with no artificial flow regulation, the outflows are subject to a larger time dependence and usually smaller variations than the inflows, but their description is similar. The rigorous mathematical description of outflows as stochastic processes is less suitable in the case of outflows regulated by reservoirs.

A systematization of types of regulated outflows from a mathematical point of view gives the following general cases:

(1) Outflow is constant and equal to the estimate of the mean. Assuming the mean inflow is equal to the mean outflow, $\overline{Q} = \overline{P}$, then,

$$P - \overline{P} = \frac{dS}{dt} . \qquad 3.4$$

(2) Outflow is constant for a given period of n-time units and is equal to the average inflow, $\overline{P}_n$, of that period, so that $Q = \overline{P}_n$, with $\overline{P}_n$ a stochastic variable. The value $\overline{P}_n$ changes from one n-time unit period to another. Its variation decreases with an increase of n. Then,

$$P - \overline{P}_n = \frac{dS}{dt} . \qquad 3.5$$

This means that after n years the reservoir storage is always at its initial stage.

(3) Outflow is prescribed only by the water demand as $Q = \overline{Q} + \overline{Q}\psi(\tau)$, with $\psi(\tau)$ a twelve-month function, with τ the time of the year, and $E\psi(\tau) = 0$. Its variation about zero depends on the seasonal patterns of water demand. Then, for $\overline{Q} = \overline{P}$,

$$P - \overline{P} \quad [1 + \psi(\tau)] = \frac{dS}{dt} . \qquad 3.6$$

The integration of eq. 3.6 depends upon how well $\psi(\tau)$ as a mathematical function, eventually with stochastic components, describes the actual water release.

(4) Outflow depends on the storage in the reservoir as $Q = \overline{Q} + \overline{Q} f(s) = \overline{Q}\,[1 + f(s)]$, so

that for $\overline{Q} = \overline{P}$

$$P - \overline{P}\,[1 + f(S)] = \frac{dS}{dt} \qquad 3.7$$

with $E\,f(S) = 0$. The variation of outflow depends on storage variation, which in turn depends on inflow variation and reservoir characteristics.

(5) Outflow depends on the inflow into the reservoir, or

$$Q = \overline{Q} + \overline{Q}\ (P) = \overline{P}\,[1 + \theta\ (P)], \text{ so that}$$

$$P - \overline{P}\,[1 + \theta\ (P)] = \frac{dS}{dt} \qquad 3.8$$

with $E\ \theta\ (P) = 0$.

(6) Outflow depends on both storage in the reservoir and inflow, or $Q = \overline{P}\,[1 + \phi\,(S, P)]$, so that

$$P - \overline{P}\,[1 + \phi\,(S, P)] = \frac{dS}{dt} \qquad 3.9$$

with $E\,\phi\,(S, P) = 0$. The variation of $\phi\,(S, P)$ depends on the type of function, and the weight by which each S and P affect the outflow.

(7) Outflow is generally prescribed by the water demand, but is also dependent on storage in reservoir and on inflow, or $Q = \overline{P}\,[1 + \psi\,(\tau)\,\phi\,(S, P)]$, so that

$$P - \overline{P}\,[1 + \psi\ (\tau)\,\phi\,(S, P)] = \frac{dS}{dt} \qquad 3.10$$

with $E\ [\psi\,(\tau)\,\phi\ (S, P)] = 0$. The variation of $[\psi\,(\tau)\,\phi\ (S, P)]$ depends on the weight by which each of the three variables: τ, S, P, affect the outflow. In practice, the demand is prescribed, but it is usually modified by the water available in reservoir storage and by the anticipated inflows.

There may be various types of the functions $\psi\,(\tau)$, $f\,(S)$, $\theta\,(P)$ and $\phi\,(S, P)$ and their combinations. Expanded in power series forms, their linear terms give first order approximations which are the simplest to investigate. When these functions become complex, they prohibit simple mathematical analysis. Usually analysis requires the use of the finite differences method in integration, as seen in eqs. 3.6 through 3.10. Outflow regimes (1) and (2), eqs. 3.4 and 3.5, are theoretical but they have practical applications as limit cases. They provide information concerning the required storage capacities and storage fluctuations for theoretical regulation patterns.

8. Infinite storage. Even though reservoir storage capacities are always finite, the theoretical concept of infinite storage is useful as a limiting factor when treating stochastic problems in the design of reservoirs. This concept may bear various names in different literature such as: infinite reservoir, infinite dam, infinite storage, infinite sum of deviations, and similar. A reservoir fulfilling the concept of infinite storage capacity requirements is assumed to be capable of storing any water surplus as incurred by the difference of inflow and outflow, and to supply any deficit for the difference between outflow and inflow.

This concept leads to the introduction of three basic and important variables into the stochastic analysis of storage problems: surplus, deficit and range. In general, the concept of infinite storage is not necessary for the definition of these three variables when applied to river flows, but it is useful as soon as these variables are associated with or applied to storage problems. It is assumed that infinite storage does not mean that the initial stage of storage is an empty reservoir. This concept does assume that on both ends of actual stage there is an infinite storage for accepting surplus or supplying the deficit.

9. Finite storage. As all reservoirs have limited storage capacities, practical problems are of the finite storage type. Finite storage is conceived as a stochastic process with two barriers, the upper with the full storage capacity, S_f, and the lower with the empty reservoir. The initial storage content, S_i, may be anywhere between 0 and S_f. This parameter, S_i, plays an important role in the operation of reservoirs until the operation becomes independent of the initial conditions.

Two factors make the analytical integration of storage differential equations or any other equation difficult: (a) The existence of two boundaries for storage, zero and S_f; and, (b) The impact of initial storage, S_i.

The interests in practical storage problems usually are in: (a) Probability distribution of water volumes stored in a reservoir at a given time, for given conditions; (b) Probabilities that a given storage volume is not exceeded in a given time; (c) Probability that the storage volume reaches either of barriers (full or empty reservoir) in a given period; (d) Probability that the reservoir is full or empty at a given moment, under given conditions; (e) Probability of time-on that a reservoir stays full or of time-off that a reservoir stays empty for a given period, once either of the two barriers are reached; (f) Probability of water excess beyond demand, once the reservoir is full and stays full, for a time period; or probability of water excess for each case of full storage; the same probabilities for the water deficiency for empty reservoirs; (g) Probabilities of range, surplus and deficit as defined above for the case of finite storage capacities; and similar problems; (h) Probability of a total water yield in a given time period under given conditions of storage operation, and similar problems.

10. Investigation of hydrologic time series. A hydrologic time series of the sample size N may be analyzed by using the properties of surplus, S_n^+, deficit, S_n^-, and range, R_n. The properties of these three parameters may be determined for simple distribution functions, simple mathematical models of sequential patterns and for stationary time series. These properties may be obtained by an analytical method, by a numerical integration of exact distribution functions, or by a data generation method. Characteristics of the basic variable and of the above three statistical parameters (S_n^+, S_n^-, and R_n) then become the bench-mark distributions and bench-mark sequential patterns. Investigators can derive conclusions on the characteristics of an observed time

series by comparing an observed time series and their S_n^+, S_n^-, and R_n (or other types of these three parameters) with the corresponding bench-mark characteristics of the variables and of their parameters S_n^+, S_n^-, and R_n. This approach permits the study of patterns in long-range hydrologic fluctuations, and especially the inference about the factors which produce the time dependence.

11. Complex hydrologic problems. When there are several storage reservoirs, many water resource problems and many water users in a river basin, the planning is usually carried out by using historic data and empirical hydrologic methods. Presently, there is a trend towards using the data generation method in hydrology. It consists of increasing the historic sample size by simulation of new data, while maintaining the distribution, stochastic and cyclic processes of the available small historic sample.

The contemporary advances in probability theory, mathematical statistics and stochastic processes permit probability methods to be used in hydrologic applications. The use of the properties of surplus, deficit and range represents potential techniques for the analysis of complex hydrologic problems.

CHAPTER IV

GENERAL CHARACTERISTICS AND METHODS OF DETERMINATION OF SURPLUS, DEFICIT AND RANGE

1. Stationarity and ergodicity conditions. In this paper the analysis of time series assumes that a time series is stationary. There are various types of stationarity in time series. The stationarity used in this analysis is specified by two basic conditions which are considered to be approximately satisfied: (1) The expected value of any X_i value in a time series is equal to the population mean which is constant, or

$$E(X_i) = \mu = \int_{-\infty}^{+\infty} X \, d[P(X)] \qquad 4.1$$

with $P(X)$ = the probability distribution function of X, and μ = the constant population mean; and (2) the expected value of covariance of X_i and X_{i+k} depends only on k and not on i; it is equal to the product of the population serial correlation coefficient ρ_k and the population variance σ^2

$$E\left[\text{cov}\ (X_i\ X_{i+k})\right] = \rho_k\ \sigma^2 \qquad 4.2$$

These two conditions make a time series second order stationary. Ergodicity is the next condition that should be satisfied. This condition means that time averages converge in probability to theoretical averages.

2. Distributions and time dependence of surplus, deficit and range. A discrete stationary time series (either independent or dependent) with a given probability distribution and size N may be considered as a random variable in N - dimensions (hyperspace). This time series may also be considered as many individual variables $X_1, X_2, \ldots X_N$ at the positions 1, 2, ... N, with the same probability distribution. Practical problems involving the dependence of X_i may be described either by the mathematical model of dependence or by using the joint distribution of N variables. The model of dependence is usually defined by its generating process and the characteristics of random independent variables involved in this process.

For a large series of size N any non-overlapping subseries of the size n has a corresponding value of each type of surplus, deficit and range: S_n^+, $S_n^+(X_o)$, $S_n^+(\overline{X}_n)$; S_n^-, $S_n^-(X_o)$, $S_n^-(\overline{X}_n)$; and R_n, $R_n(X_o)$, $R_n(\overline{X}_n)$. Their time series are stationary for a given n, if the time series of X is stationary. Each type of surplus, deficit or range is characterized by its probability distribution. In the case where the variable X is dependent, the surplus, deficit and range are also time dependent variables.

Assume that the variable X may be defined by its probability distribution and mathematical model of dependence of its stationary time series. Assume also that $a_1, a_2, \ldots$, are parameters of this probability distribution function, and $b_1, b_2, \ldots$, are parameters of the mathematical model of dependence. The general probability distribution functions of $S_n^+(X_o)$, $S_n^-(X_o)$, and $R_n(X_o)$ may be expressed in the form of families of curves, as function of the variate n, of the variable parameter X_o, and the above parameters a_i's and b_j's as

$$F[S_n^+(X_o) \le S_i] = F_s[S_n^+(X_o); a_1, a_2..; b_1, b_2..; X_o; n] \qquad 4.3$$

$$F[S_n^-(X_o) \ge S_i] = F_d[S_n^-(X_o); a_1, a_2..; b_1, b_2..; X_o; n], \qquad 4.4$$

and

$$F[R_n(X_o) \le R_i] = F_r[R_n(X_o); a_1, a_2..; b_1, b_2..; X_o; n] \qquad 4.5$$

with S_i and R_i values which $S_n^+(X_o)$, $S_n^-(X_o)$ and $R_n(S_o)$ can assume, respectively. The general mathematical expressions for dependence can be expressed in the form of families of curves as

$$\phi_s[S_n^+(X_o); a_1, a_2..; b_1, b_2..; X_o; n] = 0 \qquad 4.6$$

$$\phi_d[S_n^-(X_o); a_1, a_2..; b_1, b_2..; X_o; n] = 0 \qquad 4.7$$

$$\phi_r[R_n(X_o); a_1, a_2..; b_1, b_2..; X_o; n] = 0. \qquad 4.8$$

Suppose that X follows a normal distribution function with two parameters μ and σ. Also, suppose that the properties of a random independent variable, ϵ, are known and that the Markov first order linear model with the parameter ρ is the relationship between X and ϵ. Then, the properties of surplus, deficit and range may be described by their probability distributions and their models of time dependence which are not only functions of X_o and n, but also of μ, σ, ρ and the properties of ϵ_i.

If eqs. 4.3 through 4.8 were available for particular field conditions of hydrologic stochastic variables, they would be of value in water resources planning, design and operation. Statistical inference concerning the parameters of distribution and of time dependence is a prerequisite for the application of eqs. 4.3 through 4.8 to storage reservoirs and other water resources problems. In the case of empirical distributions and empirical models for time dependence of hydrologic variables, methods should be available for an exact or an approximate determination of the above functions.

The simplest application of eqs. 4.3 through 4.8 is when the number of parameters a_i and b_j is very small. In the case of the independent standard normal variable, none of these parameters enter into eqs. 4.3 through 4.8 so that in this case

$$F[S_n^+(X_o) \leq S_i] = F_s[S_n^+(X_o); X_o; n] \quad 4.9$$

$$F[S_n^-(X_o) \geq S_i] = F_d[S_n^-(X_o); X_o; n] \quad 4.10$$

$$F[R_n(X_o) \leq R_i] = F_r[R_n(X_o); X_o; n] \quad 4.11$$

and the surplus, deficit and range are also time independent variables.

W. Feller [4] states that it is practically impossible to analytically calculate the exact range distribution even for a simple form of the underlying probability density function $f(X)$. This is true even for a small value of n such as $n = 3$. He stresses that the sums $S_n(\overline{X})$ are normally asymptotically distributed and, therefore, the asymptotic distribution of the range is independent of the underlying function $f(X)$. Accordingly, it is sufficient to consider the case where the departures $\Delta X_i = X_i - \overline{X}$ are normally distributed.

In some cases asymptotic distributions of $S_n(\overline{X})$ have a small practical value in hydrology. This limitation is due to the fact that they depart significantly from the exact distributions of surplus, deficit and range for very small values of n. These values of n often are the most important cases in some applications.

3. Particular properties of probability distributions of surplus, deficit and range. The previous definitions and the above discussions reveal some particular properties of surplus, deficit and range. The values of surplus are always either zero or positive. Whenever the sum S_n of deviations $\Delta X_i = X_i - X_o$ has only zero and negative values for $i = 1, 2, \ldots n$, then $S_n^+ = 0$. In this case the probability distributions of S_n^+ are comprised of two parts: (1) a discrete part or probability mass for only one value, $S_n^+ = 0$; and (2) a continuous part or probability density function for values $S_n^+ \geq 0$. As the probability that S_n^+ will remain zero decreases with an increase of n, the discrete part of probability distribution for $S_n^+ = 0$ also decreases with an increase of n. This relationship takes place while the total area under the probability density curve increases with an increase of n. Therefore, the probability of S_n^+ from zero to a given value S_i is

$$F(S_n^+ \leq S_i) = F(S_n^+ = 0) + \int_{S_n^+ = 0}^{S_n^+ = S_i} f_s(S_n^+)\, dS_n^+ \quad 4.12$$

with

$$F(S_n^+ \leq \infty) = F(S_n^+ = 0) + \int_o^{\infty} f_s(S_n^+)\, dS_n^+ = 1 \quad 4.13$$

For $n \to \infty$ the value $F(S_n^+ = 0) \to 0$, and the last term of the continuous density function tends to the area unity.

The same relationship is valid for the deficit S_n^-, whose probability function is composed of a discrete part or the probability mass $F(S_n^- = 0)$, and a continuous density curve, so that

$$F(S_n^- \geq S_i) = F(S_n^- = 0) + \int_{S_n^- = S_i}^{S_n^- = 0} f_d(S_n^-)\, dS_n^- \quad 4.14$$

with the same properties of the two parts given by eq. 4.13, with S_i being a negative value.

As the range is the sum of the surplus and the deficit (deficit taken as the absolute value of S_n^-) for each value of $S_n^+ = 0$ there is a value for S_n^- which is also different from zero. The range is, according to eq. 2.11, always positive with values from zero to infinity. There is no discrete part in the probability distribution of range. As the deficit has the opposite sign of the minimum sum of deviations, all three variables (surplus, deficit and range) have values only between zero and infinity.

In the case of a symmetrical distribution of X, the distributions of S_n^+ and $-S_n^-$ are identical. The standardized variable, x, used in this study is $x = (X - \mu)/\sigma$. For an asymmetrical distribution of $f(x)$ two integrals are useful, namely

$$\int_o^{\infty} f(x)\, dx = P; \text{ and } \int_{-\infty}^{o} f(x)\, dx = Q, \text{ with } P + Q = 1 \quad 4.15$$

4. Determination of properties of surplus, deficit and range empirically from historic data. The empirical approach may produce sufficient samples of surplus, deficit and range. This approach is used when the time series of the variable X is long and n is small, thus producing a large $m = N/n$ ratio. The series is divided into m sub-series of size n. Each sub-series gives one value of a statistic, so that m sub-series gives m values of surplus, deficit and range. From these three new samples of size m, the probability distributions and time dependence for surplus, deficit and range may be empirically determined.

As the sample size m decreases by an increase in n, the smoothness of the properties determined for the surplus, deficit and range decreases with an increase of n. This decrease in smoothness

with n is the main disadvantage of empirical determination of the properties of surplus, deficit and range. A second disadvantage is the large sampling errors which are inherent to any small sample of size m.

5. Determination of properties of surplus, deficit and range by the data generation method. Empirical relationships may be used, or functions may be fitted to empirical data when a sample is characterized by its probability distribution and its time dependence. For these empirical relationships or for the mathematical expressions, the data generation method (Monte Carlo method) may be used to obtain a large sample. Samples of size $m = N/n$ may be generated as large as it is either necessary or economically feasible. Techniques (described under 4 in this chapter) are then applied to obtain distributions and time dependence of surplus, deficit and range.

6. Determination of the properties of surplus, deficit and range by the analytical method. Mathematical functions may always be fitted to empirical frequency distributions and to time dependence models, which are empirically determined. In the case of a stationary time series it is assumed that these two mathematical functions will approximate the population probability distribution and time dependence. This assumption provides a base for an analytical approach for the determination of exact or approximate probability distributions and time dependence models of surplus, deficit and range. Two cases are appropriate for consideration: (1) Exact mathematical expressions for the probability distributions and time dependence models of surplus, deficit and range may be analytically derived from the properties of X. This case is limited only to small values of n. A numerical finite difference method of integration, usually orientated to a digital computer, may be used to solve difficulties in integrating the exact mathematical equations in closed forms; (2) Statistical parameters of the probability distributions and time dependence models of surplus, deficit and range may be determined in an exact, in an approximate or in an asymptotic form as related to the parameters a_i's and b_j's of X, the base parameter X_o and the values of the variate n. In this case, a fit of approximate functions for probability distributions by using moments or statistical parameters and for time dependence by using serial correlation coefficients yields the approximate properties of surplus, deficit and range.

7. Comparison of the above three methods. An example of the series of $N = 150$ for the annual flow of a large river is used to illustrate these three methods. Two values: $n = 3$, and $n = 10$, and $X_o = \overline{X}$, are used for empirical and data generation methods. The derived samples are $m = 50$ and $m = 15$ long. The values of all three statistics (S_n^+, S_n^- and R_n) are determined respectively for $n = 3$ and $n = 10$. The analytical method is used only for $n = 3$.

The data generation method is applied to develop properties of surplus, deficit and range for normal or some non-normal but known probability distribution functions. This method is also applied to known time dependence processes. The analytical method is demonstrated by two alternatives: (a) exact distributions; and, (b) moment derivations.

The following three chapters (V, VI and VII) discuss these three methods and make detailed method comparisons. However, the information generated by each method produces a problem that warrants a brief discussion.

All three methods produce the same information if each is properly applied. In hydrology there is a contemporary trend to use the data generation method extensively. It should be noted that this method cannot produce more information than what the sample contains. When generating very large samples from data of small samples or from functions fitted to data of small samples, no new information can be obtained beyond that contained in the small sample. The same is true for the analytical method when it is applied to distribution functions and time dependence models which are derived from the available sample.

The data generation method has the following properties: (1) If the generated sample is large, the statistical characteristics of the sample converge to statistical characteristics of the small sample from which the large sample is generated. However, probability distributions and time dependence models of the large generated sample are smoother than those of the original small sample. Smoothness does not imply that the information is any better than that derived from the small sample. (2) Data generation method may be used in several problems when the mathematical equations cannot be solved in closed form. Usually, the selection is between the application of approximations in solving equations, and the use of the data generation method.

8. Systematization of variables in the analysis of surplus, deficit and range. The fitting of probability functions to empirical frequency distribution curves of hydrologic variables is practically limited to a small number of theoretical functions. These functions are: normal (Gaussian), log-normal (Galton), extreme values functions and Gamma functions (Pearson Type III included) for continuous variables, and Binomial and Poisson functions for discrete variables.

The above theoretical distributions of independent continuous variables are described: (a) by no parameters (standard normal function); (b) by one parameter (normal function with mean unity, Gamma function with one parameter); (c) by two parameters (general normal function; log-normal function with lower boundary zero; extreme values functions with lower boundary zero; Gamma function with two parameters); and, (d) by three parameters (log-normal function and extreme values function with lower boundary different from zero; Gamma with three parameters or Pearson Type III function).

Apart from the probability distribution function, a hydrologic variable and its stationary time series are characterized either as independent or dependent. This dependence is expressed by various mathematical models. The simplest dependence models in hydrology are: moving average schemes (general Markov chains); autoregressive models (Markov linear models); a combination of harmonic movement (daily or seasonal cyclic fluctuations) and the above models of moving average or autoregressive schemes, and similar.

The analysis of surplus, deficit and range is made in this paper by determining probability distributions of S_n^+, S_n^- and R_n for various values of n and X_o when the probability function and dependence model of a variable are given. The simple independent variable, symmetrically distributed in the form of independent standard normal variable, is first studied. Then, the effect of time series dependence on these three derived variables is studied for a simple mathematical model of dependence. Then, variables of various degrees of asymmetry are investigated for the influence of skewness on the properties of surplus, deficit and range.

CHAPTER V

EMPIRICAL APPROACH FOR DETERMINATION OF SURPLUS, DEFICIT AND RANGE

1. Example. In this study the annual flow series of the Rhine River at Basle, Switzerland, is used as an example to demonstrate the empirical method of obtaining the properties of the following variables: surplus, deficit, range, adjusted surplus, adjusted deficit and adjusted range. The series of annual flow is 150 years long (1808-1957), and the water year November 1 through October 31 is used for computation of annual flows.

2. Determination of new samples. Figure 5.1 gives two time series: (a) Annual flows expressed in modular coefficients $K_i = V_i/\overline{V}$, as shown in the upper graph, with K_i = modular coefficient, V_i = annual flow of i-th year, and $\overline{V}$ = average annual flow; and (b) Sums S_n of deviations $\Delta K_i = K_i - 1$, or deviations of K_i from their average value $\overline{K} = 1$, as shown in the lower graph. This lower graph of S_n, represented in a much larger scale, was used to empirically determine the surplus, deficit and range for two values of n (n = 3 and n = 10), and $S_n^+ (\overline{K}_n)$, $S_n^- (\overline{K}_n)$, and $R_n (\overline{K}_n)$ or the adjusted surplus, adjusted deficit and adjusted range. In this computation $\overline{K}_n$ was considered as being the sampling statistic or the mean for subsequent non-overlapping subseries of size n, and for two values of n (n = 3 and n = 10).

The sequence S_n in the lower graph (fig. 5.1) was divided in 50 non-overlapping subseries each 3 years long, and 15 non-overlapping subseries each 10 years long. For each subseries the values of S_n^+, S_n^- and R_n are determined. This gave three corresponding samples with size m = 50 or m = 15, respectively for n = 3 and n = 10. Similarly, the values $\overline{K}_n$ for subsequent non-overlapping subseries have been determined graphically. Then, for each subseries, $S_n^+ (\overline{K}_n)$, $S_n^- (\overline{K}_n)$ and $R_n (\overline{K}_n)$, are determined respectively for n = 3 with m = 50 and n = 10 with m = 15. Six new samples for surplus, deficit and range are thus obtained for each of the two values of n.

3. Distributions of surplus, deficit and range. Figure 5.2, left graph, gives the frequency density of surplus, S_3^+. The right graph gives the cumulative frequency (distribution) of surplus. Line (1) in both graphs represents the frequency density and distribution of surplus, respectively, as determined by the empirical method. (Lines 2 and 3 represent the same distributions as line (1), only determined by the data generation and analytical methods, respectively, as it will be described in the following two chapters). Figure 5.3 represents the same properties as fig. 5.2 except that fig. 5.3 refers to the deficit, S_3^-. Figure 5.4 refers to the range, R_3, and is analogeous to fig. 5.2.

Figures 5.5 through 5.7 represent the frequency density (left graph) and distribution (right graph) of adjusted surplus, adjusted deficit and adjusted range, respectively, for n = 3 and for $\overline{K}_3$ as a sampling statistic. Line (1) gives the distributions obtained by the empirical method (line 2 is determined by the data generation method, as it will be described in Chapter VI).

Figures 5.8 through 5.13, lines (1), give the results of the empirical method in determining distributions of the six variables for n = 10. (Lines 2 give the results of the data generation method). These figures represent the frequency density (left graph) and distribution (right graph), namely: fig. 5.8 for the surplus, S_{10}^+; fig. 5.9 for the deficit, S_{10}^-; fig. 5.10 for the range, R_{10}; fig. 5.11 for the adjusted surplus, $S_{10}^+ (\overline{K}_{10})$; fig. 5.12 for the adjusted deficit $S_{10}^- (\overline{K}_{10})$; and fig. 5.13 for the adjusted range, $R_{10} (\overline{K}_{10})$.

Table 5.1 presents statistical parameters of surplus, deficit and range for n = 3 and n = 10, respectively, and of adjusted surplus, adjusted deficit and adjusted range for n = 3 and n = 10, respectively. These statistics are: mean, variance, standard deviation, coefficient of variation, skewness coefficient, excess and the first serial correlation coefficient. The values of distribution parameters in table 5.1 are compared later in this paper with the values of the same parameters determined by one or both of the following two methods: (a) From distributions of variables S_n^+, S_n^-, R_n, $S_n^+ (\overline{K}_n)$, $S_n^- (\overline{K}_n)$ and $R_n (\overline{K}_n)$, obtained by the data generation method for both n = 3 and n = 10 (as given in Chapter VI); and (b) From exact distributions of these variables, which are integrated by the finite differences method for n = 3 (as given in Chapter VII). These comparisons, both by distributions and by their statistical parameters, are intended to show the relationship of results obtained from each of the three methods: empirical, data generation and analytical. The same example of annual flow series of the Rhine River at Basle is used for each of these three methods. The respective comparisons are discussed in Chapters VI and VII.

4. Reliability of the empirical method. Surveys of figs. 5.2 through 5.13, of statistical parameters as given in table 5.1, and general principles of statistical sampling and inference, yield two bits of knowledge: (a) Reliability of this method decreases by an increase of n; and, (b) Reliability increases with an increase of sample size N.

The data generation method may have any

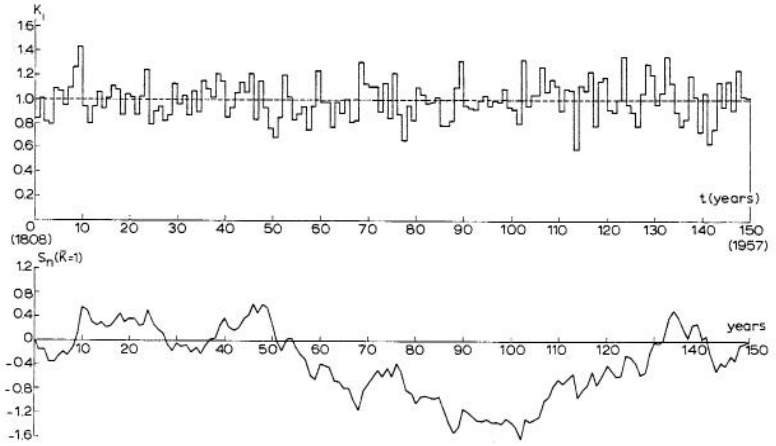


Fig. 5.1 The annual flows of the Rhine River: (a) Upper graph, the time series in modular coefficients, $K_i = V_i/\bar{V}$; (b) Sums of deviations $S_n(\bar{K}=1) = \sum_{i=1}^{n} \Delta K_i = \sum_{i=1}^{n} (K_i - 1)$.

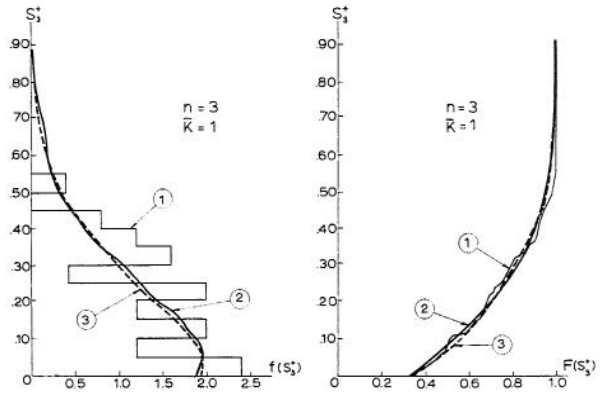


Fig. 5.2 Frequency densities (left graph) and distributions (right graph) of the surplus, S_3^+, of the annual flows of the Rhine River: (1) Determined by the empirical method; (2) Obtained by the data generation method; and (3) Obtained by the analytical method.

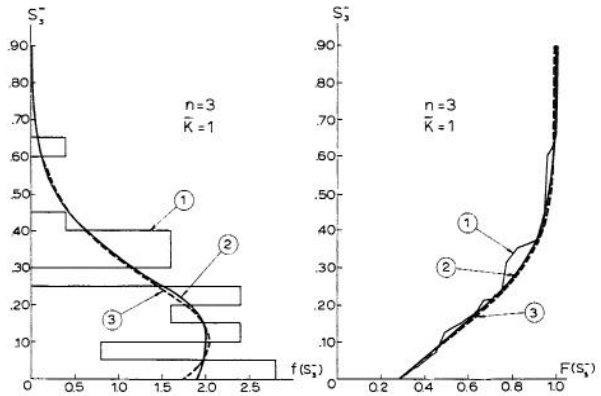


Fig. 5.3 Frequency densities (left graph) and distributions (right graph) of the deficit, S_3^-, of the annual flows of the Rhine River: (1) Determined by the empirical method; (2) Obtained by the data generation method; and, (3) Obtained by the analytical method.

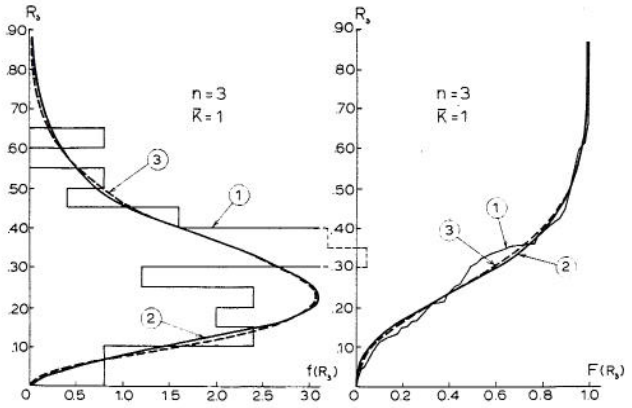


Fig. 5.4 Frequency densities (left graph) and distributions (right graph) of the range, R_3, of the annual flows of the Rhine River: (1) Determined by the empirical method; (2) Obtained by the data generation method; and, (3) Obtained by the analytical method.

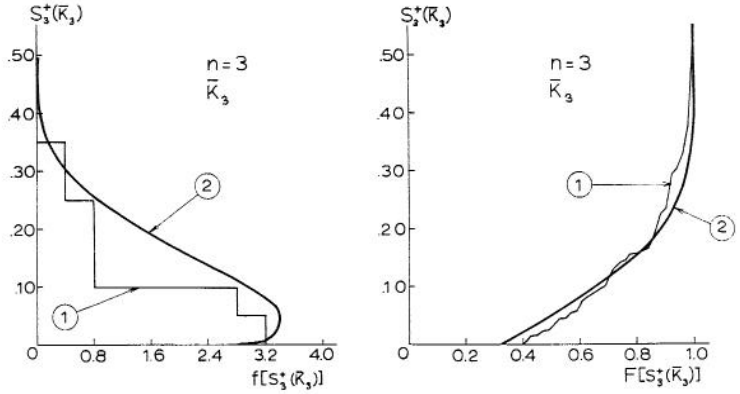


Fig. 5.5 Frequency densities (left graph) and distributions (right graph) of the adjusted surplus, $S_3^+(\bar{K}_3)$, of the annual flows of the Rhine River: (1) Determined by the empirical method; and, (2) Obtained by the data generation method.

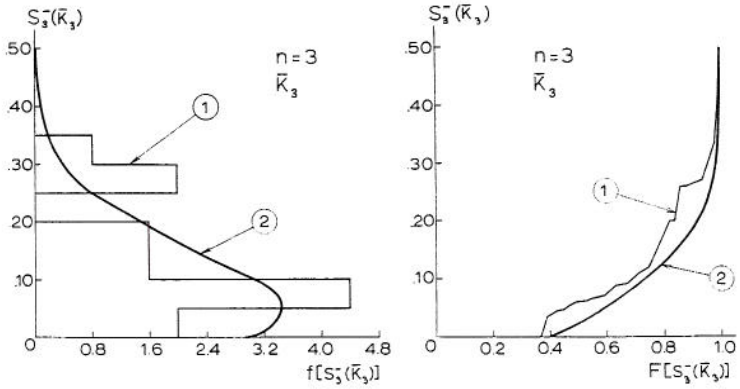


Fig. 5.6 Frequency densities (left graph) and distributions (right graph) of the adjusted deficit, $S_3^-(\bar{K}_3)$, of the annual flows of the Rhine River: (1) Determined by the empirical method; and, (2) Obtained by the data generation method.

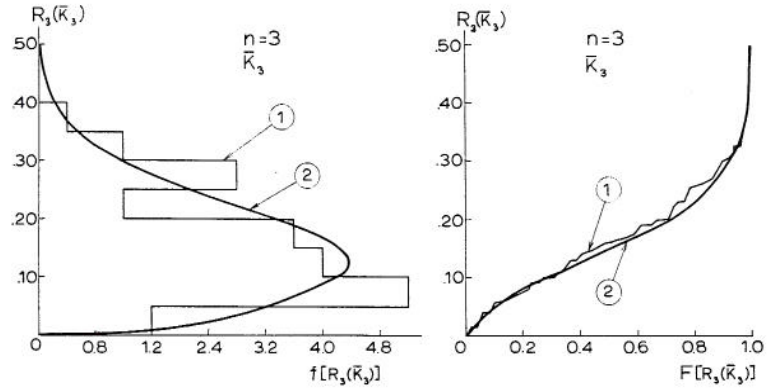


Fig. 5.7 Frequency densities (left graph) and distributions (right graph) of the adjusted range, $R_3(\bar{K}_3)$, of the annual flows of the Rhine River: (1) Determined by the empirical method; and, (2) Obtained by the data generation method.

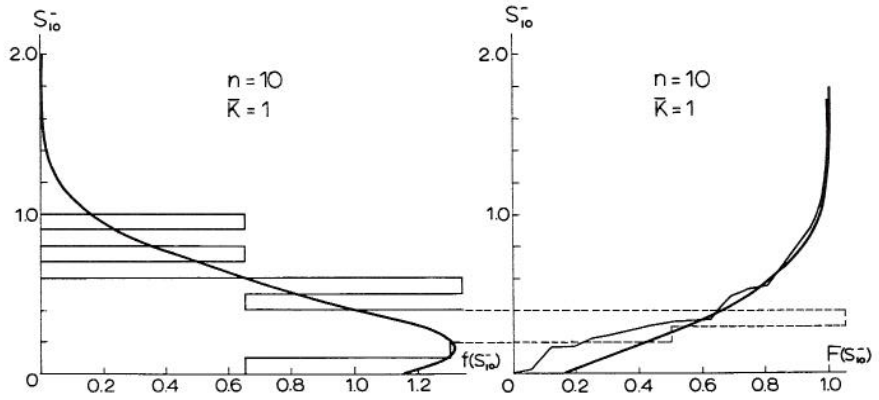


Fig. 5.8 Frequency densities (left graph) and distributions (right graph) of the surplus, n = 10 ($\bar{K}$ = 1), of the annual flows of the Rhine River: (1) Determined by the empirical method; and (2) Obtained by the data generation method.

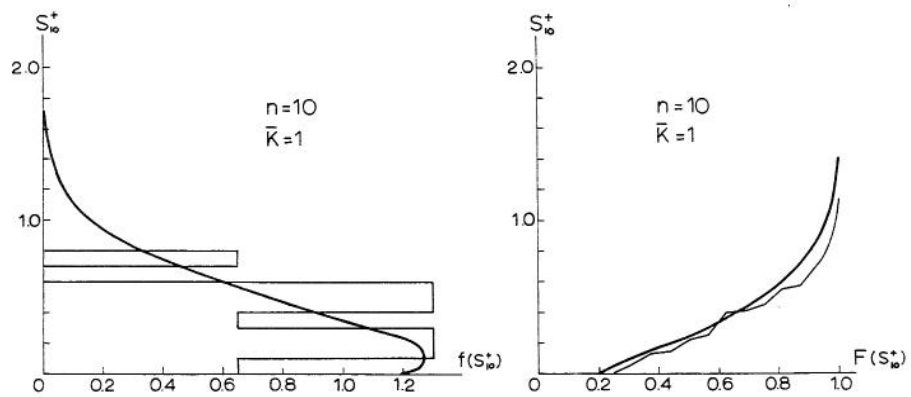


Fig. 5.9 Frequency densities (left graph) and distributions (right graph) of the deficit, n = 10 ($\bar{K}$ = 1), of the annual flows of the Rhine River: (1) Determined by the empirical method; and, (2) Obtained by the data generation method.

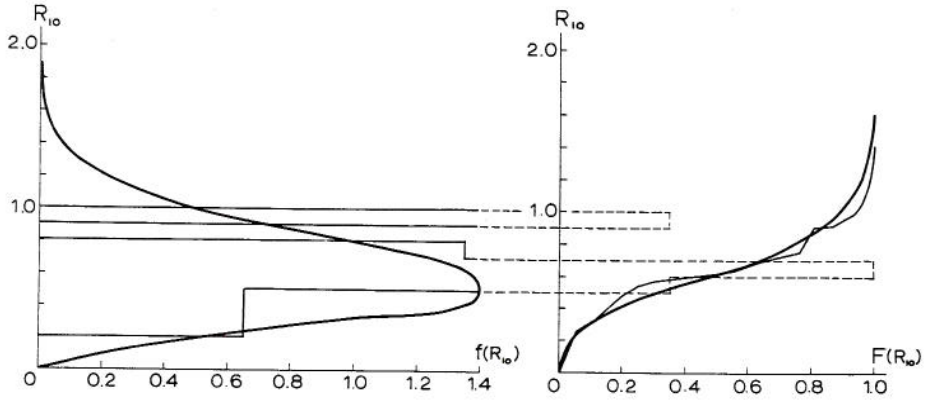


Fig. 5.10 Frequency densities (left graph) and distributions (right graph) of the range, n = 10 ($\overline{K}$ = 1), of the annual flows of the Rhine River: (1) Determined by the empirical method; and, (2) Obtained by the data generation method.

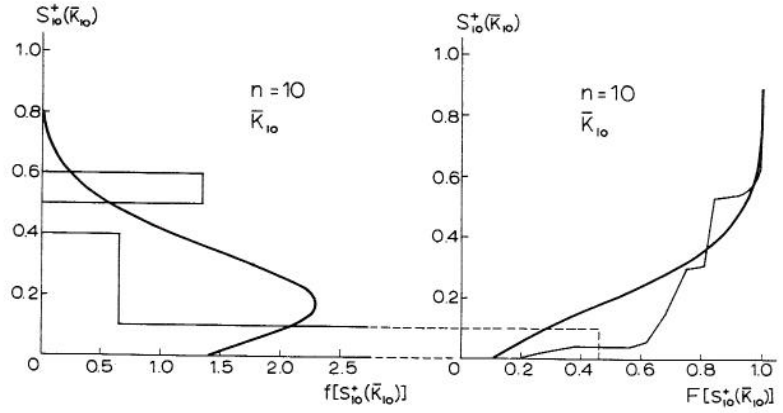


Fig. 5.11 Frequency densities (left graph) and distributions (right graph) of the adjusted surplus, n = 10 ($\overline{K}_{10}$), of the annual flows of the Rhine River: (1) Determined by the empirical method; and, (2) Obtained by the data generation method.

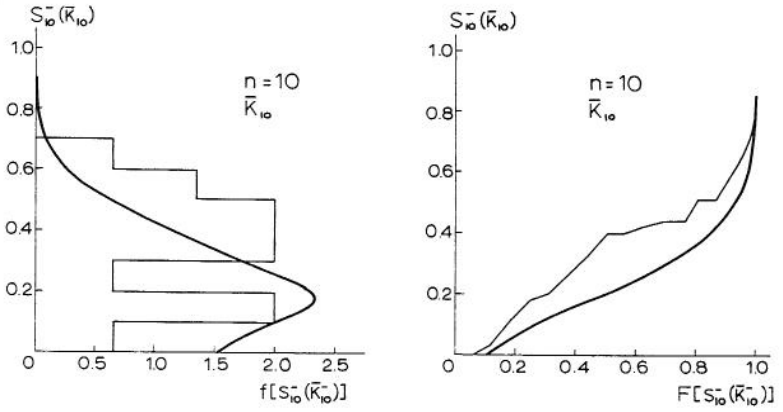


Fig. 5.12 Frequency densities (left graph) and distributions (right graph) of the adjusted deficit, n = 10 ($\overline{K}_{10}$), of the annual flows of the Rhine River: (1) Determined by the empirical method; and (2) Obtained by the data generation method.

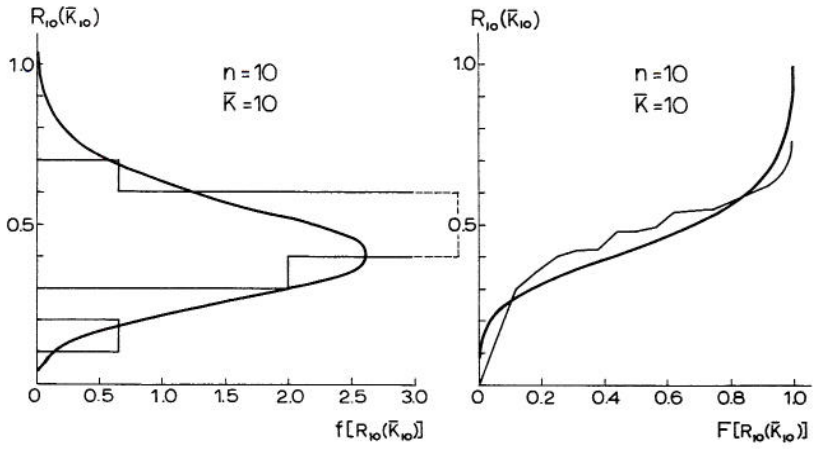


Fig. 5.13 Frequency densities (left graph) and distributions (right graph) of the adjusted range, $n = 10\,(\bar{K}_{10})$, of the annual flows of the Rhine River: (1) Determined by the empirical method; and (2) Obtained by the data generation method.

sample size N, or any value $m = N/n$. The sample size for surplus, deficit and range may be constant for any value of n (by a proportional increase of N with an increase of n). In this case the distributions and their statistical parameters for the surplus, deficit and range becomes independent of $m = N/n$. Two characteristics differentiate the data generation method from the empirical method: (a) N can be as large as it is economically feasible; and (b) m can be independent of n. However, the previously discussed problem, whether one or the other method gives more information from a given amount of sample data, should not be overlooked.

TABLE 5.1

Parameter \ Variable		S_n^+	S_n^-	R_n	$S_n^+\,(\bar{K}_n)$	$S_n^-\,(\bar{K}_n)$	$R_n\,(\bar{K}_n)$
Mean	n = 3	0.133	-0.151	0.286	0.075	-0.083	0.158
	n = 10	0.257	-0.382	0.637	0.138	-0.325	0.464
Variance	n = 3	0.022	0.026	0.065	0.009	0.009	0.008
	n = 10	0.052	0.052	0.039	0.033	0.033	0.014
Standard Deviation	n = 3	0.148	0.161	0.255	0.095	0.095	0.089
	n = 10	0.228	0.228	0.197	0.182	0.182	0.118
Coefficient of Variation	n = 3	1.113	1.066	0.892	1.267	1.144	0.563
	n = 10	0.887	0.597	0.308	1.319	0.560	0.254
Skewness Coefficient	n = 3	0.079	0.104	-0.018	0.150	0.107	-0.506
	n = 10	0.048	0.086	-0.002	0.125	-0.003	-0.092
Excess	n = 3	0.230	0.348	0.072	0.459	0.302	0.241
	n = 10	0.196	0.321	0.281	0.260	0.328	0.403
First Serial Coefficient	n = 3	0.179	0.058	-0.031	-0.085	-0.129	-0.074
	n = 10	-0.180	0.108	-0.143	-0.149	-0.218	-0.027

CHAPTER VI

DATA GENERATION METHOD FOR DETERMINATION OF SURPLUS, DEFICIT AND RANGE

1. Definition of method. The data generation method is defined as the simulation of a large sample either from data of a small sample, or from inferred characteristics of a population. These latter are usually defined by the distribution function and the mathematical dependence model of a stationary time series. Discrete series are generated by computing the independent random numbers with a given basic distribution function, and by further transformations of these numbers. Random numbers of any other distribution function with time series either independent or dependent are normally obtained by simulating them on a digital computer. The continuous series is usually generated by an analog "noise generator."

The main property of the data generation method is the absence of any limitation in the generated sample size. Size is limited by either one of the following two criteria: (a) desired accuracy of final results; and, (b) economics of generating and processing a large sample.

Independent random numbers which are readily available usually have uniform distributions. A sequence of independent random numbers with the normal (Gaussian) distribution is obtained by applying the central limit theorem for a sum of a sufficient number of uniformly distributed random variables. By further transformation, the random numbers with normal distribution may be transformed to random numbers with skewed distributions (log-normal, gamma, etc.). Dependent random numbers of various dependence models are obtained by applying these models to the series of independent random numbers.

2. Generation of large samples from empirical small samples. A small sample of a stationary series is characterized by its distribution and its time dependence. Distribution can be represented as empirical in the form of a table or graph. However, dependence models are usually described either by parameter or by equations. In this latter case, the series analysis is divided into deterministic components (trends, jumps and cycles with harmonics) and stochastic components. Dependence in series of the latter components is determined either by empirical relationships in the form of correlograms, by fitting mathematical functions to correlograms, or by the mathematical model of dependence generating process.

If random numbers of an independent variable t have a uniform density function with boundaries $t = o$ and $t = 1$, then

$$f(t) = 1 \quad \text{for} \quad o \leq t \leq 1$$
$$\text{and} \quad f(t) = o \quad \text{for} \quad t \leq o \text{ and } t > 1. \qquad 6.1$$

The transformation $t = F(X)$ makes possible the generation of random numbers of a variable with any distribution $F(X)$. Random numbers of distribution of eq. 6.1 are automatically transformed to random numbers of distribution of X when: (a) the empirical distribution function $F(X)$ is given for the range X_{min} to X_{max} as $F(X_{min}) = 0$ and $F(X_{max}) = 1$; (b) the table of $t = F(X)$ versus X is used; and, (c) the proper interpolation procedure between the discrete values of X and $F(X)$ is used in determining the X-value which corresponds to a given t-value. If the variable X is dependent in sequence, another transformation must be used.

The other approach in using the data generation method is to produce a large sample from a small sample of an independent variable X by: (a) inferring a theoretical distribution function to the empirical distribution $F(X)$; and (b) by generating a large sample of random numbers by the procedure available for that type of theoretical function.

3. Example of large sample generation. In this study the latter of the above two approaches is used when generating a large annual flow sample of the Rhine River at Basle from the available data. A log-normal distribution function is fitted to the annual flow distribution of the Rhine River with sample size $N = 150$. The log-normal function is then used to generate a large sample of $m = N/n = 10,000$ random numbers. As the serial correlation coefficients of annual flows of that river for $N = 150$ show no significant difference from an independent time series, only independent random numbers are generated. Logarithms of the Rhine River annual flows are approximately normally distributed.

The modular coefficients K_i of the Rhine River annual flows have a mean of $\overline{K} = 1$ and a standard deviation of $s_k = C_v = 0.159$. As $\ln K_i$ is normally distributed with the mean μ_n and variance σ_n, they can be obtained from $\overline{K} = 1$ and $s_k = 0.159$ as

$$\mu_n = \ln \frac{\overline{K}^2}{\sqrt{s_k^2 + \overline{K}^2}} = \ln \frac{1}{\sqrt{1 + s_k^2}} \qquad 6.2$$

and

$$\sigma_n^2 = \ln \left[1 + \frac{s_k^2}{\overline{K}^2}\right] = \ln \left[1 + s_k^2\right] \qquad 6.3$$

These two equations give $\sigma_n^2 = \ln(1 + s_k^2) = 0.02517$, or $\mu_n = -\frac{1}{2}\sigma_n^2 = -0.0126$. The variable $\ln K = 0.1582t - 0.0126$, with t the standard normal variable, so that

$$K = e^{0.1582\,t - 0.0126} \qquad 6.4$$

is the transformation equation to obtain K - variable from normal independent random numbers, t.

The log-normal independent variable, K, of the Rhine River's annual flows is generated by using the random numbers of a normal standard independent variable, t, which are transformed by eq. 6.4.

The large sample is generated with 30,000 independent random numbers in order to obtain 10,000 independent and non-overlapping subseries of n = 3. A large sample, with 100,000 numbers is also generated in order to obtain 10,000 independent and non-overlapping subseries of n = 10. The surplus, deficit, range, adjusted surplus, adjusted deficit and adjusted range for both n = 3 and n = 10 are determined from these subseries. These subseries are also used to compute distributions and statistical parameters of distributions. Both frequency density and distributions are plotted in figs. 5.2 through 5.13 as lines (2). Table 6.1 gives statistical parameters which correspond case by case to table 5.1 except the first serial correlation coefficient which was not computed in the data generation method approach. Table 6.1 illustrates the statistical parameters of frequency distributions of the following variables: surplus, deficit, range, adjusted surplus, adjusted deficit and adjusted range for n = 3 and n = 10.

TABLE 6.1

Variable / Parameter		S_n^+	S_n^-	R_n	$S_n^+ (\bar{K}_n)$	$S_n^- (\bar{K}_n)$	$R_n (\bar{K}_n)$
Mean	n = 3	0.147	-0.144	0.291	0.078	-0.078	0.155
	n = 10	0.314	-0.325	0.639	0.219	-0.223	0.442
Variance	n = 3	0.031	0.023	0.22	0.008	0.008	0.007
	n = 10	0.097	0.086	0.062	0.027	0.028	0.020
Standard Deviation	n = 3	0.175	0.153	0.147	0.088	0.087	0.085
	n = 10	0.312	0.293	0.250	0.165	0.166	0.143
Coefficient of Variation	n = 3	1.194	1.060	0.508	1.138	1.126	0.544
	n = 10	0.993	0.902	0.391	0.753	0.744	0.323
Skewness Coefficient	n = 3	1.454	1.090	1.063	1.230	1.174	0.775
	n = 10	1.166	0.999	0.960	0.742	0.712	0.679
Excess	n = 3	2.189	0.825	0.147	1.409	0.988	0.720
	n = 10	1.214	0.772	1.194	0.507	0.357	0.673

4. Comparison of the data generation method with the empirical method. Figures 5.2 through 5.13 illustrate that distributions determined by the data generation method are much smoother than distributions obtained by the empirical method. This is especially true when n = 10. Table 6.2 gives differences of statistical parameters of frequency distributions for the empirical method and the data generation method. Values in table 6.2 are the differences between the corresponding values in tables 5.1 and 6.1. These differences increase with an increase of the order of moments used in computing various statistical parameters. The greatest differences are for the skewness coefficients and the excess. However, it should be stressed that the absence of extreme large values in the frequency distributions of these six variables for the empirical method accounts for the large differences in the skewness coefficient and the excess. The smoothness of distributions obtained by the data generation method is well illustrated in figs. 5.2 through 5.13, and it is an asset of this method.

5. Generation of large samples from theoretical distribution functions and mathematical models of time dependence. Any generation of random numbers should be subjected to appropriate tests of samples generated. The two necessary tests are: (1) That the sample distribution is insignificantly different on a prescribed probability level from the distribution underlying the generation process (which is either an empirical distribution or a theoretical distribution function); and (2) That the time dependence of random numbers in the generated sample do not depart significantly on the prescribed probability level from the dependence underlying the generating process. Models for the generation of random numbers of several types of variables will be discussed in portions of this text. In these discussions the variables will be described by theoretical distribution functions which are either time independent or time dependent.

(1) Independent normal variables. Digital computer programs are already available for the random numbers generation of independent standard normal variable, t, with expected mean zero, expected standard deviation unity, and expected first serial correlation coefficient zero, (0, 1, 0). Tests of normality and independence are easy to perform.

To obtain the random numbers of an independent normal variable, X, with $\bar{X}$ different from zero, and s different from unity, the transformation to be used is

$$X_i = s\, t_i + \bar{X} \qquad 6.5$$

where t_i represents the random independent numbers.

(2) Dependent normal variables. Several

TABLE 6.2

Parameter \ Variable		S_n^+	S_n^-	R_n	$S_n^+(\bar{K}_n)$	$S_n^-(\bar{K}_n)$	$R_n(\bar{K}_n)$
Mean	n = 3	0.147	-0.144	0.291	0.078	-0.078	0.155
	n = 10	0.314	-0.325	0.639	0.219	-0.223	0.442
Variance	n = 3	0.031	0.023	0.022	0.008	0.008	0.007
	n = 10	0.097	0.086	0.062	0.027	0.028	0.020
Standard Deviation	n = 3	0.175	0.153	0.147	0.088	0.087	0.085
	n = 10	0.312	0.293	0.250	0.165	0.166	0.143
Coefficient of Variation	n = 3	1.194	1.060	0.508	1.138	1.126	0.544
	n = 10	0.993	0.902	0.391	0.753	0.744	0.323
Skewness Coefficient	n = 3	1.454	1.090	1.063	1.230	1.174	0.775
	n = 10	1.166	0.999	0.960	0.742	0.712	0.679
Excess	n = 3	2.189	0.825	0.147	1.409	0.988	0.720
	n = 10	1.214	0.772	1.194	0.507	0.357	0.673

mathematical dependence models of stationary time series are available for dependent normal variables. The selection of those models depends on the character of hydrologic process. The first order linear Markov model will often fit the dependence in time series when the change in water storage carryover is responsible for time dependence in river flows. For this reason the Markov model will be used as an example in this chapter. Moving average schemes of various types may also be used as well as the second or higher order linear Markov models. The first order linear Markov model is currently used in hydrologic sample generations in the form

$$x_i = \rho\, x_{i-1} + \sqrt{1 - \rho^2}\quad \epsilon_i \qquad 6.6$$

where ϵ_i are random numbers of an independent standard normal variable with $E(\epsilon_i) = 0$, var $\epsilon = 1$; ρ = population first autocorrelation coefficient; and x_i generated new numbers of normal standard [$E(x_i) = 0$, var $x = 1$] but dependent x-variable. Multiplication of ϵ_i variable by $\sqrt{1 - \rho^2}$ is necessary in order to obtain the x_i variable with variance unity. For the correlogram of generated series $E(r_k) = \rho_k$, with $\rho_k = \rho^k$. This is the correlogram of the first order linear Markov model. However, the model

$$x_i = \rho\, x_{i-1} + \epsilon_i \qquad 6.7$$

is also very often used. In this case, the variance of x_i is $1/(1 - \rho^2)$ if var $\epsilon_i = 1$. For all values of ρ it is greater than unity. The differences in variances of x_i between eq. 6.6 and eq. 6.7 should be taken into account whenever the two models are used interchangeably.

A correlogram $r_k = r^k$ of a generated large sample should not depart significantly from the population correlogram $\rho_k = \rho^k$ on a given probability level. This relationship may be tested by: (1) performing corresponding chi-square test, or (2) by ascertaining if $\Delta r_k = \rho_k - r_k$ differs significantly from zero on the same prescribed probability level.

(3) Independent log-normal variables. If the random numbers are needed for a log-normal independent variable, U, the following transformation can be used. For μ = mean of U and σ^2 the variance of U, the mean and variance of ln U are given as

$$\mu_u = \ln \frac{\mu^2}{(\sigma^2 + \mu^2)^{1/2}} \qquad 6.8$$

and

$$\sigma_u^2 = \ln\left(1 + \frac{\sigma^2}{\mu^2}\right) \qquad 6.9$$

with ln U normally distributed with mean μ_u and variance σ_u^2. By using random numbers of standard normal independent variable, ϵ, then $\ln U = \sigma_u \epsilon + \mu_u$, and the transformation

$$U = e^{\sigma_u \epsilon + \mu_u} \qquad 6.10$$

gives the random numbers of U, with μ_u and σ_u given by eqs. 6.8 and 6.9, respectively.

(4) Dependent log-normal variables. The first order Markov linear model of eq. 6.6 is based on the principle that the sum of two normal variables, each multiplied by a constant (and constants related), produces a standard normal dependent variable.

This principle cannot be applied to log-normal variables because the sum of two log-normal variables is not an independent or dependent log-normal variable. However, the product of two log-normal variables is a log-normal variable. In this case,

$$U_i = e^{\rho \ln U_{i-1}} \; e^{\sqrt{1-\rho^2}\;\epsilon_i} \qquad 6.11$$

If the variable x_i of eq. 6.6 has mean zero, standard deviation unity, and first serial correlation coefficient ρ, (0, 1, ρ), with $\rho_k = \rho^k$ and is transformed by

$$U = e^x \qquad 6.12$$

then $E(U) = e^{1/2}$, var $U = e(e-1)$, and the $\rho_k(U)$ becomes

$$\rho_k(U) = \frac{\text{cov } U_i\, U_{i+k}}{\text{var } U_i} = \frac{E(U_i\, U_{i+k}) - e}{e(e-1)} \qquad 6.13$$

with

$$E(U_i\, U_{i+k}) = E(e^{x_i + x_{i+k}})$$

The new variable $(x_i + x_{i+k})$ has mean zero and var $(x_i + x_{i+k}) = 2(1 + \rho^k)$ because cov $x_i\, x_{i+k} = \rho^k$. The application of eq. 6.3 gives $E(e^{x_i + x_{i+k}}) = \exp[\frac{1}{2} \text{ var } (x_i + x_{i+k})] = \exp[1 + \rho^k]$. Equation 6.13 then becomes

$$\rho_k(U) = \frac{e^{\rho^k} - 1}{e - 1} \qquad 6.14$$

This model is different from $\rho_k(x) = \rho^k$. Therefore, the transformation of the variable x of eq. 6.6 by eq. 6.12 does not produce a log-normal variable with a first order linear Markov model. This transformation does produce another sequential model of the die-away correlogram type. For various ρ, the differences of values of $\rho_k(U)$ of eq. 6.14 and $\rho_k = \rho^k$ are given in fig. 6.2 as a function of k. The model of eq. 6.14 is represented as $\rho_k = f(k)$ for various values of ρ in fig. 6.1. For comparison, the model $\rho_k = \rho^k$ is also represented in fig. 6.1 as dashed lines.

(5) Independent gamma variables. If x_j designates independent standard normal variables then the transformation

$$U = \frac{1}{2} \sum_{j=1}^{m} x_j^2 \qquad 6.15$$

gives the chi-square distribution of the variable 2U with m degrees of freedom, or Gamma distributed U-variable with the parameter m/2

$$f(U) = \frac{1}{\Gamma(\frac{m}{2})}\, U^{\frac{m}{2} - 1}\, e^{-U} \qquad 6.16$$

with mean m/2, variance m/2, and skewness coefficient $(8/m)^{1/2}$. By using integers for m, the only values of parameters obtained are 0.5, 1.0, 1.5, 2.0, ... , m/2, with m = 1, 2, ... However, these values of m/2 are sufficient to study the change in surplus, deficit and range with the change of distribution skewness. The changes inbetween the values given by $C_s = (8/m)^{1/2}$ for m = 1, 2, 3, ... may be easily interpolated. Any problem requiring a C_s-value which lies between two discrete C_s-values for two given successive m-integers may be solved for both m and m + 1, by the data generation method. Statistics are determined for both C_s-values and an interpolation gives the corresponding statistical parameters for the desired C_s-value.

(6) Dependent gamma variables. In this study the gamma distribution and the first order linear Markov model are used to generate a variable which has a skewed distribution, which is dependent in time. This concept serves the purpose of investigating the simultaneous effect of skewness and dependence on the properties of surplus, deficit and range.

Let x and ϵ be two normal standard variables. Take $x_1 = \epsilon_1$ and use the generating process

$$x_i = \sqrt{\rho}\; x_{i-1} + \sqrt{1-\rho}\; \epsilon_i \qquad 6.17$$

The dependent variable x is obtained by using the sequence $\epsilon_1, \epsilon_2, \ldots$ of the variable ϵ with a constant parameter ρ.

Using the same procedure, m variables x_j may be generated from variables ϵ_j, with j = 1. 2. 3. ... m. In this generating process, the sequences i = 1, 2, ... are obtained for each of m variables with the sample size as large as necessary or feasible. These sequences are denoted by x_{ij}, where i represents the position in the sequence of the variable x_j.

The transformation of eq. 6.15 is used to obtain the gamma distribution of a variable U from the normal standard distribution of m variables x_j

$$U_i = \frac{1}{2} \sum_{j=1}^{m} x_{ij}^2 .$$

The serial correlation coefficient of the lag k of the variable U is

$$\rho_k = \frac{\text{cov } U_i\, U_{i+k}}{\text{var } U_i} = \frac{2}{m}\left(E\, U_i\, U_{i+k} - \frac{m^2}{4}\right) \qquad 6.18$$

Replacing U_i and U_{i+k} by the corresponding values of eq. 6.15 then

$$E\, U_i\, U_{i+k} = \frac{1}{4} \sum_{j=1}^{m} \sum_{s=1}^{m} E\, x_{ij}^2\, x_{i+k,s}^2 =$$

$$= \frac{1}{4} \sum_{j=1}^{m} E\, x_{ij}^2\, x_{i+k,j}^2 + \frac{1}{4} \sum_{j \neq s} x_{ij}^2\, x_{i+k,s}^2 \qquad 6.19$$

As $E(x_{ij}^2\ x_{i+k,j}^2) = E x_{ij}^2\ E x_{i+k,j}^2 + 2(E x_{ij}\ x_{i+k,j})^2$, and as $E x_{ij}\ x_{i+k,j} = \rho^{k/2}$, then eq. 6.19 becomes

$$E U_i\ U_{i+k} = \frac{1}{4}\ m(1+2\rho^k) + \frac{1}{4}(m-1)\,m,$$

and

$$\rho_k = \frac{2}{m}\left[\frac{1}{4}\,m(1+2\rho^k) + \frac{1}{4}(m-1)\,m - \frac{m^2}{4}\right] = \rho^k. \qquad 6.20$$

When eq. 6.17 is used in generating x_j variables, it has been proven that the variable U_j is dependent and the mathematical process is Markov first order linear scheme.

The desired degrees of skewness and dependence of the model of eq. 6.17 is obtained by changing the number m of variables x_j, thus varying the skewness coefficient $(8/m)^{1/2}$, and by changing the parameter ρ. Therefore, eqs. 6.17 and 6.15 may be used with changing parameters m and ρ for the generation of sequences with various degrees of skewness and dependence, respectively.

6. Examples. Chapters VIII and IX offer examples of large sample generation for variables whose distribution functions and mathematical expressions for time dependence of stationary series are given. The normal standard independent or dependent variables, and the gamma independent variables are used in these examples. The example of the log-normal independent variable has been shown in this chapter for the Rhine River's annual flows.

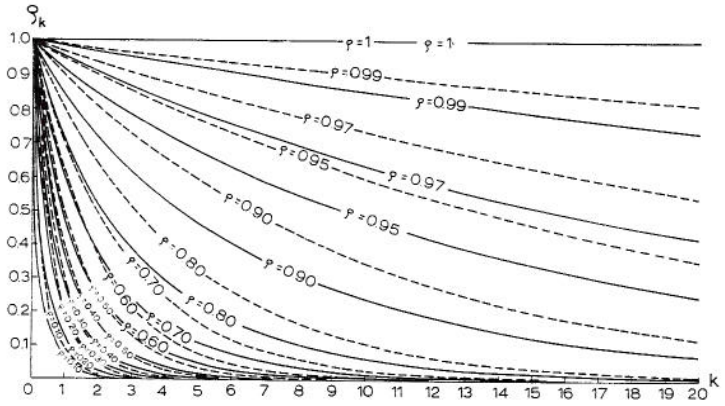


Fig. 6.1 The correlograms of two dependence models for various values of the first autocorrelation coefficient, ρ: (1) $\rho_k = \rho^k$ (first order linear Markov dependence model), dashed lines; (2) $\rho_n = (e^{\rho^k} - 1)/(e-1)$, of eq. 6.14, solid lines.

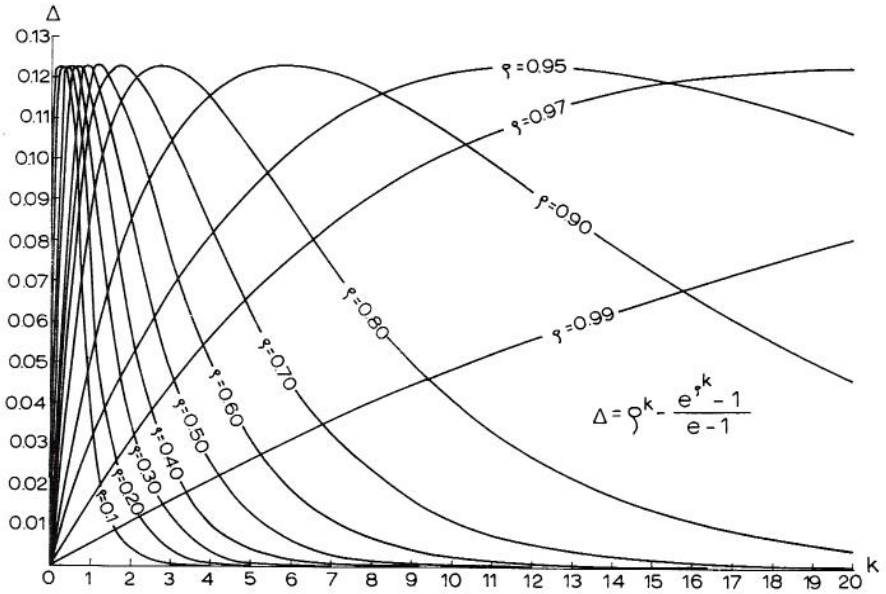


Fig. 6.2 Differences $\Delta = \rho^k - (e^{\rho^k} - 1)/(e-1)$ of the two models of fig. 6.1 as functions of ρ and n.

CHAPTER VII

EXACT DISTRIBUTIONS OF SURPLUS, DEFICIT AND RANGE DETERMINED ANALYTICALLY FOR AN INDEPENDENT VARIABLE

1. Types of variable distributions. It is assumed that the probability density of a standardized variable x is given either by an empirical frequency density curve, by a mathematical function that has been fitted to this empirical curve, or in general by a population probability density function f(x). However, it is assumed that the empirical frequency density curve has been locally smoothed. Whenever f(x) is used in the following equations, it should be replaced either by data of an empirical curve or by a mathematical function fitted to data which is, or assumed to be, the population density curve.

In this study, distributions of surplus, deficit and range for a given n and a given x_o will be expressed in general terms by using the function f(x). Integration of equations of their exact distributions is carried out for the example given by the finite differences numerical method. In this case, integrals are replaced by summations. When using the empirical frequency density curve f(x), it is represented by a table for all computational purposes. This table presents data as frequency (or probability) densities versus x-values spaced at selected differences Δx.

TABLE 7.1

x_i	$f(x_i)$	x_i	$f(x_i)$
-3.00	0.0000	0.10	0.3804
-2.90	0.0002	0.20	0.3669
-2.80	0.0010	0.30	0.3503
-2.70	0.0020	0.40	0.3328
-2.60	0.0040	0.50	0.3132
-2.50	0.0060	0.60	0.2921
-2.40	0.0095	0.70	0.2710
-2.30	0.0146	0.80	0.2499
-2.20	0.0221	0.90	0.2294
-2.10	0.0311	1.00	0.2098
-2.00	0.0427	1.10	0.1902
-1.90	0.0567	1.20	0.1706
-1.80	0.0733	1.30	0.1516
-1.70	0.0939	1.40	0.1340
-1.60	0.1174	1.50	0.1190
-1.50	0.1420	1.60	0.1044
-1.40	0.1696	1.70	0.0923
-1.30	0.1993	1.80	0.0813
-1.20	0.2289	1.90	0.0713
-1.10	0.2580	2.00	0.0617
-1.00	0.2866	2.10	0.0537
-0.90	0.3127	2.20	0.0462
-0.80	0.3363	2.30	0.0397
-0.70	0.3563	2.40	0.0341
-0.60	0.3749	2.50	0.0291
-0.50	0.3885	2.60	0.0246
-0.40	0.3975	2.70	0.0206
-0.30	0.4020	2.80	0.0171
-0.20	0.4025	2.90	0.0141
-0.10	0.3985	3.00	0.0110
0.00	0.3910	3.10	0.0080
		3.20	0.0055
		3.30	0.0035
		3.40	0.0015
		3.50	0.0000

In this study, the analytical method of determining exact distributions is related only to surplus (S_n^+), deficit (S_n^-), and range (R_n). The exact distributions of adjusted surplus, adjusted deficit and adjusted range are not investigated in this paper.

2. Example to be used. This chapter employs the same sample as Chapters V and VI to illustrate the analytical approach in the determination of distributions of surplus, deficit and range. The independent variable is the annual flow of the Rhine River at Basle, Switzerland, with $N = 150$ years (1808 - 1957), average annual flow $\bar{Q} = 36250$ cfs and the coefficient of variation $C_v = 0.159$. Figure 7.1 gives the fitted log-normal functions to annual flows of the Rhine River. Figure 7.2 shows the log-normal probability density curve of standardized variable $x = (V_i - \bar{V})/s$ of the Rhine River's annual flows. Table 7.1 gives the values of f(x) for x at the intervals of $\Delta x = 0.10$, of the standardized variable $x = (V_i - \bar{V})/s$. This example is used in this chapter to show the analytical method of computations. To compare distributions obtained by the analytical method with distributions obtained by empirical and data generations methods, the values of S_n^+, S_n^-, and R_n must be multiplied by $C_v = 0.159$, and their densities divided by it. This factor will yield values that are comparable with those given in Chapters V and VI. This example will be shown after the theoretical analysis of exact distributions is completed.

3. The approach to analytical determination of exact distributions. The analytical determination of exact distributions of surplus, deficit and range may be approached by either of the following methods: (a) By analyzing all possible combinations of cases between $x_1, x_2, \ldots, x_n$ for $n = 1, 2, \ldots$; (b) By using the distributions of $x_1, x_2, \ldots, x_n$, but with changing integration regions; and (c) By using the joint distribution of sums $S_1, S_2, \ldots, S_n$, in the form of $F(S_1, S_2, \ldots, S_n)$; and defining the probability of S_n^+ so that none of the $S_1, S_2, \ldots, S_n$ variables exceeds a given S, or $S_n^+ \leq S_i$. Similarly, it can be done for S_n^- and R_n.

4. Exact distributions of surplus, deficit and range for n = 1. The basic value $x_o = \bar{x} = o$ of a standardized variable x is used in the derivation of probability densities or probability mass of surplus, deficit, and range. The surplus has the probability density

$$f_1(S_1^+) = f(x), \text{ for } x \geq 0$$

and probability mass for $S_1^+ = 0$

$$F_1(S_1^+ = 0) = Q$$

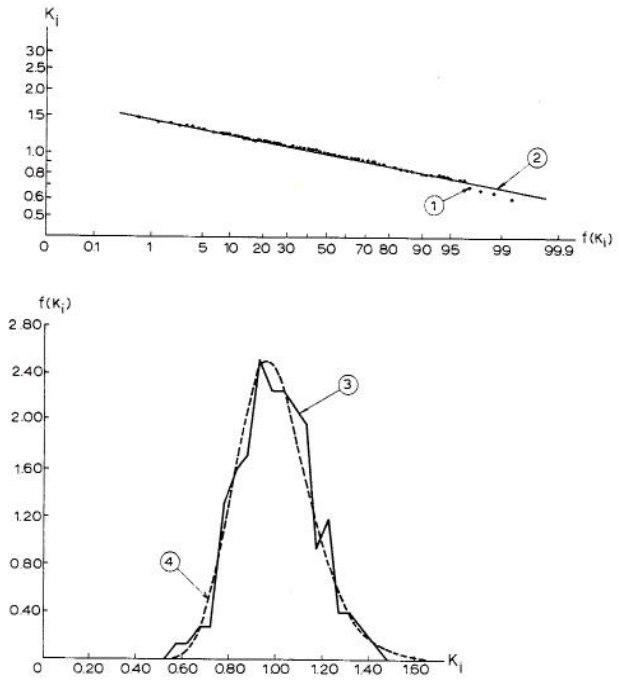


Fig. 7.1 Frequency distribution (upper graph) and frequency density curve (lower graph) of the Rhine River's annual flows at Basle, in modular coefficients, K_i: (1) Observed; (2) Fitted log-normal function; (3) Observed densities; and (4) Fitted log-normal probability density function.

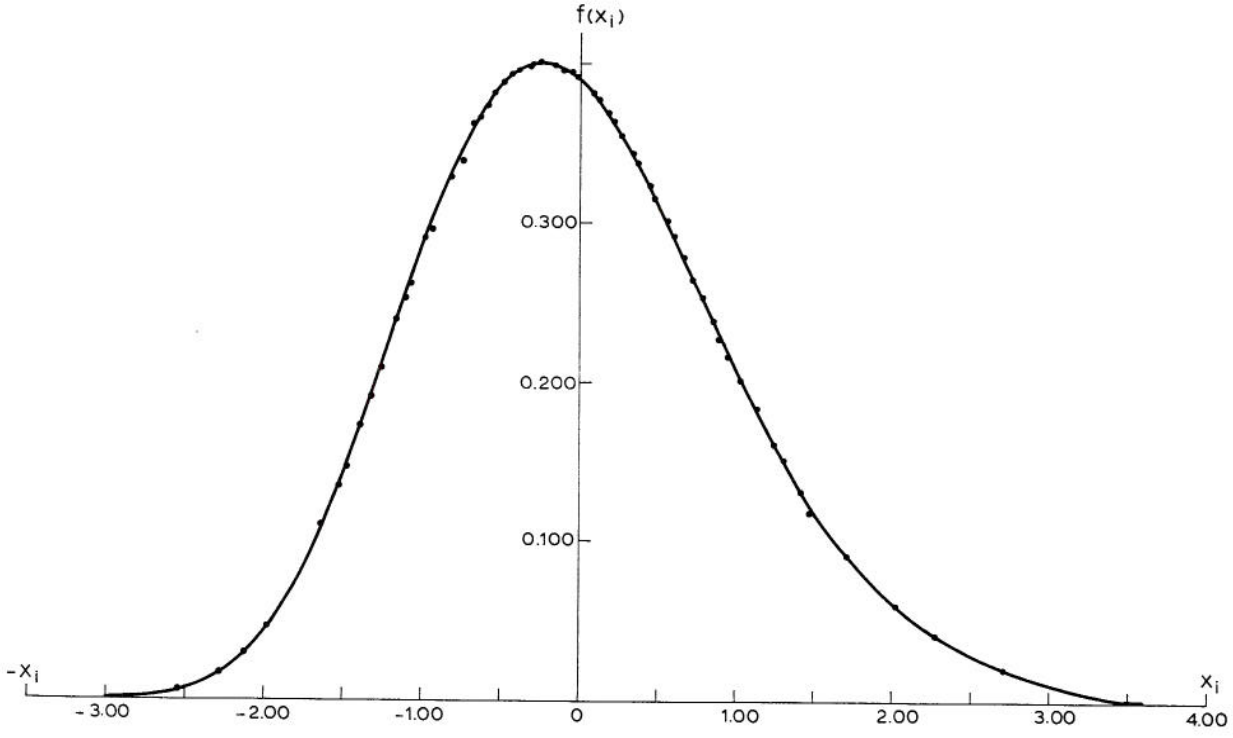


Fig. 7.2 Fitted log-normal probability density curve to standardized variable $X_i = (V_i - \overline{V})/s$ for the annual flow of the Rhine River at Basle, Switzerland (1808-1957), N = 150 years.

with Q defined by equation 4.15. Similarly,

$$f_1(S_1^-) = f(x), \text{ for } x \leq 0$$

is the probability density of S_1^-, and

$$F_1(S_1^- = 0) = P$$

the probability mass of $S_1^- = 0$, with P defined by eq. 4.15.

As $R_1 = S_1^+ - S_1^-$ the probability density of R_1 is

$$f_1(R_1) = f(x) + f(-x) = f(|x|)$$

because $R_1 = |x|$, with R_1 in the limits from zero to the maximum absolute value of x.

Distributions of S_1^+, S_1^- and R_1 are, therefore,

$$F_1(S_1^+ \leq S) = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{S} f(x)\,dx \qquad 7.1$$

$$F_1(S_1^- \leq S) = \int_{0}^{\infty} f(x)\,dx + \int_{S}^{0} f(x)\,dx \qquad 7.2$$

and

$$F_1(R_1 \leq R) = \int_{0}^{R} f(|x|)\,dx \qquad 7.3$$

Figure 7.3 shows probability densities and the probability mass of S_1^+, S_1^- and R_1 for the above example. For this example, P = 0.46764 and Q = 0.53236.

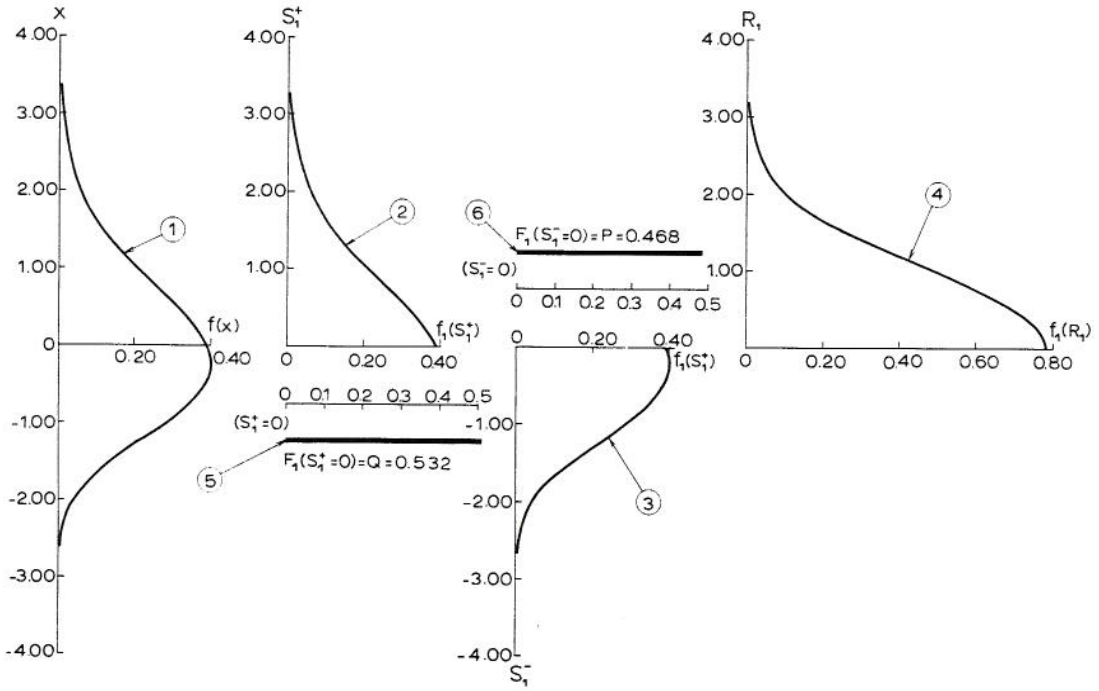
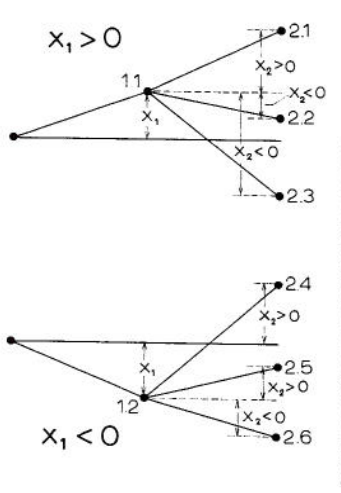


Fig. 7.3 Probability density curves of x, S_1^+, S_1^-, and R_1, determined for the standard log-normal probability density curve, f(x), of the Rhine River's annual flows: (1) Probability density of x_i; (2) Probability density of S_1^+; (3) Probability density of S_1^-; (4) Probability density of R_1; (5) Probability mass for $S_1^+ = 0$; and (6) Probability mass for $S_1^- = 0$.

5. Distributions of surplus, deficit and range for n = 2. Figure 7.4 shows six possible cases for different combinations of x_1 and x_2, where x_1 is the variable value for the first time interval, and x_2 for the second time interval. Numbers of subcases are: 1.1 and 1.2 for x_1 and, 2.1 through 2.6 for x_2. The first number in this designation refers to n-value and the second number to the subcase. Figure 7.4 also gives S_2^+, S_2^- and R_2 for each of the six subcases as expressed in x_1 and x_2.



$x_1+x_2=$ $S_1^+ + x_2$	0	$x_1+x_2=$ R_1+x_2
$x_1 = S_1^+$	0	$x_1 = R_1$
$x_1 = S_1^+$	$x_1 = x_2$	$\lvert x_2 \rvert$

S_2^+	S_2^-	R_2
$x_1 = x_2$	$x_1 = S_1^-$	x_1
0	$x_1 = S_1^-$	$\lvert x_1 \rvert = R_1$
0	$x_1+x_2=$ $S_1^- + x_2$	$\lvert x_1+x_2 \rvert =$ $R_1 + \lvert x_2 \rvert$

Fig. 7.4 Six possible cases for different combinations of x_1 and x_2 in the determination of exact distributions of S_2^+, S_2^- and R_2; $(n = 2, \bar{x} = 0)$.

Subcases 2.1 through 2.6 have the following probability densities ($f_{2.i}$) and probability mass ($F_{2.i}$) for the surplus, deficit and range

(2.1)

$$f_{2.1}(S_2^+ = S) \int_0^S f(x)\, f(S-x)\, dx$$

$$F_{2.1}(S_2^- = 0) = \int_0^\infty f(x) \left[\int_0^\infty f(y)\, dy \right] dx$$

$$f_{2.1}(R_2 = R) = \int_0^R f(x) f(R-x)\, dx$$

(2.2)

$$f_{2.2}(S_2^+ = S) = f(S) \int_{-S}^0 f(x)\, dx$$

$$F_{2.2}(S_2^- = 0) = \int_0^\infty f(x) \left[\int_{-x}^0 f(y)\, dy \right] dx$$

$$f_{2.2}(R_2 = R) = f(R) \int_{-R}^0 f(x)\, dx$$

(2.3)

$$f_{2.3}(S_2^+ = S) = f(S) \int_{-\infty}^{-S} f(x)\, dx$$

$$f_{2.3}(S_2^- = S) = \int_0^\infty f(x)\, f(S-x)\, dx$$

$$f_{2.3}(R_2 = R) = f(-R) \int_0^R f(x)\, dx$$

(2.4)

$$f_{2.4}(S_2^+ = S) = \int_{-\infty}^0 f(x)\, f(S-x)\, dx$$

$$f_{2.4}(S_2^- = S) = f(S) \int_{-S}^\infty f(x)\, dx$$

$$f_{2.4}(R_2 = R) = f(R) \int_{-R}^0 f(x)\, dx$$

(2.5)

$$F_{2.5}(S_2^+ = 0) = \int_{-\infty}^0 f(x) \left[\int_0^{-x} f(y)\, dy \right] dx$$

$$f_{2.5}(S_2^- = S) = f(S) \int_0^{-S} f(x)\, dx$$

$$f_{2.5}(R_2 = R) = f(-R) \int_0^R f(x)\, dx$$

(2.6)
$$F_{2.6}(S_2^+ = 0) = \int_{-\infty}^{0} f(x) \left[\int_{-\infty}^{0} f(y)\, dy \right] dx$$

$$f_{2.6}(S_2^- = S) = \int_{S}^{0} f(x)\ f(S - x)\, dx$$

$$f_{2.6}(R_2 = R) = \int_{-R}^{0} f(x)\, f(-R - x)\, dx$$

The subcases for S_2^+, S_2^-, and R_2 are combined in the form

$$f_2(Y) = \sum_{i=1}^{6} f_{2.i}(Y)$$

with $Y = S_2^+$, $Y = S_2^-$, and $Y = R_2$, respectively.

The probability density curve, the probability mass and the probability distributions of S_2^+, S_2^- and R_2 are

$$f_2(S_2^+ = S) = Q f(S) + \int_{-\infty}^{S} f(x)\, f(S - x)\, dx \qquad 7.4$$

$$F_2(S_2^+ = 0) = \int_{-\infty}^{0} f(x) \left[\int_{-\infty}^{-x} f(y)\, dy \right] dx \qquad 7.5$$

and

$$F_2(S_2^+ \leq S) = \int_{-\infty}^{0} f(x) \left[\int_{-\infty}^{-x} f(y)\, dy \right] dx + \int_{0}^{S} \left[Q f(s) + \int_{-\infty}^{S} f(x)\, f(S - x)\, dx \right] dS,$$

for S_2^+; 7.6

$$f_2(S_2^- = S) = P f(S) + \int_{S}^{\infty} f(x)\, f(S - x)\, dx, \qquad 7.7$$

$$F_2(S_2^- = 0) = \int_{0}^{\infty} f(x) \left[\int_{-x}^{\infty} f(y)\, dy \right] dx, \qquad 7.8$$

and

$$F_2(S_2^- \leq S) = \int_{0}^{\infty} f(x) \left[\int_{-x}^{\infty} f(y)\, dy \right] dx + \int_{S}^{0} \left[P f(S) + \int_{S}^{\infty} f(x)\, f(S - x)\, dx \right] dS$$

for S_2^-; and 7.9

$$f_2(R_2 = R) = \int_{0}^{R} f(x)\, f(R - x)\, dx + \int_{-R}^{0} f(x)\, f(-R - x)\, dx + 2f(R) \int_{-R}^{0} f(x)\, dx + 2f(-R) \int_{0}^{R} f(x)\, dx; \qquad 7.10$$

with

$$F_2(R_2 \leq R) = \int_{0}^{R} f_2(R)\, dR$$

for R_2, with $f_2(R)$ given by eq. 7.10.

In the case of a symmetrical density curve $f(x)$, $P = Q$, and eqs. 7.4 through 7.6 are identical to eqs. 7.7 through 7.9 for $S_n^+ = -S_n^-$. Equation 7.10 for a symmetrical $f(x)$ becomes

$$f_2(R_2 = R) = 2 \int_{0}^{R} f(x)\, f(R - x)\, dx + 4f(R) \int_{0}^{R} f(x)\, dx \qquad 7.12$$

For numerical computation of distributions, the integrals of the above equations are replaced by summations, and differentials dx, dS_2^+, dS_2^-, and dR_2 are replaced by differences Δx, ΔS_2^+, ΔS_2^-, and ΔR_2, respectively. All four differences are taken to be 0.10 for the Rhine River example, which is given in table 7.1.

Figure 7.5 gives the distributions of S_2^+, S_2^-, and R_2 in the form of probability density curves and probability mass (for $S_2^+ = 0$ and $S_2^- = 0$) for the example of figs. 7.1 and 7.2 and table 7.1.

The requirement that the sum of the areas under the probability density curves plus the probability mass (for $S_2^+ = 0$ and $S_2^- = 0$) are unities has been verified for all three probability distributions (S_n^+, S_n^-, and R_n).

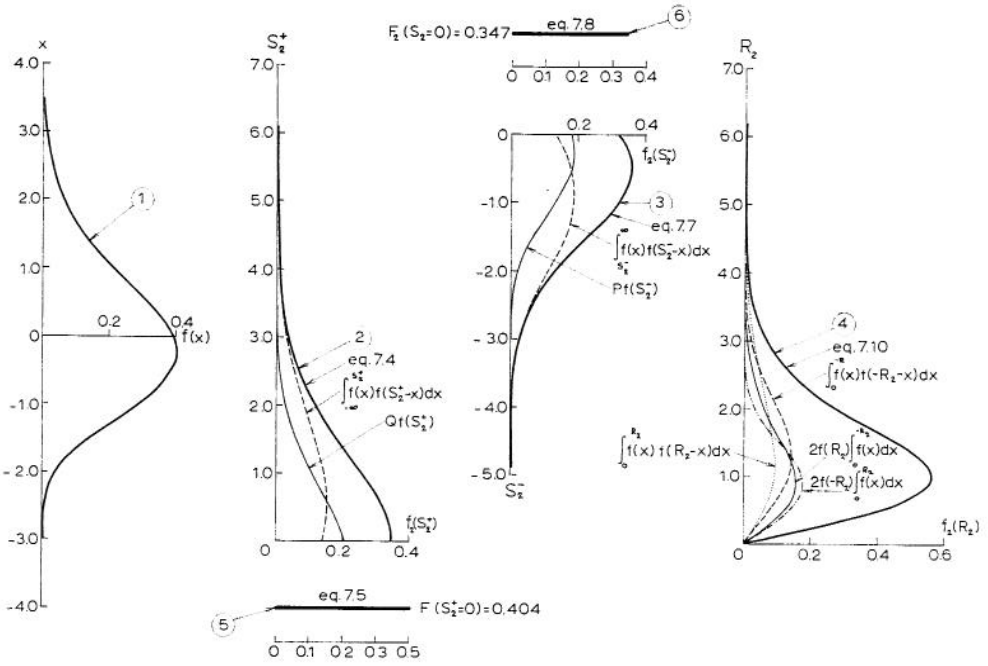


Fig. 7.5 Probability density curves of: x, surplus, deficit and range for n = 2, determined from the exact distributions by the finite difference method of integration for the standardized log-normal probability density, f(x), of the Rhine River's annual flows: (1) Probability density of x; (2) Probability density of S_2^+, and densities of two individual integrals; (3) Probability density of S_2^-, and densities of two individual integrals; (4) Probability density of R_2, and four individual integrals; (5) Probability mass of $S_2^+ = 0$; and (6) Probability mass of $S_2^- = 0$.

6. Distributions of surplus, deficit and range for n = 3. Figure 7.6 shows 18 subcases for different combinations of x_1, x_2 and x_3 (values of x for intervals 1, 2, and 3), numbered 3.1 through 3.18. The first number 3 refers to n, and the second number to the subcases. Figure 7.6 gives the corresponding values of S_3^+, S_3^-, and R_3. There is an inversion between S_3^+ and S_3^- for subcases 3.1 - 3.9 and 3.10 - 3.18. For R_3 these cases mean only a change of signs and limits. This is an important property for asymmetric distributions. For symmetric functions there are only 9 subcases for R_3, and 18 for S_3^+, but then $f_3(S_3^+) = f_3(-S_3^-)$.

Subcases 3.10 through 3.18 are inverts of subcases 3.1 through 3.9, respectively. Inversion is performed in such a way that only the inequality signs have been changed. This fact enables only the analysis of subcases 3.1 through 3.9. The probability density equations and probability mass developed in these equations are used to obtain, by a change of signs and integral limits, the corresponding expressions for subcases 3.10 through 3.18. These 18 subcases for $S_3^+ = S$ are:

$$(3.1)\quad \int_0^S f(S-y)\left[\int_0^y f(x)\, f(y-x)\,dx\right] dy$$

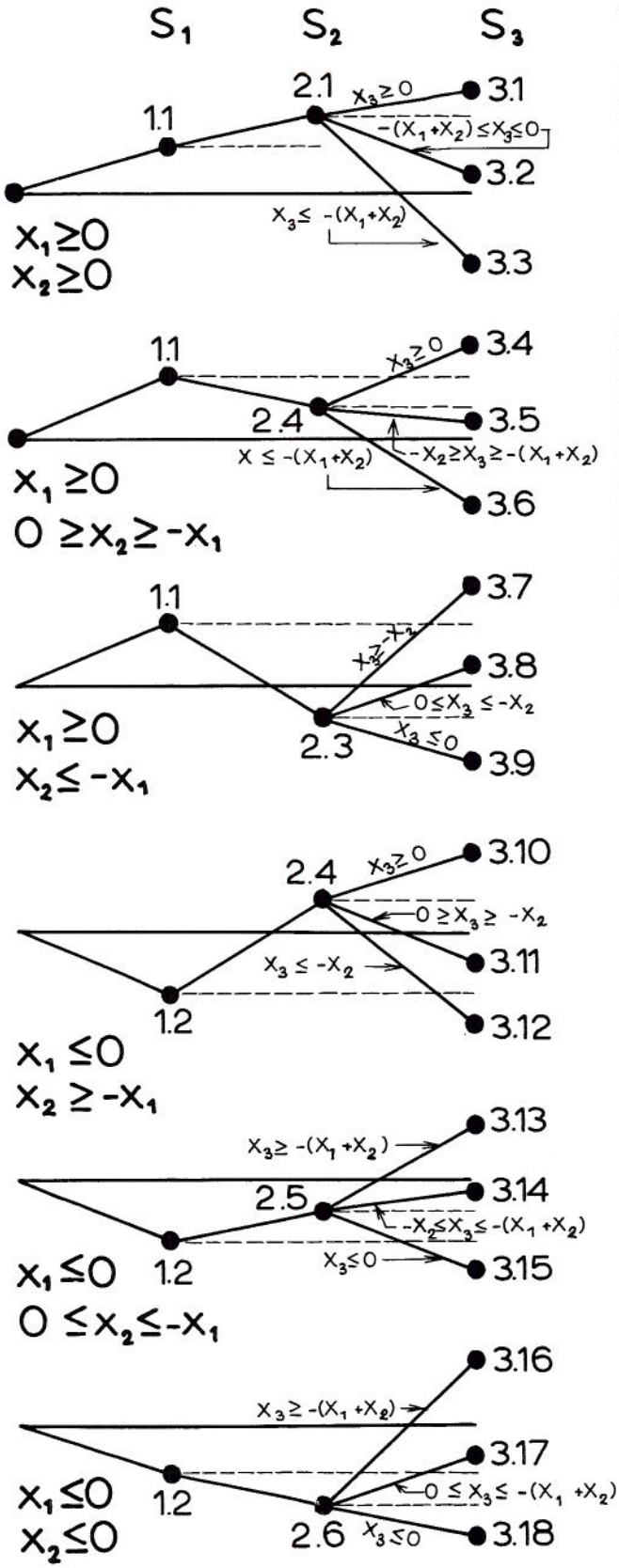
$$(3.2)\quad \int_0^S f(x)\, f(S-x)\,dx \int_{-S}^{0} f(x)\,dx$$

$$(3.3)\quad \int_0^S f(x)\, f(S-x)\,dx \int_{-\infty}^{-S} f(x)\,dx$$

$$(3.4)\quad \int_0^S f(S-y)\left[\int_y^S f(x)\, f(y-x)\,dx\right] dy$$

$$(3.5)\quad f(S)\int_{-S}^{0} f(x)\left[\int_{-(S+x)}^{-x} f(y)\,dy\right] dx$$

$$(3.6)\quad f(S)\int_{-S}^{0} f(x)\left[\int_{-\infty}^{-(S+x)} f(y)\,dy\right] dx$$



S_3^+	S_3^-	R_3
$x_1+x_2+x_3$	0	$x_1+x_2+x_3$
x_1+x_2	0	x_1+x_2
x_1+x_2	$x_1+x_2+x_3$	$\lvert x_3\rvert$
$x_1+x_2+x_3$	0	$x_1+x_2+x_3$
x_1	0	x_1
x_1	$x_1+x_2+x_3$	$\lvert x_2+x_3\rvert$
$x_1+x_2+x_3$	x_1+x_2	x_3
x_1	x_1+x_2	$\lvert x_2\rvert$
x_1	$x_1+x_2+x_3$	$\lvert x_2+x_3\rvert$
$x_1+x_2+x_3$	x_1	x_2+x_3
x_1+x_2	x_1	x_2
x_1+x_2	$x_1+x_2+x_3$	$\lvert x_3\rvert$
$x_1+x_2+x_3$	x_1	x_2+x_3
0	x_1	$\lvert x_1\rvert$
0	$x_1+x_2+x_3$	$\lvert x_1+x_2+x_3\rvert$
$x_1+x_2+x_3$	x_1+x_2	x_3
0	x_1+x_2	x_1+x_2
0	$x_1+x_2+x_3$	$\lvert x_1+x_2+x_3\rvert$

Fig. 7.6 Eighteen possible cases for different combinations of x_1, x_2 and x_3 in the determination of exact distributions of S_3^+, S_3^- and R_3 ($n = 3$, $\bar{x} = 10$).

$$(3.7)\quad \int_{-\infty}^{0} f(S-y)\left[\int_{0}^{S} f(x)\, f(y-x)\, dx\right] dy$$

$$(3.8)\quad f(S)\int_{-\infty}^{-S} f(x)\left[\int_{0}^{-x} f(y)\, dy\right] dx$$

$$(3.9)\quad f(S)\int_{-\infty}^{-S} f(x)\left[\int_{-\infty}^{0} f(y)\, dy\right] dx$$

$$(3.10)\quad \int_{0}^{S} f(S-y)\left[\int_{-\infty}^{0} f(x)\, f(y-x)\, dx\right] dy$$

$$(3.11)\quad \int_{-\infty}^{0} f(x)\, f(S-x)\, dx \int_{-(S-x)}^{0} f(x)\, dx$$

$$(3.12)\quad \int_{-\infty}^{0} f(x)\, f(S-x)\, dx \int_{-\infty}^{-(S-x)} f(x)\, dx$$

$$(3.13)\quad \int_{-\infty}^{0} f(S-y)\left[\int_{-\infty}^{y} f(x)\, f(y-x)\, dx\right] dy$$

$$(3.14)\quad F_3(S_3^+ = 0) = \int_{-\infty}^{0}\left[\int_{-\infty}^{y} f(x)\, f(y-x)\, dx \int_{-y}^{0} f(x)\, dx\right] dy$$

$$(3.15)\quad F_3(S_3^+ = 0) = \int_{-\infty}^{0}\left[\int_{-\infty}^{y} f(x)\, f(y-x)\, dx \int_{-\infty}^{0} f(x)\, dx\right] dy$$

$$(3.16)\quad \int_{-\infty}^{0} f(S-y)\left[\int_{0}^{y} f(x)\, f(y-x)\, dx\right] dy$$

$$(3.17)\quad F_3(S_3^+ = 0) = \int_{-\infty}^{0}\left[\int_{-y}^{0} f(x)\, f(y-x)\, dx \int_{-y}^{0} f(x)\, dx\right] dy$$

$$(3.18)\quad F_3(S_3^+ = 0) = \int_{-\infty}^{0}\left[\int_{y}^{0} f(x)\, f(y-x)\, dx \int_{-\infty}^{0} f(x)\, dx\right] dy$$

Subcases 3.1, 3.4, 3.7, 3.10, 3.13 and 3.16, then subcases 3.2, 3.3, 3.11 and 3.12, and finally subcases 3.5, 3.6, 3.8 and 3.9 are combined in integrals of the same type. They give the probability density of S_3^+ in the form of

$$f_3(S_3^+ = S) = \int_{-\infty}^{S} f(S-y)\left[\int_{-\infty}^{S} f(x)\, f(y-x)\, dx\right] dy +$$
$$+ Q\int_{-\infty}^{S} f(x)\, f(S-x)\, dx + f(S)\int_{-\infty}^{0} f(x)\left[\int_{-\infty}^{-x} f(y)\, dy\right] dx \qquad 7.13$$

By combining subcases 3.14, 3.15, 3.17 and 3.18 the probability mass for $S_3^+ = 0$ is obtained as

$$F_3(S_3^+ = 0) = \int_{-\infty}^{0}\left[\int_{-\infty}^{0} f(x)\, f(y-x)\, dx \int_{-\infty}^{-y} f(x)\, dx\right] dy \qquad 7.14$$

The distribution of S_3^+ is then

$$F_3(S_3^+ \le S) = F_3(S_3^+ = 0) + \int_{0}^{S} f_3(S)\, dS \qquad 7.15$$

with $F_3(S_3^+ = 0)$ given by eq. 7.14 and $f_3(S)$ given by eq. 7.13.

The probability densities of eq. 7.13 and the probability mass of eq. 7.14 are computed for the example of figs. 7.1 and 7.2 and table 7.1. The sum under the curve of eq. 7.13 plus the probability mass of eq. 7.14 gives 0.999345.

Similarly, the probability density of S_3^- is obtained as

$$f_3(S_3^- = S) = \int_{S}^{\infty} f(S-y)\left[\int_{S}^{\infty} f(x)\, f(y-x)\, dx\right] dy +$$
$$+ P\int_{S}^{\infty} f(x)\, f(S-x)\, dx + f(S)\int_{0}^{\infty} f(x)\left[\int_{-x}^{\infty} f(y)\, dy\right] dx \qquad 7.16$$

with the probability mass for $S_3^- = 0$

$$F_3(S_3^- = 0) = \int_{0}^{\infty}\left[\int_{0}^{\infty} f(x)\, f(y-x)\, dx \int_{-y}^{\infty} f(x)\, dx\right] dy \qquad 7.17$$

The distribution of S_3^- is then

$$F_3(S_3^- \le S) = F_3(S_3^- = 0) + \int_{S}^{0} f_3(S)\, dS \qquad 7.18$$

with $F_3(S_3^- = 1)$ given by eq. 7.17 and $f_3(S)$ given by eq. 7.16.

The probabilities of eq. 7.16 and 7.17 are computed similarly as for S_3^+. The sum is 0.999346.

Figure 7.7 depicts the probability density curves, and the probability mass of S_3^+ and S_3^- as obtained by the finite differences integration of eqs. 7.13 through 7.18.

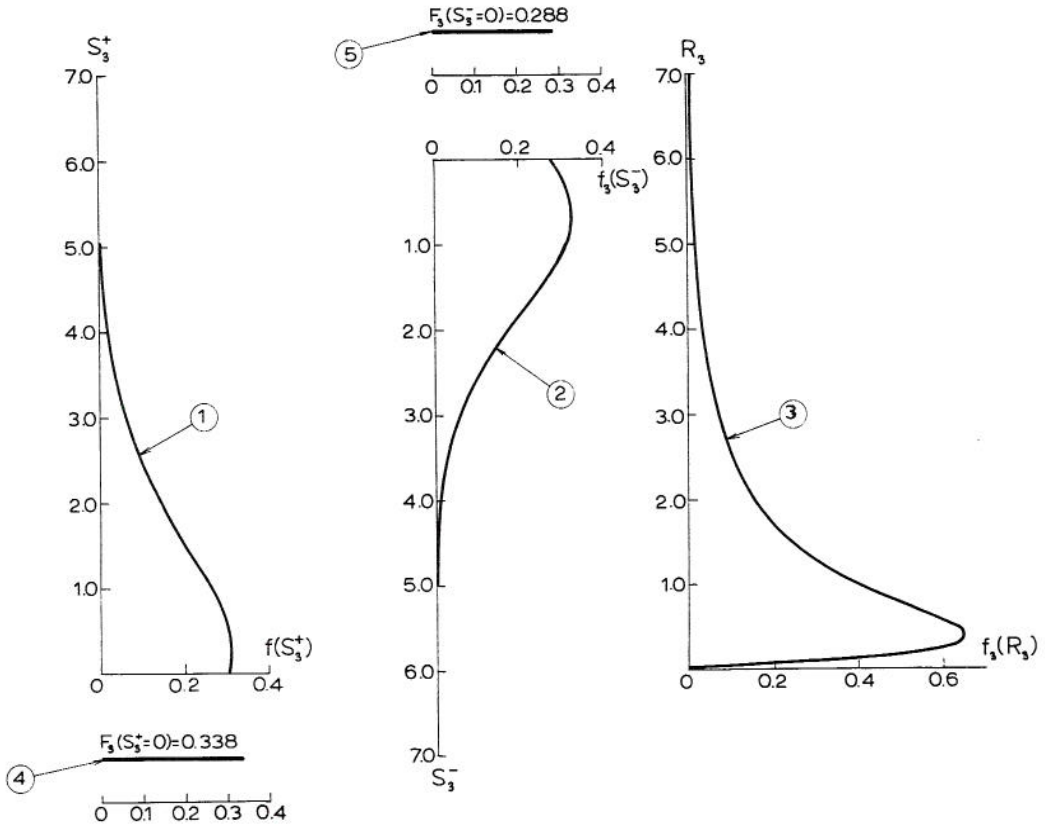


Fig. 7.7 Probability density curves of S_3^+, S_3^- and R_3 determined from the exact distributions by the finite difference method of integration for the independent standardized log-normal probability density curve, $f(x)$, of the Rhine River's annual flows: (1) Probability densities of S_3^+; (2) Probability densities of S_3^-; (3) Probability densities of R_3; (4) Probability mass for $S_3^+ = 0$; and (5) Probability mass for $S_3^- = 0$.

Similarly as for S_3^+, the 18 subcases of $R_3 = R$ have the following expressions for its probability densities

$$(3.1)\quad \int_0^R f(R-y)\left[\int_0^y f(y-x)\,f(x)\,dx\right] dy$$

$$(3.2)\quad \int_{-R}^0 f(x)\,dx \int_0^R f(R-y)\,f(y)\,dy$$

$$(3.3)\quad f(-R)\int_0^R \left[\int_0^y f(y-x)\,f(x)\,dx\right] dy$$

$$(3.4)\quad \int_0^R f(R-y)\left[\int_y^R f(y-x)\,f(x)\,dx\right] dy$$

$$(3.5)\quad f(R)\int_{-R}^0 f(y)\left[\int_{-(R+y)}^{-y} f(x)\,dx\right] dy$$

$$(3.6)\quad \int_{-y}^R f(x)\,dx \int_{-R}^0 f(-R-y)\,f(y)\,dy$$

$$(3.7)\quad f(R)\int_{-R}^0 \left[\int_0^{R+y} f(y-x)\,f(x)\,dx\right] dy$$

(3.8) $f(-R)\int_0^R f(x)\,dx \int_0^R f(x)\,dx$

(3.9) $\int_0^{-y} f(x)\,dx \int_{-R}^{0} f(-R-y)\,f(y)\,dy$

(3.10) $\int_{-y}^{0} f(x)\,dx \int_0^R f(R-y)\,f(y)\,dy$

(3.11) $f(R)\int_{-R}^{0} f(x)\,dx \int_{-R}^{0} f(x)\,dx$

(3.12) $f(-R)\int_0^R \left[\int_{-R+y}^{0} f(y-x)\,f(x)\,dx\right] dy$

(3.13) $\int_{-R}^{-y} f(x)\,dx \int_0^R f(R-y)\,f(y)\,dy$

(3.14) $f(-R)\int_0^R f(y)\left[\int_{-y}^{R-y} f(x)\,dx\right] dy$

(3.15) $\int_{-R}^{0} f(-R-y)\left[\int_{-R}^{y} f(y-x)\,f(x)\,dx\right] dy$

(3.16) $f(R)\int_{-R}^{0}\left[\int_{y}^{0} f(y-x)\,f(x)\,dx\right] dy$

(3.17) $\int_0^R f(x)\,dx \int_{-R}^{0} f(-R-y)\,f(y)\,dy$

(3.18) $\int_{-R}^{0} f(-R-y)\left[\int_{y}^{0} f(y-x)\,f(x)\,dx\right] dy$

By combining identical integrals: subcases 3.1, 3.4, 3.15 and 3.18; then subcases 3.5, 3.7 and 3.16; subcases 3.3, 3.12, and 3.14; subcases 3.2, 3.10 and 3.13; subcases 3.6, 3.9 and 3.17; and finally 3.8 and 3.11, then

$$f_3(R_3 = R) = \int_0^R f(R-y)\left[\int_0^R f(y-x)\,f(x)\,dx\right] dy +$$

$$+ \int_{-R}^{0} f(-R-y)\left[\int_{-R}^{0} f(y-x)\,f(x)\,dx\right] dy +$$

$$+ f(R)\int_{-R}^{0}\left[f(y)\int_{-(R+y)}^{-y} f(x)\,dx + \int_{y}^{0} f(y-x)\,f(x)\,dx + \right.$$

$$\left. + \int_0^{R+y} f(y-x)\,f(x)\,dx\right] dy + f(-R)\int_0^R \left[f(y)\int_{-y}^{R-y} f(x)\,dx + \right.$$

$$\left. + \int_0^{y} f(y-x)\,f(x)\,dx + \int_{-R+y}^{0} f(y-x)\,f(x)\,dx\right] dy +$$

$$+ 2\int_{-R}^{0} f(x)\,dx \int_0^R f(R-x)\,f(x)\,dx +$$

$$+ 2\int_0^R f(x)\,dx \int_{-R}^{0} f(-R-x)\,f(x)\,dx +$$

$$+ f(-R)\left[\int_0^R f(x)\,dx\right]^2 + f(R)\left[\int_{-R}^{0} f(x)\,dx\right]^2 \qquad 7.19$$

The distribution of R_3 is computed as

$$F_3(R_3 \le R) = \int_0^R f_3(R)\,dR \qquad 7.20$$

and $f_3(R)$ is given by eq. 7.19.

When the probability density $f(x)$ is symmetrical, the above integrals of eq. 7.19 may be simplified to:

$$f_3(R_3 = R) = 2\int_0^R f(R-y)\left[\int_0^R f(y-x)\,f(x)\,dx\right] dy +$$

$$+ 2f(R)\int_0^R \left[f(y)\int_{-y}^{R-y} f(x)\,dx + \int_0^{y} f(y-x)\,f(x)\,dx + \right.$$

$$\left. + \int_0^{R+y} f(y-x)\,f(x)\,dx\right] dy +$$

$$+ 4\int_0^R f(x)\,dx \int_0^R f(R-x)\,f(x)\,dx +$$

$$+ 2f(R)\left[\int_0^R f(x)\,dx\right]^2 \qquad 7.21$$

Eight probability density curves for the eight integrals of eq. 7.19 are obtained by digital computer integrations. This integration is accomplished by passing from integrals to summation in eq. 7.19 with $\Delta x = \Delta y = \Delta R_3$, and using $\Delta x = 0.10$. The probability density curve of R_3 is plotted in fig. 7.7. The area under the total probability density curve is 0.99797. The main reason for these equations of exact distributions of S_3^+, S_3^-, and R_3 to be given in this study is to show the complexity of exact distributions even for $n = 3$. Equation 7.19 with eight different integrals and limits shows that the exact distribution of range is difficult to obtain even for n as small as 3, 4, or 5. W. Feller [4] pointed out this same fact in his study of asymptotic distributions.

7 Comparison of the analytical method with the data generation and the empirical methods, by S_3^+, S_3^- and R_3 distributions (for $n = 3$). At this point the densities, mass and distributions of S_3^+, S_3^-, and R_3, for the annual flow of the Rhine River as obtained by the empirical, data generation and analytical methods should be compared. In this comparison the values of S_3^+, S_3^- and R_3, in the above integrations, are multiplied and densities are divided by $C_v = 0.159$. The corresponding curves are plotted as lines (3) in figs. 5.2 through 5.4 for $n = 3$. This comparison by distributions illustrates two factors: (a) The empirical method gives distributions which are not smooth; and (b) The data generation method gives values which are very close to those obtained by integration which uses the finite differences method of exact distributions. The selection of $\Delta x = \Delta y = \Delta S_3^+ = \Delta R_3$ plays a significant role in the integration accuracy of exact distributions.

8. Distributions for the $n = 4$. Figure 7.8 shows the 54 cases of S_4^+, S_4^- and R_4. These cases are various combinations of x_1, x_2, x_3 and x_4, numbered 4.1 through 4.54. Cases 4.28 through 4.54 are identical to the corresponding cases 4.27 through 4.1. Exceptions to this statement are that S_4^+ and S_4^- are interchanged and the signs of variables and limits of integrals or sums have to be modified appropriately. The first number refers to $n = 4$, and the second number to the case as designated in fig. 7.8. In the last four right hand columns fig. 7.8 yields: the case number, S_4^+, S_4^- and R_4 as expressed in values x_1, x_2, x_3 and x_4.

Figure 7.8 is presented to illustrate how complicated this method becomes even for $n = 4$. Many cases can be combined as being of the same type of integrals but with different integral limits. In these cases the expressions for the distributions of S_4^+, S_4^-, and R_4 become complicated, and as such are not reported in this study. This involvement supports Feller's [4] statement that even for $n = 4$ the exact distributions are difficult to obtain. For $n = 5$ there are 162 cases ($3 \times 54 = 162$) and any attempt to derive the exact distributions of surplus, deficit and range become intractable from a practical point of view.

The other approach for the derivation of exact distributions of surplus, deficit and range is in combining all similar elementary cases of the previously mentioned systematic method of analysis into the cases of the same type of integrals. It reduces the number of cases to be separately investigated. However, this new approach is likely to omit some elementary cases.

9. Distribution of surplus, deficit and range as obtained from x_m variables by using the changing integration region. The probability distribution $F_n(S_n^+)$ may be expressed as

$$F_n(S_n^+) = \int_I (n) \int \prod_{m=1}^{m=n} f(x_m)\, dx_m \qquad 7.22$$

with the region of integration defined as

$$I: \left\{ \sum_{i=1}^{m} x_i = S_m \leq S_n^+, \text{ with } m = 0, 1, \ldots, n. \right\} \qquad 7.23$$

Difficulties in integrating eq. 7.22 come from the changing integration region I. Equation 7.22 is approximated by the summation

$$F_n(S_n^+) = \sum_{-\infty}^{S_1 \leq S_n^+} \sum_{-\infty}^{S_2 \leq S_n^+} \cdots\cdots \sum_{-\infty}^{S_n \leq S_n^+} f(x_1).$$

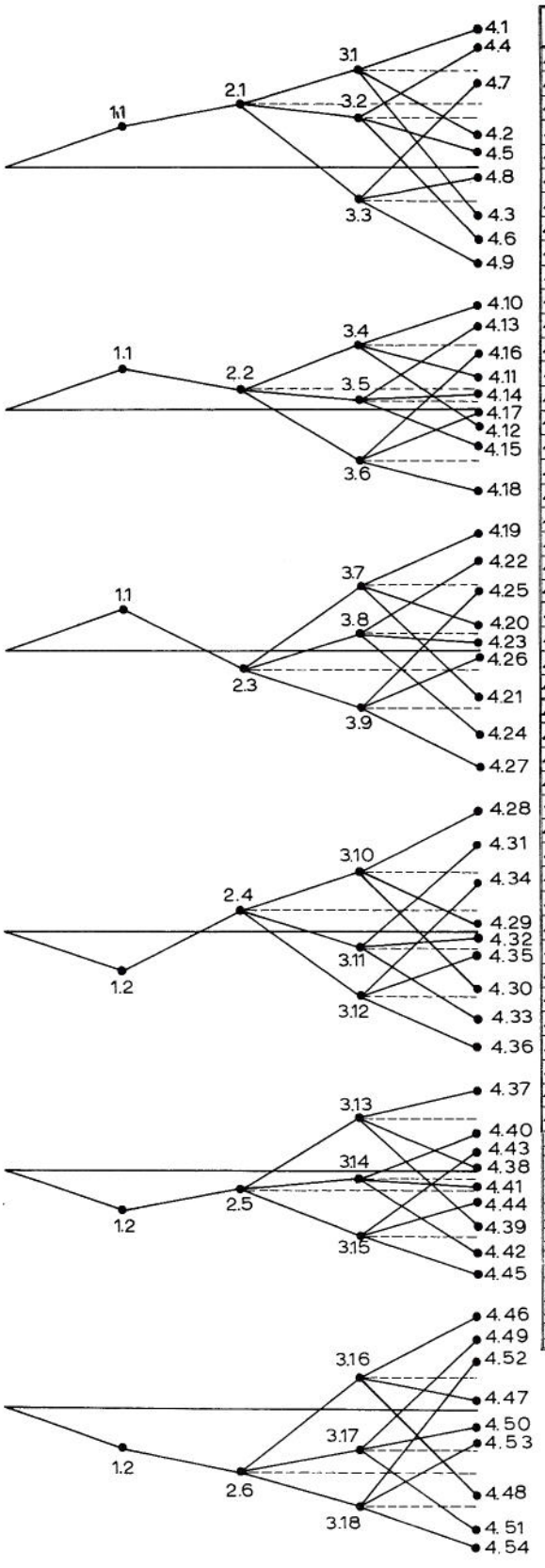
$$f(x_2) \ \ldots \ f(x_n)\ \Delta x_1 . \Delta x_2 . \ldots . \Delta x_n \qquad 7.24$$

As S_n^+ cannot be smaller than $S_o = 0$, or S_n^+ is always positive, the probability mass for $S_n^+ = 0$ is equal to eq. 7.24 with the lower limit $-\infty$ and the upper limit zero. In summing eq. 7.24 the value $S_n = S_{n-1} + x_n$, or $S_{n-1} = S_n - x_n$; $S_{n-2} = S_n - x_n - x_{n-1}$ and so on, with $S_1 = S_n - x_n - x_{n-1} - \ldots - x_2 = x_1$. Taking $\Delta x_1 = \Delta x_2 = \ldots = \Delta x_n = \Delta x$, for a given n, there is a constant $(\Delta x)^n$ in eq. 7.24, or

$$F_n(S_n^+) = (\Delta x)^n \sum_{-\infty}^{S_1 \leq S_n^+} \sum_{-\infty}^{S_2 \leq S_n^+} \cdots \sum_{-\infty}^{S_n \leq S_n^+} f(x_1).$$

$$f(x_2). \ldots . f(x_n). \qquad 7.25$$

The following steps are feasible when using a digital computer to sum up eq. 7.25 for a given S_n^+: (a) the sum of $f_n(x_n)$ is determined in the limits $-\infty$ to S_n^+; (b) this determined value is multiplied by the sum of $f_{n-1}(x_{n-1})$ for each S_n of the previous sum; and (c) x_n is then selected in such a way that $S_{n-1} = (S_n - x_n) \leq S_n^+$ with x_n in the limits $-\infty$ to $+\infty$ or $x_n \geq (S_n - S_n^+)$; and so on. This summation requires fast digital computer with a very large core storage capacity. When n is greater than $n = 5$ the complexity of this integration may discourage even the digital computer approach.



	S_4^+	S_4^-	R_4
4.1	$x_1+x_2+x_3+x_4$	0	$x_1+x_2+x_3+x_4$
4.2	$x_1+x_2+x_3$	0	$x_1+x_2+x_3$
4.3	$x_1+x_2+x_3$	$x_1+x_2+x_3+x_4$	$\lvert x_4 \rvert$
4.4	$x_1+x_2+x_3+x_4$	0	$x_1+x_2+x_3+x_4$
4.5	x_1+x_2	0	x_1+x_2
4.6	x_1+x_2	$x_1+x_2+x_3+x_4$	x_3+x_4
4.7	$x_1+x_2+x_3+x_4$	$x_1+x_2+x_3$	x_4
4.8	x_1+x_2	$x_1+x_2+x_3$	$\lvert x_3 \rvert$
4.9	x_1+x_2	$x_1+x_2+x_3+x_4$	$\lvert x_3+x_4 \rvert$
4.10	$x_1+x_2+x_3+x_4$	0	$x_1+x_2+x_3+x_4$
4.11	$x_1+x_2+x_3$	0	$x_1+x_2+x_3$
4.12	$x_1+x_2+x_3$	$x_1+x_2+x_3+x_4$	$\lvert x_4 \rvert$
4.13	$x_1+x_2+x_3+x_4$	0	$x_1+x_2+x_3+x_4$
4.14	x_1	0	x_1
4.15	x_1	$x_1+x_2+x_3+x_4$	$\lvert x_1+x_2+x_3 \rvert$
4.16	$x_1+x_2+x_3+x_4$	$x_1+x_2+x_3$	x_4
4.17	x_1	$x_1+x_2+x_3$	$\lvert x_2+x_3 \rvert$
4.18	x_1	$x_1+x_2+x_3+x_4$	$\lvert x_1+x_2+x_3 \rvert$
4.19	$x_1+x_2+x_3+x_4$	x_1+x_2	x_3+x_4
4.20	$x_1+x_2+x_3$	x_1+x_2	x_3
4.21	$x_1+x_2+x_3$	$x_1+x_2+x_3+x_4$	$\lvert x_4 \rvert$
4.22	$x_1+x_2+x_3+x_4$	x_1+x_2	x_3+x_4
4.23	x_1	x_1+x_2	$\lvert x_2 \rvert$
4.24	x_1	$x_1+x_2+x_3+x_4$	$\lvert x_2+x_3+x_4 \rvert$
4.25	$x_1+x_2+x_3+x_4$	$x_1+x_2+x_3$	x_4
4.26	x_1	$x_1+x_2+x_3$	$\lvert x_2+x_3 \rvert$
4.27	x_1	$x_1+x_2+x_3+x_4$	$\lvert x_2+x_3+x_4 \rvert$
4.28	$x_1+x_2+x_3+x_4$	x_1	$x_2+x_3+x_4$
4.29	$x_1+x_2+x_3$	x_1	x_2+x_3
4.30	$x_1+x_2+x_3$	$x_1+x_2+x_3+x_4$	$\lvert x_4 \rvert$
4.31	$x_1+x_2+x_3+x_4$	x_1	$x_2+x_3+x_4$
4.32	x_1+x_2	x_1	x_2
4.33	x_1+x_2	$x_1+x_2+x_3+x_4$	$\lvert x_3+x_4 \rvert$
4.34	$x_1+x_2+x_3+x_4$	$x_1+x_2+x_3$	x_4
4.35	x_1+x_2	$x_1+x_2+x_3$	$\lvert x_3 \rvert$
4.36	x_1+x_2	$x_1+x_2+x_3+x_4$	$\lvert x_3+x_4 \rvert$
4.37	$x_1+x_2+x_3+x_4$	x_1	$x_2+x_3+x_4$
4.38	$x_1+x_2+x_3$	x_1	x_2+x_3
4.39	$x_1+x_2+x_3$	$x_1+x_2+x_3+x_4$	$\lvert x_4 \rvert$
4.40	$x_1+x_2+x_3+x_4$	x_1	$x_2+x_3+x_4$
4.41	0	x_1	$\lvert x_1 \rvert$
4.42	0	$x_1+x_2+x_3+x_4$	$\lvert x_1+x_2+x_3+x_4 \rvert$
4.43	$x_1+x_2+x_3$	$x_1+x_2+x_3$	x_4
4.44	0	$x_1+x_2+x_3+x_4$	$\lvert x_1+x_2+x_3 \rvert$
4.45	0	$x_1+x_2+x_3+x_4$	$\lvert x_1+x_2+x_3+x_4 \rvert$
4.46	$x_1+x_2+x_3+x_4$	x_1+x_2	x_3+x_4
4.47	$x_1+x_2+x_3$	x_1+x_2	x_3
4.48	$x_1+x_2+x_3$	$x_1+x_2+x_3+x_4$	$\lvert x_4 \rvert$
4.49	$x_1+x_2+x_3+x_4$	x_1+x_2	x_3+x_4
4.50	0	x_1+x_2	$\lvert x_1+x_2 \rvert$
4.51	0	$x_1+x_2+x_3+x_4$	$\lvert x_1+x_2+x_3+x_4 \rvert$
4.52	$x_1+x_2+x_3+x_4$	$x_1+x_2+x_3$	x_4
4.53	0	$x_1+x_2+x_3$	$\lvert x_1+x_2+x_3 \rvert$
4.54	0	$x_1+x_2+x_3+x_4$	$\lvert x_1+x_2+x_3+x_4 \rvert$

Fig. 7.8 Fifty-four possible cases for different combinations of x_1, x_2, x_3 and x_4 in the determination of exact distributions of S_4^+, S_4^- and R_4 ($n = 4$, $\bar{x} = 10$).

10. Use of joint distribution of sums of x_m. Assume that the independent variable x is distributed according to the probability density function f(x). The sum of m independent standard variables x_1, x_2, ... , x_m is

$$S_m = x_1 + \ldots + x_m \qquad 7.26$$

with $S_o = 0$ for $m = 0$, and with $m = 0, 1, 2, \ldots n$. This new variable S_m satisfies the first order stationarity; namely, the expected mean is zero and it is independent of m. It does not satisfy the second or higher order stationarity. The variance increases with an increase of m, and for the standard normal variable, var $S_m = m$, for $m = 0, 1, 2, \ldots n$. The variable S_m is correlated serially and the expected serial correlation coefficients depend on the position m. For the position m, the first serial correlation coefficient $r_1(m)$ is between S_m and S_{m+1}. The second serial correlation coefficient is between S_m and S_{m+2}, and so on for the higher order coefficients. For the position m the covariance of S_m and S_{m+1} is m. Then,

$$r_1(m) = \frac{\operatorname{cov} S_m S_{m+1}}{(\operatorname{var} S_m \operatorname{var} S_{m+1})^{1/2}} = \sqrt{\frac{m}{m+1}} \qquad 7.27$$

The dependence of the sequence of the variable S_m increases with an increase of m. Then $r_1(1) = 1/\sqrt{2}$, and $r_1(\infty) = 1$. Similarly,

$$r_k(m) = \frac{m}{\sqrt{m(m+k)}} = \sqrt{\frac{m}{m+k}}$$

In this case, the serial correlation matrix of $r_k(m)$ is

m		S_1	S_2	S_3	S_4 ---	S_n
1	S_1	1	$\sqrt{\frac{1}{2}}$	$\sqrt{\frac{1}{3}}$	$\sqrt{\frac{1}{4}}$ ---	$\sqrt{\frac{1}{n}}$
2	S_2		1	$\sqrt{\frac{2}{3}}$	$\sqrt{\frac{1}{2}}$ ---	$\sqrt{\frac{2}{n}}$
3	S_3			1	$\sqrt{\frac{3}{4}}$ ---	$\sqrt{\frac{3}{n}}$
4	S_4				1 ---	$\sqrt{\frac{4}{n}}$
	⋮					⋮
n	S_n					1

By using the convolution integral, the distribution of S_{m+1} can be obtained from that of S_m as

$$f_{m+1}(S_{m+1}) = \int_{-\infty}^{+\infty} f(x) f_m(S_{m+1} - x)\, dx \qquad 7.30$$

The surplus S_n^+, the deficit S_n^- and the range R_n are

$$S_n^+ = {}_m\mathrm{Max} \left[S_m\right] \qquad 7.31$$

$$S_n^- = {}_m\mathrm{Min} \left[S_m\right] \qquad 7.32$$

$$R_n = {}_m\mathrm{Max} \left[S_m\right] - {}_m\mathrm{Min} \left[S_m\right] \qquad 7.33$$

for $m = 0, 1, \ldots, n$.

$$F_n(S_n^+) = P_r(S_m \leq S_n^+, \text{ for } m = 0, 1, \ldots, n) \qquad 7.34$$

which is the probability distribution of S_n^+. In other words, the probability of a given S_n^+ is the probability that all dependent S_m are smaller than or equal to S_n^+ for any $m = 0, 1, \ldots, n$.

Equations 7.27 through 7.29 enable a derivation of joint analytical distribution $f(S_1, S_2, \ldots, S_n)$ when the serial correlation coefficients are known. Then the surplus S_n^+ is

$$F_n(S_n^+ \leq S_n) = \int_{-\infty}^{S_n} \int_{-\infty}^{S_n} \cdots \int^{(n)} f_n(S_1, S_2, \ldots, S_n)\, dS_1 .\, dS_2 . \ldots dS_n \qquad 7.35$$

The problem is in deriving a proper analytical equation for the joint distribution of S_m ($m = 1, 2, \ldots, n$). This is feasible only for a normal function. Integration is then accomplished by a finite difference procedure. With an increase of n the expressions for the exact distributions and their integrations become more and more complex.

11. Comparison of three methods of exact distribution computations. A comparison of the above three methods of obtaining the exact distributions of surplus, deficit and range shows that all three approaches are similarly complex. The difficulties in obtaining these distributions grow by a geometric progression of n by an increase of n.

These difficulties and an increase in the computations needed by a geometric progression of n lead to the following conclusions:

(1) The practical aspects of obtaining the above distributions for n approximately five or greater, do not justify the use of any of the three methods;

(2) The determination of exact expressions for parameters of exact distributions of surplus, deficit and range, and the fitting of functions to exact distributions by the use of the above parameters, becomes an attractive practical solution; and,

(3) The data generation method, with large generated samples of time series of a given distribution and a given time dependence, is the attractive method of obtaining the distributions of surplus, deficit and range, closest to the exact distributions.

CHAPTER VIII

DISTRIBUTION OF SURPLUS, DEFICIT AND RANGE FOR INDEPENDENT AND DEPENDENT STANDARD NORMAL VARIABLES

1. Independent Normal variables. The distribution of an independent normal variable, X, is described by its mean μ and its standard deviation σ. The surplus, deficit and range of an independent normal variable are equal to S_n^+, S_n^-, and R_n of the independent standard normal variable, $x = (x - \mu)/\sigma$, multiplied by the standard deviation σ of X, and their probability densities divided by σ. Therefore, it suffices to investigate the case of the independent standard normal variable to cover all independent normal variables.

To simplify the text, the independent standard normal variable x is often designated by (0, 1, 0), which means $\mu = 0$, $\sigma = 1$, and $\rho_k = 0$ (all autocorrelation coefficients are zeros). The dependent standard normal variable is designated by (0, 1, ρ), with ρ representing the time dependence. The independent normal variables are designated by (μ, σ, 0), with $\rho_k = 0$ for all k. The dependent normal variables are represented by (μ, σ, ρ), with ρ the symbol of dependence. As the surplus and deficit have the identical distributions for $S_n^+ = -S_n^-$ of a symmetrical (0, 1, 0) - variable, it is sufficient in this report to give only the properties of the surplus. Therefore, whatever is stated about S_n^+ and $S_n^+(\bar{X}_n)$ is also valid for $-S_n^-$ and $-S_n^-(\bar{X}_n)$.

2. Asymptotic mean and variance of surplus, range, adjusted surplus and adjusted range for (0, 1, 0) - variable. W. Feller [4] developed the asymptotic distribution of the range, R_t, of the continuous sum, S_t. This distribution is based on the concept of a continuously changing normal variable, S_t, as cumulative sums of x. It should be noted that x is subjected to a Bachelier-Wiener process which uses the distribution function that occurs in the Kolmogorov-Smirnov theorem on empirical distribution functions. He also obtained the asymptotic mean and asymptotic variance of the range, R_t. These asymptotic parameter values of the range, R_t, are determined as approximations. The sum S_n can be considered as the value at time $t = n$ of the continuously changing variable S_t, and R_n the value of R_t at $t = n$. The expected asymptotic range is

$$E(R_n) = 2\left(\frac{2n}{\pi}\right)^{1/2} = 1.5958\ldots\sqrt{n} \simeq 1.6\sqrt{n} \qquad 8.1$$

and the asymptotic variance of range is

$$\sigma_n^2 = \operatorname{var} R_n = 4n\left(\ln 2 - \frac{2}{\pi}\right) \simeq 0.218\, n \qquad 8.2$$

where ln is the natural logarithm, and n is any position in a discrete series. The expressions: the expected value and the (population) mean are used in this text as interchangeable synonymous terms.

According to W. Feller [4], the asymptotic mean and asymptotic variance for the range given by eqs. 8.1 and 8.2 seem to agree with the exact values computed for extreme cases. This statement is true when the values of x_i are only ± 1, having probabilities 1/2, and n is small (6, 10, 12). Considering the smallness of n, and the fact that the assumed distribution of variable [+1, -1] is most unfavorable for the approximation, according to Feller, the above equations appear surprisingly good. But, they also bear out the expectation that the ranges of discrete sums S_n should be smaller than those of the corresponding continuously varying sum, S_t.

As $E(S_n^+) = -E(S_n^-)$, and $E(R_n) = E(S_n^+) - -E(S_n^-)$, the asymptotic mean of S_n^+ is

$$E(S_n^+) = \frac{1}{2}E(R_n) = \left(\frac{2n}{\pi}\right)^{1/2} = 0.8\sqrt{n} \qquad 8.3$$

The asymptotic variance of S_n^+ is

$$\operatorname{var} S_n^+ = \frac{\operatorname{var} R_n}{2(1-\rho_n)} = \frac{2n(\ln 2 - \frac{2}{\pi})}{1-\rho_n} \simeq \frac{0.109\, n}{1-\rho_n} \qquad 8.4$$

with ρ_n the correlation coefficient between S_n^+ and S_n^-.

Figure 8.1 gives the computed values of ρ_n obtained by the data generation method from 100,000 independent normal numbers. The points show a convergence of ρ_n to the asymptotic value $\rho_\infty = 0.70$ with an increase of n. The curve fitting in fig. 8.1 gives the following approximate expression

$$\rho_n = 1 - \frac{0.30\sqrt{n}}{\sqrt{n} - 0.37} \qquad 8.5$$

with the constants 0.30 and 0.37 obtained by the least squares method. The relationship of eq. 8.1 is plotted in fig. 8.1 as a solid line.

By using ρ_n of eq. 8.5, the approximation of eq. 8.4 becomes

$$\operatorname{var} S_n^+ = 0.363\, n - 0.134\sqrt{n} \qquad 8.6$$

Equation 8.4 may be written as

$$\rho_n = 1 - \frac{\operatorname{var} R_n}{2 \operatorname{var} S_n^+} \qquad 8.7$$

The asymptotic variance of R_n is given by eq. 8.2. The asymptotic variance of S_n^+ was not available for this analysis. However, A. Anis [6] gives an

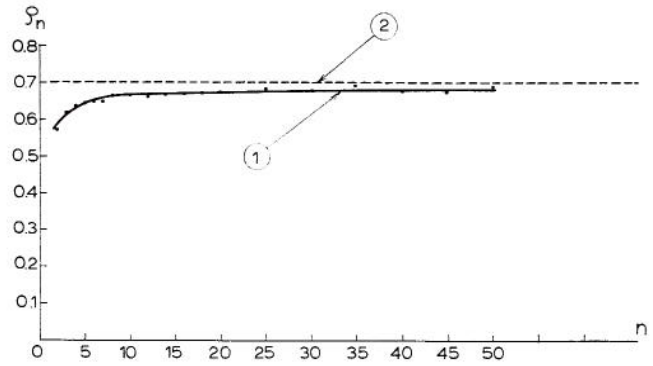


Fig. 8.1 The correlation coefficient ρ_n between the surplus (S_n^+) and the deficit (S_n^-) of an independent standard normal variable, as function of n. Points are obtained by the data generation method (100,000 independent numbers): (1) Curve, fitted to points by least squares method; and, (2) Asymptotic value $\rho_n = 0.70$ for very large n.

approximation to the second moment about zero of the exact distribution of S_n^+

$$\mu_2' (S_n^+) \simeq n - \frac{2+\sqrt{2}}{\pi}\sqrt{n} \qquad 8.8$$

As the asymptotic mean of S_n^+ is given by eq. 8.2, then by using eq. 8.8 the approximate variance of S_n^+ becomes

$$\operatorname{var} S_n^+ = n\left(1-\frac{2}{\pi}\right) - \frac{2+\sqrt{2}}{\pi}\sqrt{n} = 0.363n - 1.09\sqrt{n} \qquad 8.9$$

Equation 8.9 is obtained by using the asymptotic mean of eq. 8.3 and an approximation to the exact second moment about zero of eq. 8.8. As the asymptotic mean for small n (as it will be shown later in this text) is greater than the exact mean, eq. 8.9 gives negative values for small n. However, the purpose of eq. 8.9 is only a derivation of asymptotic value of ρ_n.

With eqs. 8.2 and 8.9, eq. 8.7 becomes

$$\rho_n = 1 - \frac{2\left(\ln 2 - \frac{2}{\pi}\right)}{\left(1-\frac{2}{\pi}\right) - \frac{2+\sqrt{2}}{\pi\sqrt{n}}} \simeq 1 - \frac{0.30\sqrt{n}}{\sqrt{n} - 3} \qquad 8.10$$

When n is very large $\rho_n = 0.70$, approximately. However, the convergence to $\rho_\infty = 0.70$ is very slow. Even for $n = 144$, $\rho_n = 0.60$. For $n = 9$ the value ρ_n becomes negative infinite. Eq. 8.10 is useful only for its asymptotic value $\rho_\infty = 0.70$. Figure 8.1 shows that ρ_n converges more rapidly to the value $\rho_\infty = 0.70$ than eq. 8.10 indicates. It should be stressed that eq. 8.10 is derived from the approximate or asymptotic values of variances, while fig. 8.1 for $\rho_n = f(n)$ is very close to the exact relationship.

The asymptotic distribution of range for any value of x_o is based on the sums

$$S_n(x_o) = S_n - n\,x_o \qquad 8.11$$

where S_n is the cumulative sum of x, either positive or negative, and the range is

$$R_n(x_o) = S_n^+(x_o) - S_n^-(x_o) \qquad 8.12$$

with $S_n^+(x_o)$ the maximum positive cumulative sum and $S_n^-(x_o)$ the minimum negative cumulative sum of $S_n(x_o)$ for $n = 0, 1, 2, \ldots$. The value $S_n^+(x_o)$ takes a position in i_1, and $S_n^-(x_o)$ in i_2, when i is between 0 and n. These values and positions may be different from the values and positions of S_n^+ and S_n^- for $x_o = 0$. The approximate expected value of the asymptotic distribution of range $R_n(x_o)$ is

$$E[R_n(x_o)] \simeq 1.6\sqrt{n} + n\,|x_o| \qquad 8.13$$

where the last term is always taken positive. The asymptotic variance is the same as in eq. 8.2.

H. E. Hurst was the first [1, 2, 3] to develop the expression for the expected value (asymptotic, however) of the adjusted range as

$$E[R_n(\bar{x}_n)] = \sqrt{\frac{\pi n}{2}} \simeq 1.25\sqrt{n} \qquad 8.14$$

W. Feller [4] employed other means to develop the asymptotic mean and variance of the adjusted range of (0, 1, 0) - variable. The asymptotic mean is the same as in eq. 8.14 and the asymptotic variance is

$$S_n^2 = \operatorname{var}[R_n(\bar{x}_n)] = \left(\frac{\pi^2}{6} - \frac{\pi}{2}\right)n \simeq 0.07414\,n \simeq 0.074\,n \qquad 8.15$$

Equations 8.14 and 8.15 are developed as approximations by W. Feller using Doob's approach to the Kolmogorov-Smirnov theorem. This approach yields the distribution of the adjusted range for the continuously changing sum, S_t.

The asymptotic expected value of $S_n^+(\bar{X}_n)$ for (0, 1, 0) - variable is half of a value given by eq. 8.14, or

$$E[S_n^+(\bar{x}_n)] = \sqrt{\frac{\pi n}{8}} \simeq 0.625\sqrt{n} \qquad 8.16$$

The coefficients of variation for both the range R_n and the adjusted range $R_n(\bar{x}_n)$ are constants for asymptotic distributions, namely,

$$C_v[R_n] = \frac{\sqrt{\text{var } R_n}}{\bar{R}_n} = \sqrt{\frac{\pi \ln 2}{2}} - 1 \simeq 0.292; \qquad 8.17$$

and

$$C_v[R_n(\bar{x}_n)] = \frac{\sqrt{\text{var } R_n(\bar{x}_n)}}{\bar{R}_n(\bar{x}_n)} = \sqrt{\frac{\pi}{3}} - 1 \simeq 0.213. \qquad 8.18$$

Equations 8.17 and 8.18 show that the exact C_v - values of range and adjusted range have horizontal asymptotics. The comparison of eqs. 8.2 and 8.15 shows that the adjusted range $R_n(\bar{x}_n)$ has the advantage of a greater sampling stability than the range R_n. Equations 8.1 and 8.14 show that the difference between the asymptotic means of range and adjusted range is $\Delta R_n = 0.35\sqrt{n}$. These equations also indicate that the difference of asymptotic variances given by eqs. 8.2 and 8.15 is $\Delta \text{var } R_n = 0.144\,n$. The use of asymptotic values of mean and variance for surplus, deficit, range, adjusted surplus, adjusted range and adjusted deficit is useful because they may be considered as good approximations for large values of n.

3. Exact means of surplus and range for (0, 1, 0) - variable. The expected value of R_n for a given n of (0, 1, 0) - variable was determined by A. A. Anis and E. H. Lloyd [5] in 1953 as

$$E(R_n) = \sqrt{\frac{2}{\pi}} \sum_{i=1}^{i=n} i^{-1/2} \qquad 8.19$$

As $\bar{R}_n = \bar{S}_n^+$ for a (0, 1, 0) - variable, then,

$$E(S_n^+) = \sqrt{\frac{1}{2\pi}} \sum_{i=1}^{i=n} i^{-1/2} \qquad 8.20$$

For n = 1, eq. 8.19 gives $E(R_1) = \sqrt{2/\pi}$. For a large n the expression $\sum_{i=1}^{n} i^{-1/2}$ becomes approximately $2\sqrt{n}$, and the asymptotic mean of range is $2\sqrt{n}\sqrt{2/\pi}$ which is in agreement with Feller's results [4] and eq. 8.1. Equation 8.1 for n = 1 gives $E(R_1) = 2\sqrt{\frac{2}{\pi}}$, which is twice the value of eq. 8.19.

4. Comparison of various expressions and methods of computing means of range and adjusted range. Figure 8.2 gives the comparison between $E(R_n)$ values computed by: (1) the formula of asymptotic mean, eq. 8.1; (2) the formula for exact values, eq. 8.19; and, (3) means obtained by the data generation method from 100,000 independent numbers of (0, 1, 0) - variable, for various values of n.

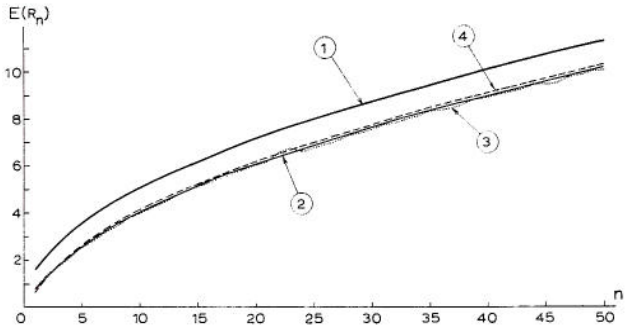


Fig. 8.2 Comparison of means of range: (1) Mean of asymptotic distribution, eq. 8.1; (2) Exact means, eq. 8.19; (3) Means determined by the data generation method, $\bar{R}_n$; and, (4) An approximation given by eq. 8.21.

For small values of n, an approximation to exact values is given in a recent Ph. D. dissertation at Colorado State University [11].

$$E(R_n) = \sqrt{\frac{8n}{\pi}} - 1 = 1.6\sqrt{n} - 1. \qquad 8.21$$

The means of eq. 8.21 are also given in fig. 8.2, line (4). The differences between these mean values are given in fig. 8.3. The absolute differences between the exact values obtained by eq. 8.19 and the values of the data generation method are within the sampling errors of this latter method. The difference of the asymptotic mean and the exact mean of range in percent of the exact mean, as given in fig. 8.4, decreases with an increase of n. They are relatively high for values of n encountered in hydrologic applications. Figure 8.5 gives the differences between various means of R_n in relation to the exact mean (eq. 8.19) in percent of the exact mean.

In conclusion, the above comparison indicates that eq. 8.1 should not be used for small values of n, even for n as large as 25 - 30, because the error

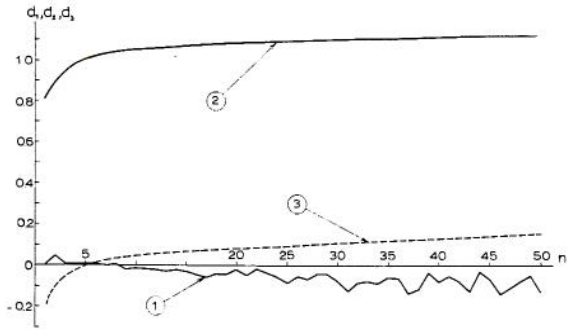


Fig. 8.3 Differences of various means of the range: (1) $d_1 = \bar{R}_n - \frac{2}{\pi} \sum_{i=1}^{n} i^{-1/2}$ or the data generation means minus exact means (eq. 8.19); (2) $d_2 = 1.6\sqrt{n} - \frac{2}{\pi} \sum_{i=1}^{n} i^{-1/2}$, or asymptotic mean (eq. 8.1) minus exact mean (eq. 8.19; and (3) $d_3 = (1.6\sqrt{n} - 1) - \frac{2}{\pi} \sum_{i=1}^{n} i^{-1/2}$, or an approximation (eq. 8.21) minus exact means (eq. 8.19).

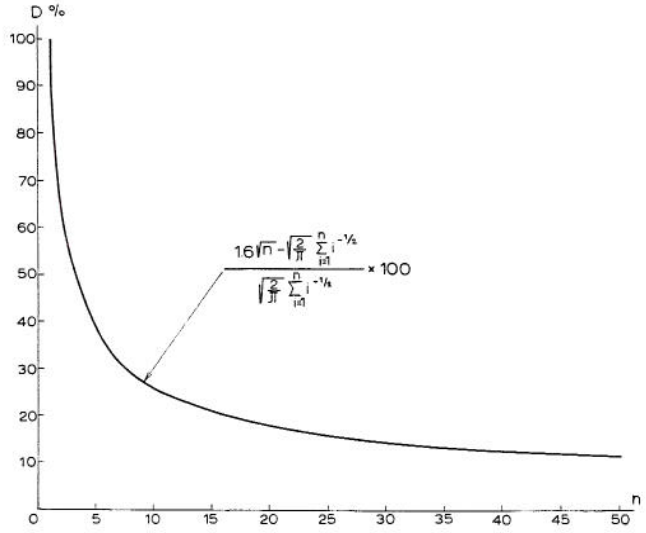


Fig. 8.4 The relative difference, D in %, of the asymptotic and exact means of range.

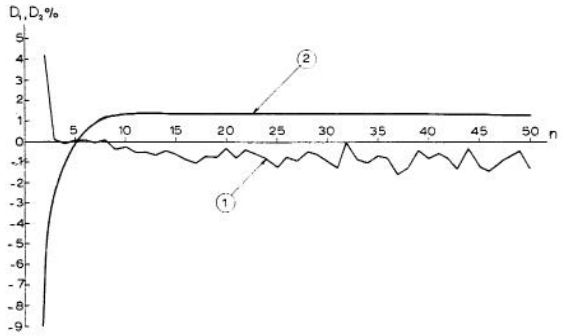


Fig. 8.5 The relative difference of ranges: (1) The difference of mean obtained by the data generation method and the exact mean (eq. 8.19) to the exact mean, D_1 in %; (2) The difference of approximate mean (eq. 8.21) and the exact mean (eq. 8.19) to the exact mean, D_2 in %.

at these n values is about 15 percent. Equation 8.21 does not have any advantage when comparing it to the exact mean of eq. 8.19. There is one exception to the foregoing statement and that is that the approximate $\overline{R}_n$ values may be computed readily from eq. 8.21 for a given n, while the derivation of $\overline{R}_n$ from eq. 8.19 is based on the computation of all previous $\overline{R}_i$ values, with i = 1, 2, ... , n - 1.

To the writer's knowledge, an expression for the exact means of adjusted range, $\overline{R}_n(\overline{x}_n)$, is not available in the literature. Therefore, fig. 8.6 gives a comparison between the following four curves of the expected mean of adjusted range, computed by: (1) expression for the asymptotic mean, eq. 8.13; (2) mean obtained by the data generation method from 100,000 independent normal numbers; (3) first approximation to the means obtained by the data generation method

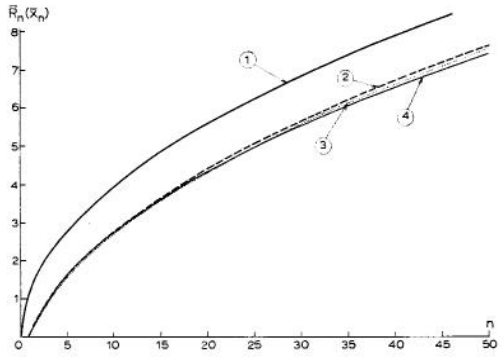


Fig. 8.6 Comparison of means of adjusted range: (1) mean of asymptotic distribution, eq. 8.13; (2) mean determined by the data generation method (100,000 independent numbers); (3) means of the approximation, $\overline{R}_n(\overline{x}_n) = 1.25\ (\sqrt{n} - 1)$, eq. 8.22; and (4) means of the approximation

$$\overline{R}_n(\overline{x}_n) = \frac{\sqrt{\pi}}{3} \sum_{i=1}^{n} (i-1)^{-1/2}, \text{ eq. 8.23.}$$

in the form

$$E[R_n(\overline{x}_n)] = 1.25\ (\sqrt{n} - 1) \qquad 8.22$$

and (4) second approximation to means obtained by the data generation method in the form

$$E[R_n(\overline{x}_n)] = \frac{\sqrt{\pi}}{3} \sum_{i=1}^{n} (i-1)^{-1/2} \qquad 8.23$$

For n = 1 these two equations give $\overline{R}_1(\overline{x}_1) = 0$ which satisfies the conditions of adjusted range that $\overline{R}_1 = 0$.

Figure 8.7 gives the following relative differences: (1) Asymptotic mean minus the mean obtained by the data generation method in percent of this latter value; (2) Mean of eq. 8.22 minus the mean obtained by the data generation method in percent of this latter value; and (3) Mean of eq. 8.23 minus the mean obtained by the data generation method in percent of this latter value.

Equations 8.22 and 8.23 fit the means obtained by the data generation method relatively well, particularly for n > 10. For n = 1 - 10 the largest relative differences are not greater than 8 percent and 5 percent for n = 2, respectively for eqs. 8.22 and 8.23.

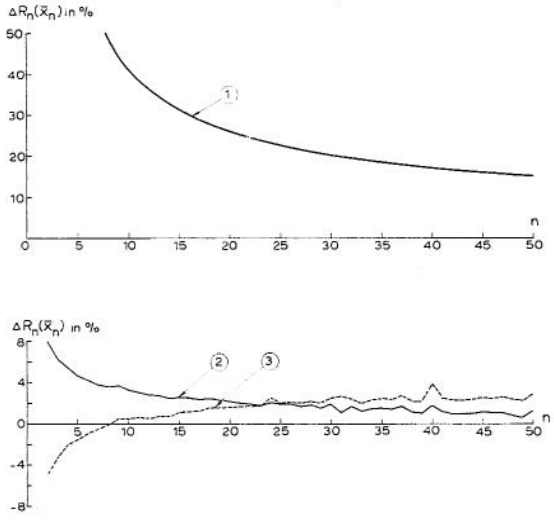


Fig. 8.7 Relative differences of means of adjusted range: (1) Asymptotic mean minus the mean obtained by the data generation method, in percent of this latter value; (2) Mean of eq. 8.22 minus the mean obtained by the data generation method, in percent of this latter value; and (3) Mean of eq. 8.23 minus the mean obtained by the data generation method, in percent of this latter value.

5. Exact variances of surplus and range for (0, 1, 0) - varaible. A. Anis [6] gives the exact second moment of S_n^+ about zero as

$$\mu_2'(S_n^+) = \frac{1}{2}(n+1) + \frac{1}{2\pi}\sum_{i=1}^{n-2}\sum_{j=1}^{i}\left[j(i-j+1)\right]^{1/2} \qquad 8.24$$

As n ordinates have n - 1 intervals, n in eq. 7.1 of ref. 6 is replaced by n - 1. Using the expression for the exact mean of S_n^+, eq. 8.20, the variance of S_n^+, becomes

$$\text{var } S_n^+ = \frac{n}{2} + \frac{1}{2\pi}\sum_{i=1}^{n-3}\sum_{j=1}^{i}\left[(j(i-j+1)\right]^{1/2} - \frac{1}{2\pi}\left(\sum_{i=1}^{n} i^{-1/2}\right)^2 \qquad 8.25$$

Using this expression and eqs. 8.5 and 8.7, the variance of range becomes

$$\text{var } R_n = \frac{0.60\sqrt{n}}{\sqrt{n} - 0.37}\left[\frac{n}{2} + \frac{1}{2\pi}\sum_{i=1}^{n-3}\sum_{j=1}^{i}\left\{j(i-j+1)\right\}^{-1/2} - \frac{1}{2\pi}\left(\sum_{i=1}^{n} i^{-1/2}\right)^2\right] \qquad 8.26$$

Table 8.1 gives the computed values of: (1) Exact values of $\bar{R}_n$, eq. 8.19; (2) Exact values of Var S_n^+, eq. 8.25; and (3) Approximations to exact values of var R_n, eq. 8.26.

6. Comparison of various expressions and methods of computing variances of range and adjusted range. Figure 8.8 gives the comparison between variances of R_n computed by: (1) expression for the asymptotic variance of R_n, eq. 8.2; (2) expression for the variance of R_n, computed by eq. 8.26; and (3) variances of R_n computed by the data generation method (100,000 independent normal numbers).

TABLE 8.1

n	$\bar{R}_n$ eq. 8.19	var S_n^+ eq. 8.25	var $\bar{R}_n$ eq. 8.26
1	0.7979	0.3408	0.3246
2	1.3621	0.5361	0.4364
3	1.8228	0.6692	0.5106
4	2.2218	0.9249	0.6809
5	2.5786	1.2216	0.8783
6	2.9043	1.5387	1.0876
7	3.2059	1.8671	1.3025
8	3.4880	2.2028	1.5206
9	3.7540	2.5438	1.7410
10	4.0063	2.8886	1.9628
11	4.2582	3.2360	2.1856
12	4.4886	3.5855	2.4087
13	4.7099	3.9368	2.6321
14	4.9232	4.2895	2.8559
15	5.1292	4.6433	3.0804
16	5.3287	4.9980	3.3047
17	5.5222	5.3536	3.5291
18	5.7103	5.7099	3.7531
19	5.8933	6.0667	3.9773
20	6.0717	6.4241	4.2020
21	6.2458	6.7819	4.4265
22	6.4159	7.1401	4.6511
23	6.5823	7.4986	4.8756
24	6.7451	7.8575	5.0995
25	6.9047	8.2167	5.3236
26	7.0612	8.5760	5.5478
27	7.2148	8.8358	5.7725
28	7.3655	9.2995	5.9965
29	7.5137	9.6555	6.2210
30	7.6594	10.0156	6.4450
31	7.8027	10.3760	6.4450
32	7.9438	10.7364	6.8928
33	8.0827	11.0968	7.1164
34	8.2195	11.4575	7.3408
35	8.3543	11.8184	7.5638
36	8.4873	12.1792	7.7874
37	8.6185	12.5403	8.0120
38	8.7479	12.9014	8.2350
39	8.8757	13.2627	8.4590
40	9.0018	13.6241	8.6826
41	9.1265	13.9854	8.9059
42	9.2496	14.3469	9.1289
43	9.3713	14.7083	9.3530
44	9.4916	15.0697	9.5753
45	9.6105	15.4312	9.7988
46	9.7282	15.7929	10.0222
47	9.8446	16.1545	10.2452
48	9.9597	16.5165	10.4682
49	10.0737	16.8782	10.6923
50	10.1865	17.2402	10.9148

Figure 8.9 gives: (a) differences between the variance of eq. 8.2 and var R_n obtained by the data generation method, in percent of this latter value; and (b) differences between the variance of eq. 8.26 and var R_n obtained by the data generation method, in percent of this latter value. This figure shows that neither the variance obtained by eq. 8.2, nor the variance obtained by eq. 8.26 fit closely the variances obtained by the data generation method for small values of n (2 - 15).

Figure 8.10 gives the comparison between variances of adjusted range computed by: (1) expression for the asymptotic variance of adjusted range, eq. 8.15; and (2) variance of adjusted range obtained by the data generation method. No expression was available for the exact values or approximation to the exact values of the variance of adjusted surplus or adjusted range.

Figure 8.11 gives the differences between the asymptotic variance of range, eq. 8.15, and the variance of adjusted range obtained by the data generation method in percent of this latter value, for $n \geq 2$. A comparison of figs. 8.10 and 8.11 show that the asymptotic variance of adjusted range does not depart significantly from the variance obtained by the data

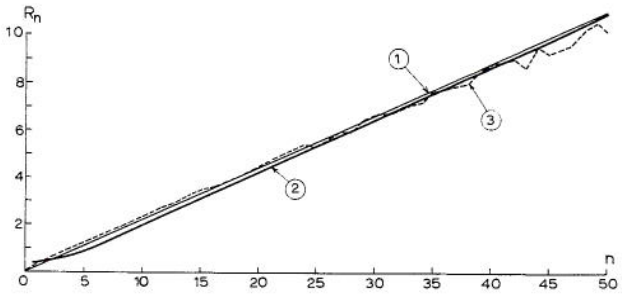


Fig. 8.8 Comparison of variances of range computed in the following ways: (1) asymptotic variance of range, eq. 8.2; (2) approximation to exact values, eq. 8.26; and (3) values obtained by the data generation method.

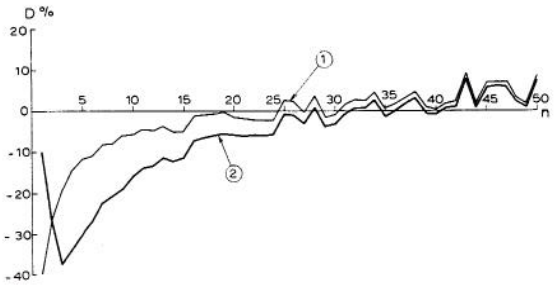


Fig. 8.9 Differences of variances of range computed in the following ways: (1) asymptotic variance of range (eq. 8.2) minus the variance obtained by the data generation method in relation to this latter method; and (2) approximation to exact variance of range (eq. 8.26) minus the variance obtained by the data generation method in relation to this latter value.

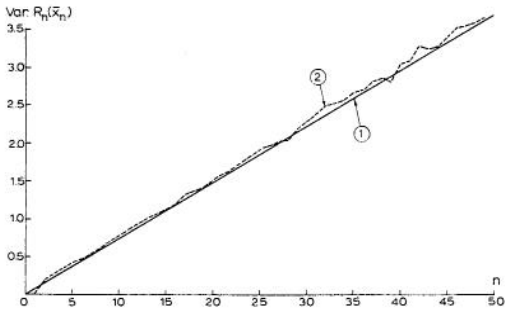


Fig. 8.10 Comparison of variances of adjusted range: (1) asymptotic variance, computed by eq. 8.15; and (2) values obtained by the data generation method.

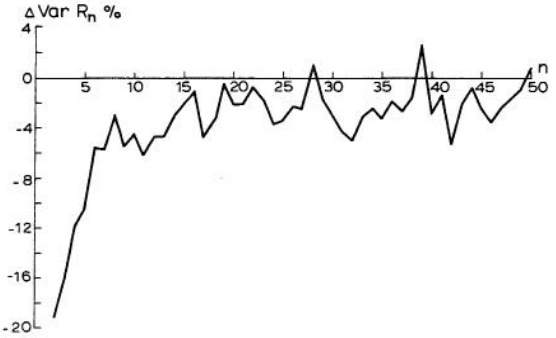


Fig. 8.11 Difference of asymptotic variance of adjusted range (eq. 8.15) and the variance of adjusted range obtained by the data generation method, in percent of this latter value.

generation method. Therefore, the asymptotic variance can be used in practical cases, except for very small values of n, such as n = 2 - 6.

The approximation var $R_n(\bar{X}_n) = 0.074\,n + 0.05$ approaches better the variance of adjusted range obtained by the data generation method than eq. 8.15.

7. Skewness and excess coefficients of surplus, range, adjusted surplus and adjusted range. Figure 8.12, upper graph, gives the skewness and excess coefficients as functions of n for the independent standard normal variable (0, 1, 0) for both the surplus and range. This figure presents these parameters as computed by the data generation method

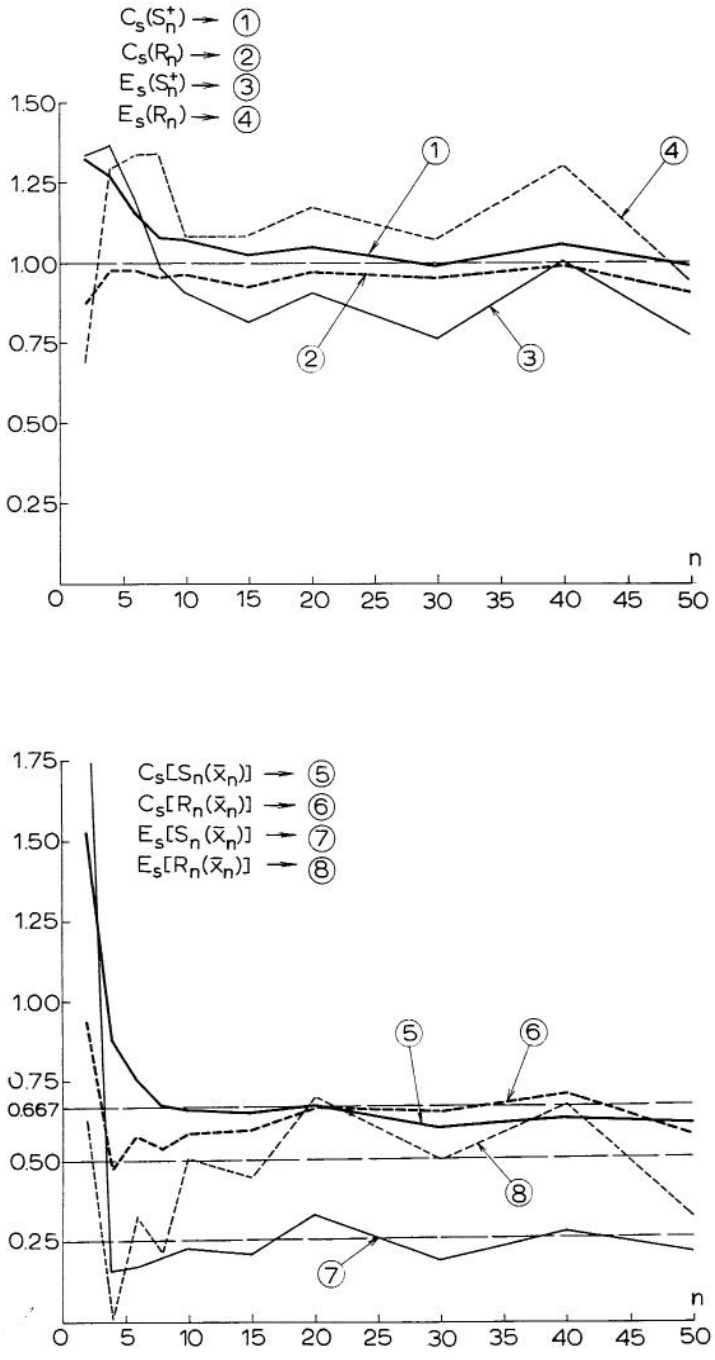


Fig. 8.12 Skewness and excess coefficients of surplus, range, adjusted surplus and adjusted range obtained by the data generation method for (0, 1, 0) - variable (100,000 independent normal numbers): (1) Skewness coefficient of surplus; (2) skewness coefficient of range; (3) excess coefficient of surplus; (4) excess coefficient of range; (5) skewness coefficient of adjusted surplus; (6) skewness coefficient of adjusted range; (7) excess coefficient of adjusted surplus; and (8) excess coefficient of adjusted range.

(100,000 independent normal numbers) as they change with n. Figure 8.12 leads to the following conclusions:

(a) Even a sample of 100,000 is not sufficient to produce reliable values of skewness coefficients for the surplus and range of (0, 1, 0) - variable. In the case n = 5, there was a sample of 20,000 for either the surplus or the range.

(b) Skewness coefficients of both the surplus and the range seems to converge to the asymptotic value of about $C_s(S_n^+) = C_s(R_n) = 1.00$ for an increase in n. For a small n the skewness coefficient of surplus converges slower to $C_s = 1.0$ than the skewness coefficient of range. $C_s(S_n^+)$ seems to converge to unity from the above and $C_s(R_n)$ from below of the asymptotic value of unity.

(c) Excess coefficients of surplus and range are less reliable than the skewness coefficients for this sample of 100,000 random numbers. The fact is true because the values $E_s(S_n^+)$ and $E_s(R_n)$ fluctuate more around a smooth curve imagined to be drawn for the computed values than is the case for $C_s(S_n^+)$ or $C_s(R_n)$.

(d) It may be concluded that the excess coefficient for both the surplus and the range converge to the asymptotic value of $E_s = 1.00$ with an increase in n though with a slower convergency than C_s; the $E_s(S_n^+)$ seems to converge to $E_s = 1.00$ from below and $E_s(R_n)$ from above, opposite to the direction of convergence for the skewness coefficients.

(e) A much larger sample than 100,000 is necessary to obtain smoother curves of C_s - and E_s - coefficients.

(f) The use of values $C_s = 1.0$ and $E_s = 1.0$ for deriving distributions of surplus and range of (0, 1, 0) - variable may be considered as reasonable approximations, even for n as small as 5 - 10.

By using the best available values for $\overline{S}_n^+$, var S_n^+, $C_s(S_n^+) = 1$, $E_s(S_n^+) = 1$; $\overline{R}_n$, var R_n, $C_s(R_n) = 1$ and $E_s(R_n)$, it is possible to obtain the approximate distributions of S_n^+ and R_n. It suffices to use the procedures and criteria for selecting fitting functions when the first three or four moments, or their corresponding parameters of mean, variance, skewness and excess coefficients, are available.

Figure 8.12, lower graph, gives the skewness and excess coefficients for the adjusted surplus and the adjusted range, obtained by the data generation method, as they change with n. This figure leads to the following conclusions:

(a) Skewness coefficients $C_s[S_n(\overline{X}_n)]$ and $C_s[R_n(\overline{X}_n)]$ seem to converge to an approximate asymptotic value of $C_s = 2/3$.

(b) The convergence trends of $E_s[S_n(\overline{X}_n)]$ and $E_s[R_n(\overline{X}_n)]$ are not as clearly indicated on fig. 8.12 as are the skewness coefficients. This occurs because the sampling error associated with E is larger than that associated with C_s. However, it seems that $E_s[S_n(\overline{X}_n)]$ converges to the approximate value of $E_s = 1/4$ and $E_s[R_n(\overline{X}_n)]$ to the approximate value of $E_s = 1/2$.

8. Exact distributions of surplus and range for (0, 1, 0) - variable. The exact distributions of surplus and range for n = 2 and n = 3 are computed from the following equations given in Chapter VII: (1) eq. 7.6 for the surplus and for n = 2; (2) eq. 7.10 for the range for n = 2; (3) eq. 7.15 for the surplus and n = 3; and (4) eq. 7.20 for the range and n = 3. The only difference between the above equations and the equations used for the computation of exact distributions of surplus and range of (0, 1, 0) - variable is that the symmetry reduces the number of integrals in equations 7.6, 7.10, 7.15 and 7.20.

Figure 8.13 gives the exact probability densities and distributions for S_2^+ and R_2 of (0, 1, 0) variable. For densities, the curves for the basic two integrals as parts of probability densities of R_2 are also given. The probability mass of $S_2^+ = 0$ is shown on the S_2^+ - probability distribution.

Figure 8.14 gives the exact probability densities and distributions for S_3^+ and R_3 of (0, 1, 0) variable. For densities, the basic integrals as parts of probability densities of S_3^+ and R_3 are also given. The probability mass of $S_3^+ = 0$ is shown on the S_3^+ probability distribution.

The above distributions of S_2^+, R_2, S_3^+ and R_3 are obtained by the finite differences method of integrating the exact equations of distributions. The differences were $\Delta x = \Delta S_2^+ = \Delta R_2 = \Delta S_3^+ = \Delta R_3 = 0.10$.

9. Distributions of surplus and range of (0, 1, 0) - variable, obtained by the data generation method. The distributions of surplus and range and their parameters are computed for (0, 1, 0) - variable from 100,000 independent normal numbers for the following n - values: 2, 4, 6, 8, 10, 15, 20, 30, 40 and 50. Both the surplus and deficit were computed. Surplus and deficit for the population are equal because of distribution symmetry of (0, 1, 0) - variable. Sampling errors of the data generation method make for some small differences in computed values of surplus and deficit. The average values of the two are plotted in fig. 8.15. This figure gives the distributions of surplus with the upper most graph for $\rho = 0$. Figure 8.16 gives the probability mass of $F(S_n^+ = 0)$ for $S_n^+ = 0$ as a function of n, with the lowest curve for $\rho = 0$. Values presented in these two figures are averages between those obtained for surplus and deficit, respectively. Figure 8.17 gives distributions of range, R_n, with the upper most graph for $\rho = 0$. Figures 8.18, 8.19 and 8.20 give the distributions of adjusted surplus, probability mass of adjusted surplus for $S_n^+ = 0$, and distributions of adjusted range, respectively, for various values of n and for $\rho = 0$ on the proper graphs.

10. Properties of dependent variables. The distributions of surplus, deficit and range of dependent variables may be divided into the following two

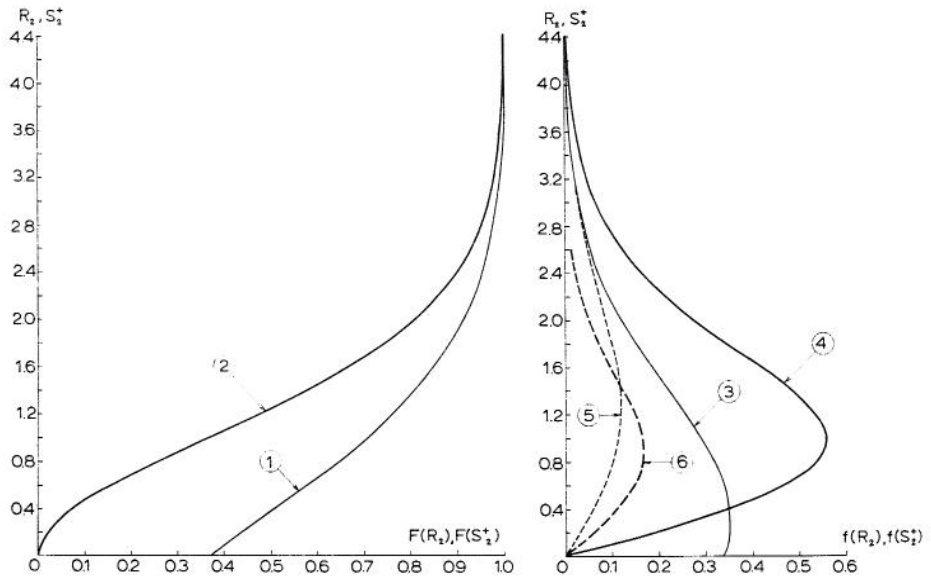


Fig. 8.13 Exact distributions for surplus and range for n = 2 of the independent standard normal variable (0, 1, 0), obtained by the finite difference method of integration of exact equations: (1) Probability distribution of surplus, S_2^+; (2) Probability distribution of range, R_2; (3) Probability density of S_2^+; (4) Probability density of R_2; (5) and (6) Component densities, each multiplied by two and summed up gives the probability density of R_2.

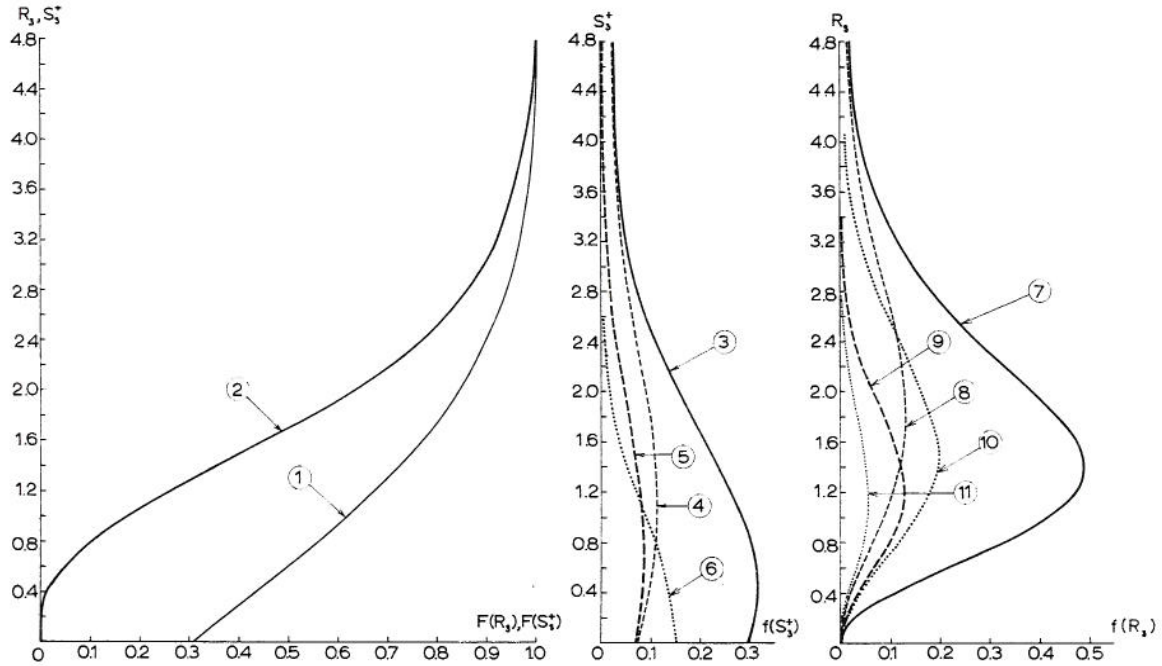


Fig. 8.14 Exact distributions for surplus and range for n = 3 of the independent standard normal variable (0, 1, 0), obtained by the finite difference method of integration of exact equations: (1) Probability distribution of surplus S_3^+; (2) Probability distribution of range, R_3; (3) Probability density of S_3^+; (4), (5) and (6) Component probability densities, when summed up give the total probability density of S_3^+; (7) Probability density of R_3; (8), (9), (10) and (11) Component probability densities, when summed up give the total probability density of R_3.

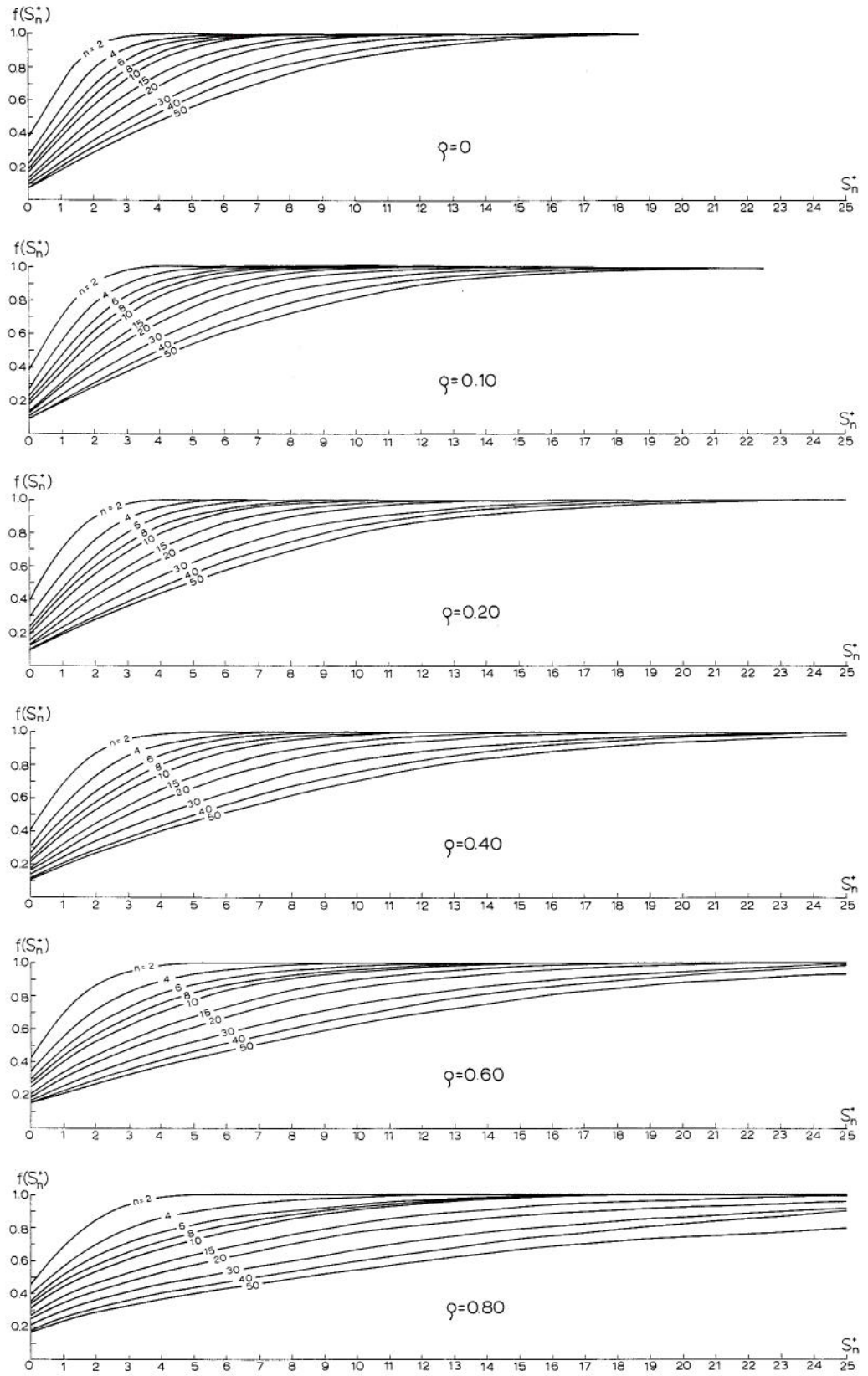


Fig. 8.15 Distributions of surplus, S_n^+, of standard normal variables for various values of n and the following values of ρ, in the case of Markov first order linear dependence: 0, 0.10, 0.20, 0.40, 0.60 and 0.80, obtained by the data generation method.

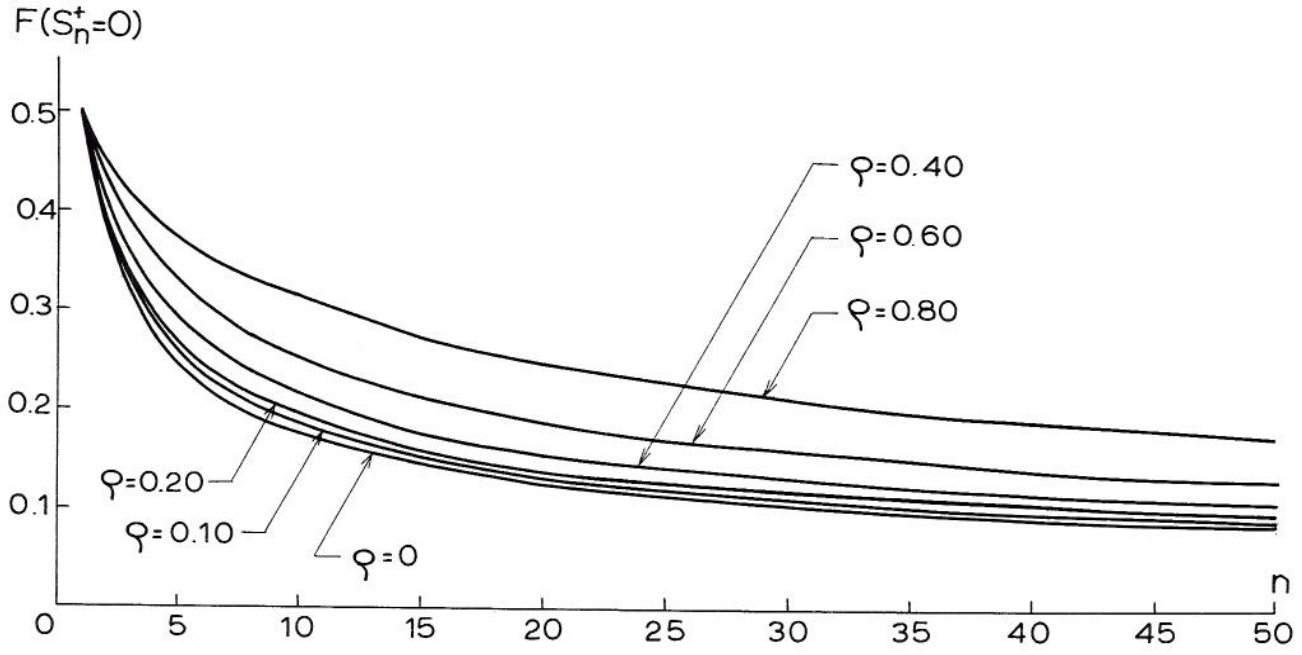


Fig. 8.16 Probability mass for surplus being zero, $F(S_n^+ = 0)$, of standard normal variables for various values of n, and the following values of ρ, in the case of Markov first order linear dependence: 0, 0.10, 0.20, 0.40, 0.60 and 0.80, obtained by the data generation method.

categories: (1) The mathematical model of dependence of a variable may be expressed as a function of an independent variable. Furthermore, the distributions of surplus, deficit and range of this dependent variable are related to distributions of surplus, deficit and range of the independent variable through the parameters of the dependence model. As an example, the dependent variable X may be expressed as

$X_i = a_1 X_{i-1} + a_2 X_{i-2} + \epsilon_i$, with ϵ_i an independent variable, and a_1 and a_2 the dependence parameters. Then the distributions of surplus, deficit and range of X are related to distributions of surplus, deficit and range of ϵ_i, respectively, via parameters a_1 and a_2. The same procedure is valid for the statistical parameters of distributions and for the mathematical dependence models of surplus, deficit and range of dependent variables; (2) Mathematical dependence model of a variable is either very complex or is given in an empirical form so that the above procedure, under (1), cannot be applied. As a result, it is necessary to obtain the distributions and mathematical models of dependence for surplus, deficit and range by a direct method of computation.

The general mathematical dependence model for those stochastic variables in hydrology, which are transformed by storage effects, is of the general moving average type

$$X_i = b_o \epsilon_i + b_1 \epsilon_{i-1} + \cdots + b_m \epsilon_{i-m} \qquad 8.27$$

where $\epsilon_i, \epsilon_{i-1}, \cdots, \epsilon_{i-m}$ represent the values of an independent variable at intervals, i, i-1, ... , i-m, either concurrent or previous to the time interval i during which the value X_i of the dependent variable occurs.

The usual characteristics of b_j coefficients in the case of water storage effects are: (a) their sum is unity (but not necessarily); (b) they are monotonically decreasing; (c) they are positive; and (d) they are either finite or infinite in number (but in the latter case they can be approximated by a finite number of coefficients for all practical purposes).

Let ϵ in eq. 8.27 be an independent variable, with the mean μ_ϵ and variance σ_ϵ^2. For a very large sample $\overline{X} = \overline{\epsilon}$, or $\mu_x = \mu_\epsilon$, because $\sum_{j=o}^{m} b_j = 1$. This means that for a sufficient time period the average output from a storage facility is equal to the average storage input. As the expected values of all crossproducts $\Delta\epsilon_p \cdot \Delta\epsilon_s$ with $p \neq s$ are zeros, because ϵ_i is an independent variable, then

$$\sigma_x^2 = \sigma_\epsilon^2 \sum_{j=o}^{m} b_j^2 \qquad 8.28$$

Denote $D^2 = 1 / \sum_{j=o}^{m} b_j^2$, with D greater than unity because $\sum_{j=o}^{m} b_j^2 < 1$. D is equal to unity either when m=o, or when $\sum_{j=o}^{m} b_j^2 = 1$. As $D^2 > 1$, then $\sigma_x^2 < \sigma_\epsilon^2$. The condition for D = 1, or that $\sum_{j=o}^{m} b_j^2 = 1$, can be obtained only when the first of the

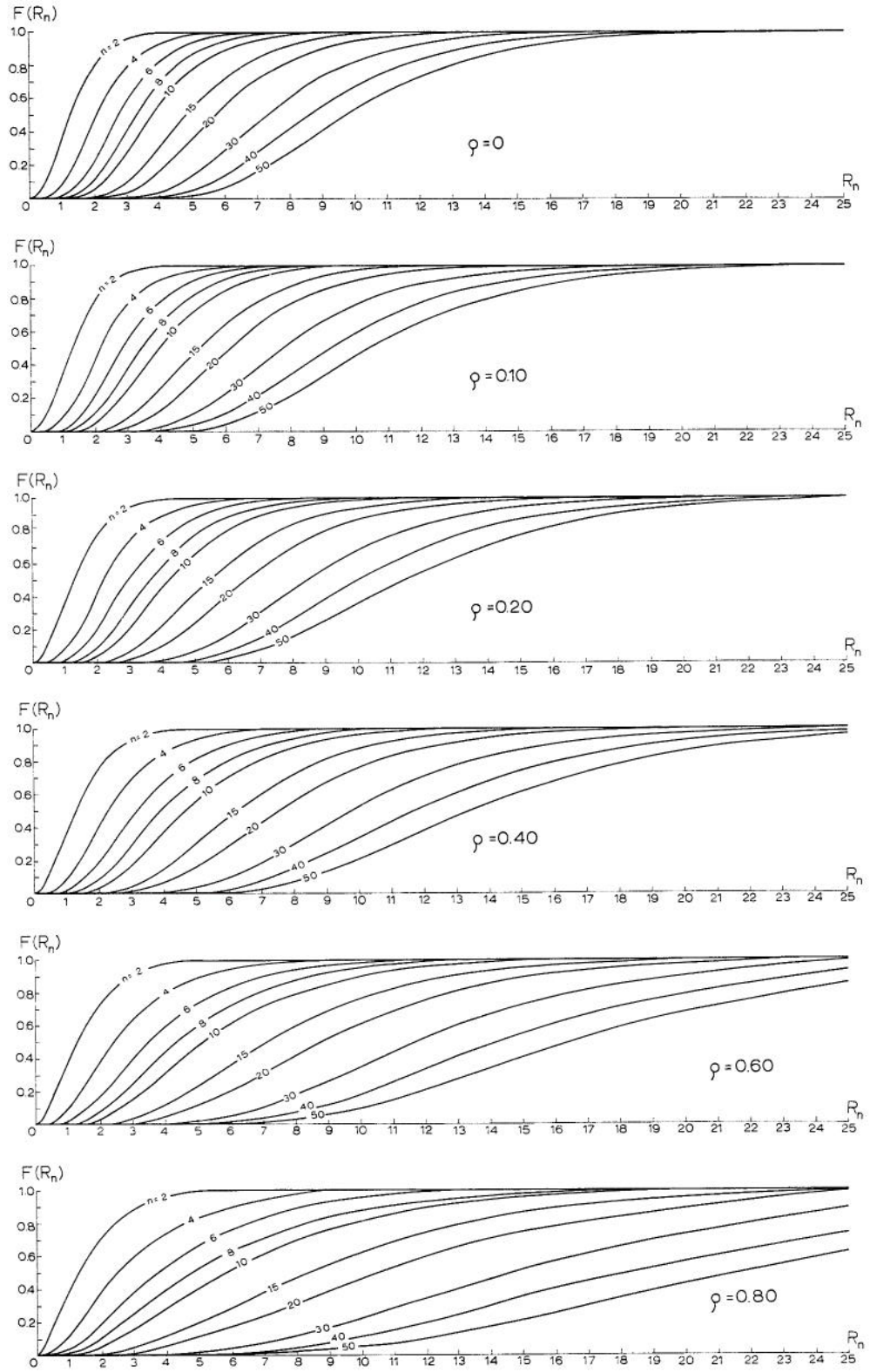


Fig. 8.17 Distributions of range, R_n, of standard normal variables for various values of n and the following values of ρ, in the case of Markov first order linear dependence: 0, 0.10, 0.20, 0.40, 0.60 and 0.80, obtained by the data generation method.

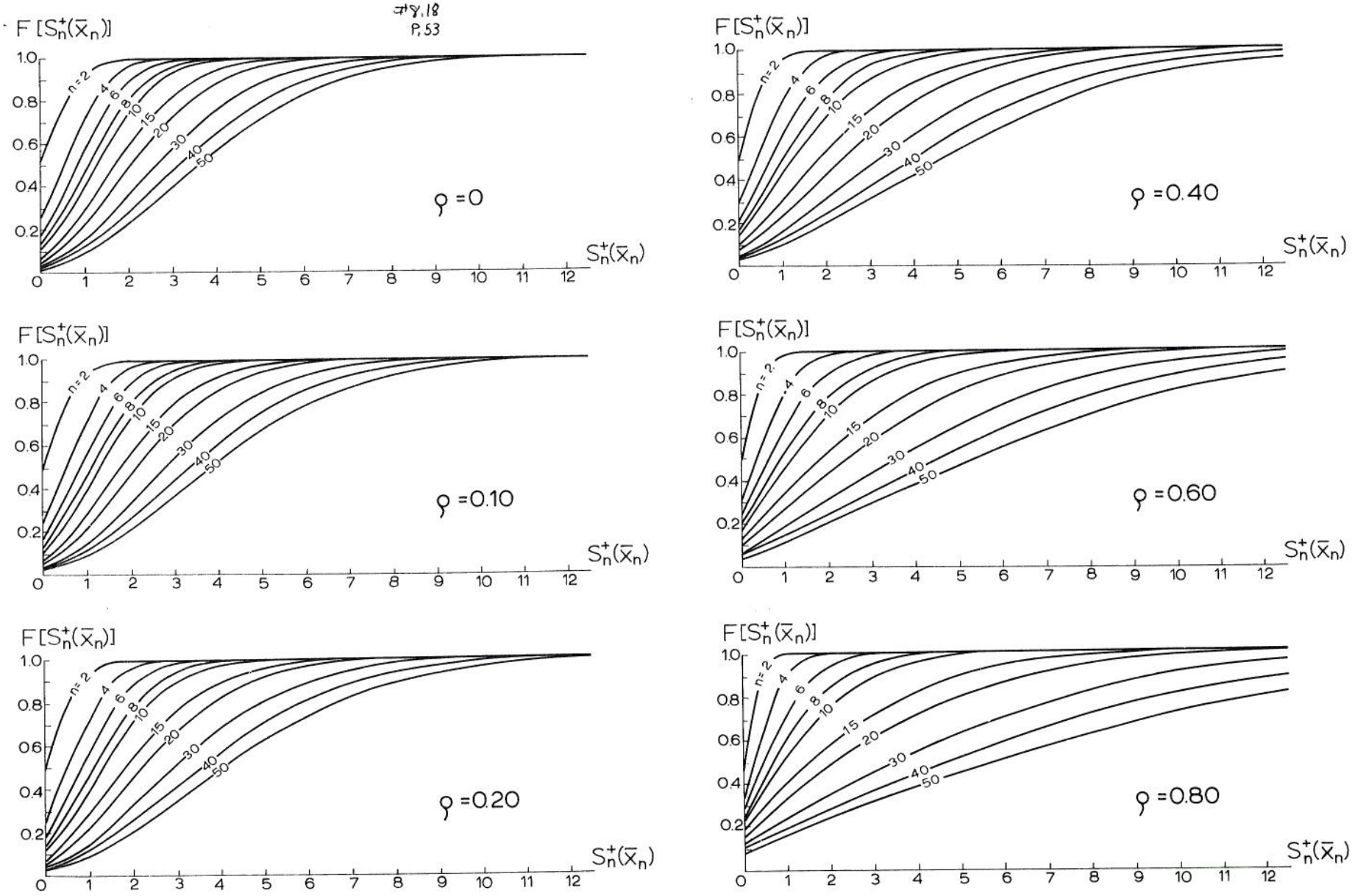


Fig. 8.18 Distributions of adjusted surplus, $S_n^+(\bar{X}_n)$, of standard normal variables for various values of n and the following values of ρ, in the case of Markov first order linear dependence: 0, 0.10, 0.20, 0.40, 0.60, 0.80, obtained by the data generation method.

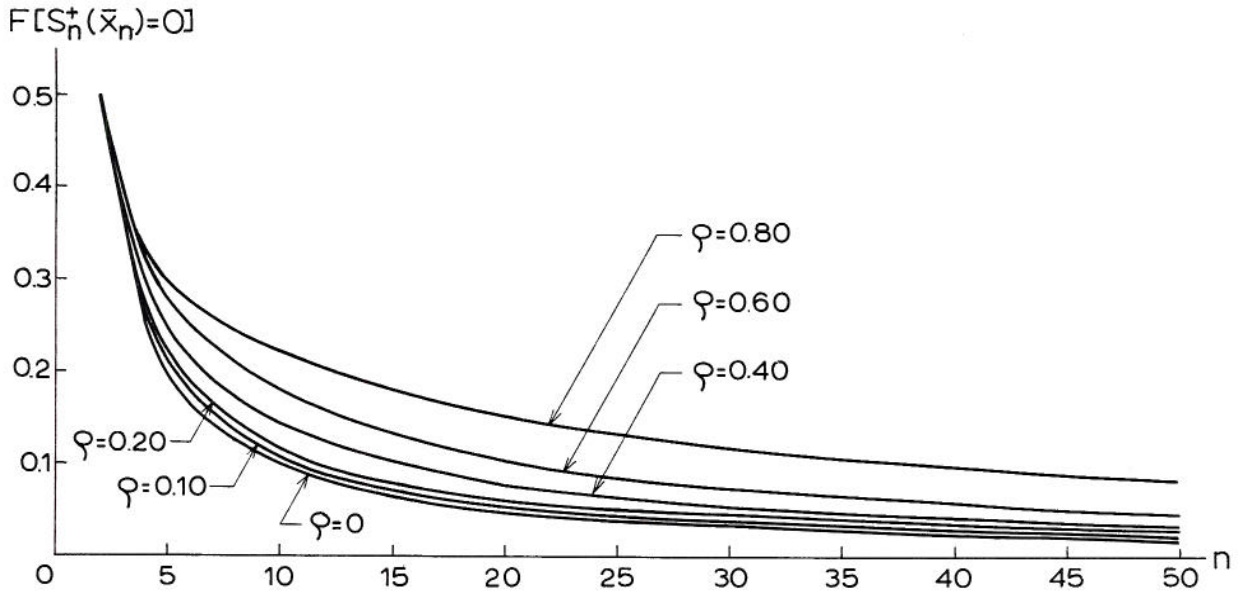


Fig. 8.19 Probability mass for adjusted surplus being zero, $F[S_n^+(\bar{X}_n)] = 0$, of standard normal variables for various values of n, and the following values of ρ, in the case of Markov first order linear dependence: 0, 0.10, 0.20, 0.40, 0.60 and 0.80, obtained by the data generation method.

previously mentioned four conditions, or $\sum_{j=o}^{m} b_j = 1$, is removed.

The relationship of σ_ϵ^2 and σ_x^2 depends only on b_j - coefficients. The factor D may be derived from the properties of X - series. For a X - series generated from the independent ϵ - series by eq. 8.27, with b_j - coefficients satisfying the four mentioned conditions, the first m autocorrelation coefficients are positive. The ρ_k - autocorrelation coefficient is

$$\rho_k = \frac{1}{\sigma_x^2 (N - k)} \sum_{1}^{N-k} (X_i - \mu_x)(X_{i+k} - \mu_k) \qquad 8.29$$

for $N = \infty$.

Replacing X by ϵ - values of eq. 8.27 in eq. 8.29, with the expected crossproducts of ϵ - series being zeros, and using eq. 8.28, then

$$\rho_k = \frac{\sum_{j=o}^{m-k} b_j \, b_{j+k}}{\sum_{j=o}^{m} b_j^2} \qquad 8.30$$

This relation was given by Cramer in 1933.

As there are m positive values of ρ_k for the mathematical model of eq. 8.27, with the four conditions for b_j values, and as $\sum_{o}^{m} b_j = 1$, there are (m + 1) equations with (m + 1) unknowns: b_o to b_m. Theoretically, it should be possible to determine all b_j - values from first m values of ρ_k from the autocorrelogram. As soon as $m > 2$, the analytical solutions for b_j as a function of ρ_k become difficult.

It is not necessary to pass through b_j values in order to derive D for some special cases. The use of eq. 8.30 gives

$$\sum_{1}^{m} \rho_k = D^2 \, B \qquad 8.31$$

where B is the sum of all crossproducts of b_j values. As

$$1 = (b_o + b_1 + \cdots + b_m)^2 =$$

$$\sum_{o}^{m} b_j^2 + 2B = \frac{1}{D^2} + 2B \qquad 8.32$$

the square of factor D is

$$D^2 = 1 + 2 \sum_{1}^{m} \rho_k \, . \qquad 8.33$$

To obtain the factor D of a time series of eq. 8.27, it is sufficient to compute all initial positive m serial correlation coefficients r_k as estimates of ρ_k.

The use of eq. 8.33 requires computation of m values of r_k. Instead, a simple procedure, based on the models assumed for decreasing r_k values may be used.

For all b_j - coefficients equal (simple moving average model), with $j = o, \ldots, m$, the correlogram decreases linearly from ρ_1 to $\rho_m = o$.

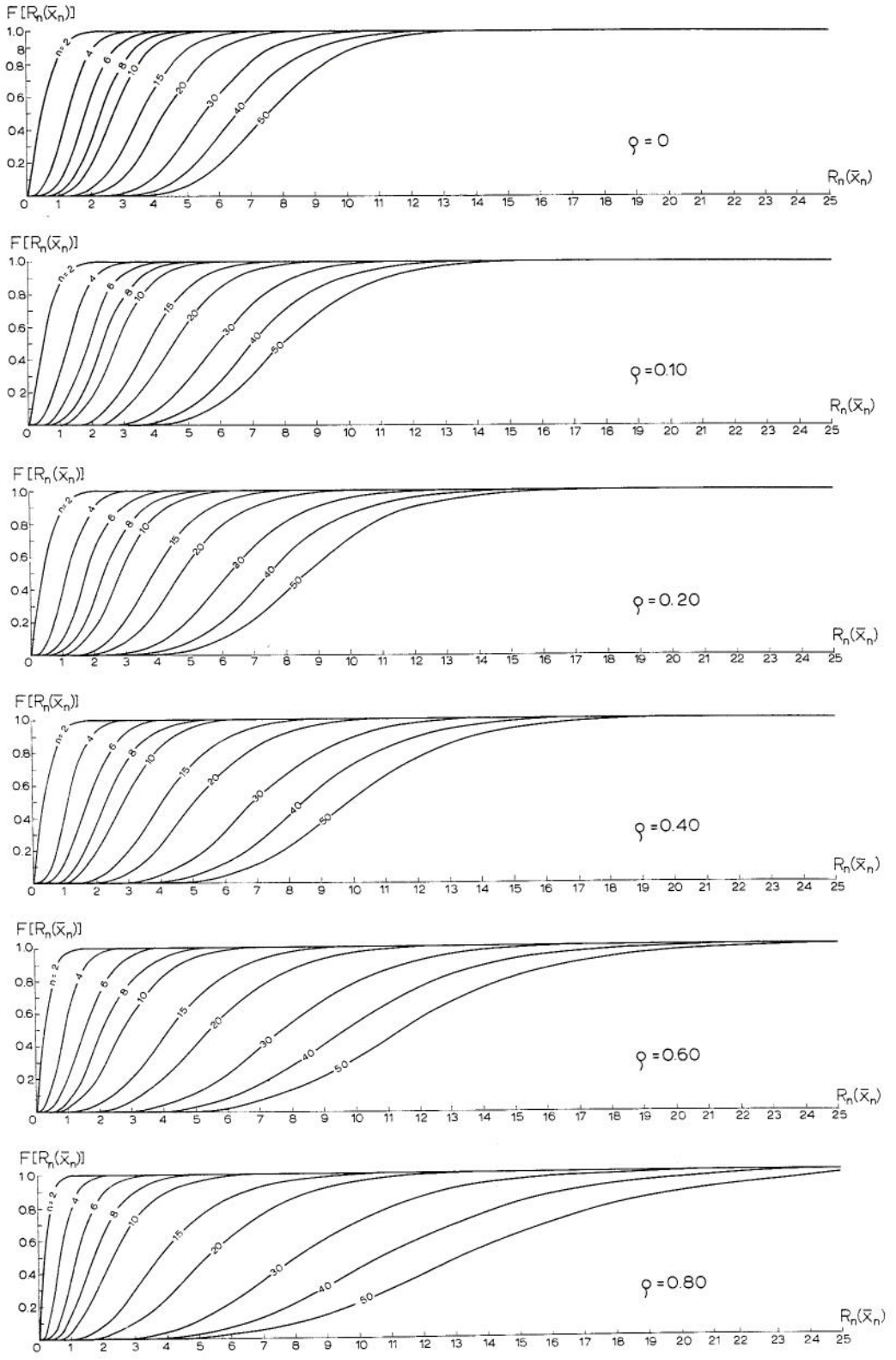


Fig. 8.20 Distributions of adjusted range, $R_n(\bar{X}_n)$, of standard normal variables for various values of n and the following values of ρ, in the case of Markov first order linear dependence: 0, 0.10, 0.20, 0.40, 0.60 and 0.80, obtained by the data generation method.

For the sum $\sum_{j=o}^{m} b_j = 1$, D^2 becomes

$$D^2 = 1 + 2 \sum_{k=o}^{m} \rho_k = 1 + m \rho_1 \qquad 8.34$$

Equation 8.33 is valid only for the case in which $\sum_{j=o}^{m} b_j = 1$. If this is not a case, another equation must be derived for every moving average model with $\sum_{j=o}^{m} b_j \neq o$. The example is the first order Markov linear model of dependence in the form

$$X_i = \rho X_{i-1} + \epsilon_i \qquad 8.35$$

Equation 8.35 may be written as

$$X_i = \epsilon_i + \rho\epsilon_{i-1} + \rho^2 \epsilon_{i-2} + \cdots + \rho^m \epsilon_{i-m} + \cdots \qquad 8.36$$

In this case,

$$\sum_{j=o}^{\infty} b_j = 1 + \rho + \rho^2 + \cdots + \rho^m + \cdots = \frac{1}{1-\rho} \qquad 8.37$$

for $\rho < 1$, which condition is always satisfied.

The factor D is given by

$$D^2 = \frac{1}{\sum_{j=o}^{\infty} b_j^2} = \frac{1}{1 + \rho^2 + \rho^4 + \cdots} = 1 - \rho^2 \qquad 8.38$$

The sum, B, of all cross products of b_j - coefficients of eq. 8.36 becomes,

$$B = \frac{\sum_{k=1}^{\infty} \rho_k}{D^2} = \frac{\frac{1}{1-\rho} - 1}{1 - \rho^2} = \frac{\rho}{(1-\rho)(1-\rho^2)} \qquad 8.39$$

Equation 8.28 gives

$$\sigma_x^2 = \frac{\sigma_\epsilon^2}{1 - \rho^2} \qquad 8.40$$

which also comes directly out of eq. 8.35.

In the case the Markov first order linear model of the following type is used

$$X_i = \rho X_{i-1} + \sqrt{1 - \rho^2}\ \epsilon_i \qquad 8.41$$

with $\sigma_x^2 = \sigma^2$, then

$$\sum_{j=o}^{\infty} b_j = \sqrt{1 - \rho^2}\,(1 + \rho + \rho^2 + \cdots) = \sqrt{\frac{1+\rho}{1-\rho}} \qquad 8.42$$

$$D^2 = \frac{1}{\sum_{j=o}^{\infty} b_j^2} = \frac{1}{(1-\rho)(1+\rho^2+\rho^4+\cdots)} = 1 \qquad 8.43$$

which also comes out of eq. 8.28.

11. Distributions of surplus, range, adjusted surplus and adjusted range of dependent normal variables. Investigation probing the effect of time dependence on the distributions of surplus, deficit and range is shown here for the standard normal but dependent variable (0, 1, ρ). Dependence is of the first order linear Markov model of the type given by eq. 8.41. Both the independent variable, ϵ, and the dependent variable, X, have mean zero, and variances unities. The only difference is the degree of dependence measured by the first autocorrelation coefficient, ρ, as the only parameter of dependence.

As the surplus and deficit have the same characteristics for any symmetrical distribution, only the surplus and range are investigated as they depend on the parameter ρ. For comparative purposes, the characteristics of surplus and range are given for $\rho = o$ of independent standard normal variable (0, 1, 0). The following properties of surplus and deficit for dependent standard normal variables, for various values of ρ, are obtained by the data generation method, by using 100,000 normal independent numbers: (1) Mean; (2) Mean range; (3) Variance of surplus; (4) Variance of range; (5) Skewness coefficient of surplus; (6) Skewness coefficient of range, all as functions of the variate n and the parameter ρ.

A clear distinction should be made for these two cases: (1) Distributions of surplus and range of a dependent standard normal variable (0, 1, ρ) are compared with the corresponding distributions of the independent standard normal variable (0, 1, 0); and (2) Distributions of surplus and range of a dependent normal variable (μ, σ, ρ) are compared with the corresponding distributions of independent standard normal variable (0, 1, 0). Because $\sigma \neq o$ for (μ, σ, ρ) - variable, the distributions of surplus and range of this variable should be first transformed by multiplying S_n^+ and R_n by σ, and dividing their probability densities by σ.

When a mathematical dependence model transforms a (0, 1, 0) - variable into a dependent variable (0, σ^2, ρ) with the variance σ^2, the resulting surplus and range are affected by two factors: (a) The effect of a standard deviation different from unity; and (b) The effect of dependence model. These two factors may produce the effects of the same or of the opposite direction.

For the dependence of Markov first order linear model, eq. 8.35, var X is given by eq. 8.40, so that var $X_i >$ var ϵ_i, and $\sigma_x = (1 - \rho^2)^{-1/2} \sigma_\epsilon$. For example, if $\rho = 0.8$, $\sigma_x = 1.333\ \sigma_\epsilon$. In this case, the variable x has larger values of surplus and range than ϵ, because both factors: the greater value of σ_x than σ_ϵ and the dependence model increase the surplus and the range of x. The case is opposite for the moving average type of dependence, given by eq. 8.27. As $\sum b_j^2 < 1$, eq. 8.28 gives $\sigma_x^2 < \sigma_\epsilon^2$. Surplus and range of the variable x are effected by two opposite forces: (a) the small value of σ_x decreases the surplus and range of x; and (b) the dependence model increases the surplus and range of x in comparison with those of ϵ. Which effect is stronger depends on the type of dependence model. As the auto regressive schemes can be shown in the form of

moving average models, the sum $\sum_{-\infty}^{\infty} b_j^2 \gtrless 1$ determines whether var $x \gtrless$ var ϵ, as shown by eq. 8.28.

To avoid the effect of a $\sigma_x \neq 1$, the dependence model used in this study is that of eq. 8.41, so that $\sigma_x = \sigma_\epsilon = 1$. To fit this case in practical applications, first the original variable, y, is standardized so that $x = (y - \mu_y)/\sigma_y$. Then ρ in eq. 8.41 is estimated so that the independent variable, ϵ, has the mean zero and variance unity.

As $\sigma_x = (1 - \rho^2)^{-1/2} \sigma_\epsilon$ in the model of eq. 8.35, it is sufficient to multiply all values of surplus and range by σ_x and divide their densities by σ_x, to reduce the case of the model of eq. 8.41 to that of eq. 8.35.

Distributions of dependent standard normal variables $(0, 1, \rho)$ are determined by the data generation method for the dependence model given by eq. 8.41. The following values of ρ were used: 0.10, 0.20, 0.40, 0.60 and 0.80. Figure 8.15 gives the distributions of surplus, S_n^+, with the case of $\rho = o$ included. Figure 8.16 gives the probability mass for $S_n^+ = o$ with the case $\rho = o$ included. Figure 8.17 gives the distributions of range, R_n, with the value $\rho = o$ included. Figures 8.18, 8.19 and 8.20 correspond case by case to figs. 8.15, 8.16 and 8.17, respectively, but for the adjusted surplus and the adjusted range.

Figure 8.21 gives the mean, the variance and the skewness coefficient for the surplus and the range as they change both with the change of the variate n and the parameter ρ (ρ = 0, 0.10, 0.20, 0.40, 0.60 and 0.80), all obtained by the data generation method from 100,000 numbers of the independent standard normal variable and the dependence model of eq. 8.41. Figure 8.22 gives the same graphs as fig. 8.21, but for the adjusted surplus and adjusted range. The values for the surplus (or the adjusted surplus) as given in fig. 8.21 (or in fig. 8.22) are the averages of two computed values: that obtained for the surplus and that obtained for the deficit by the data generation method.

Figures 8.21 and 8.22 represent the results obtained by the data generation method. These results contain the sampling errors. A sample size of N = 100,000 generated dependent numbers gave m = N/n as the derived sample size for surplus, deficit and range, for each n. For n = 2, the size is m = 50,000. For n = 50, the size is m = 2,000. The accuracy of results, therefore, decreases with an increase of n. Also, the accuracy decreases with an increase of the order of statistical moments used in computing the three parameters: the mean, the variance and the skewness coefficient. It is clearly seen from these two figures (8.21 and 8.22) that the skewness coefficients are subject to larger sampling errors than the mean or the variance even though the sample sizes of the above three variables are relatively large for any value of n.

The results given in figs. 8.21 and 8.22 show a great effect of the dependence parameter, ρ, as used in the model of eq. 8.41, on the distributions of surplus, range, adjusted surplus and adjusted range. This effect increases non-linearly with an increase of ρ, and the relative effect is greatest on the variance rather than on the mean, and the smallest effect seems to be on the skewness coefficient. However, for a small n (n = 2 - 25) the means of adjusted surplus and adjusted range (see fig. 8.22) may be smaller for a large ρ than the means for a small ρ.

The analytical derivations of distributions for surplus, deficit and range of dependent standard normal variables are not attempted in this study. The comparison of the results obtained by the data generation method and analytical approach are not therefore discussed in this paper.

Hydrologic variables usually have a time dependence which is of various moving average schemes (autoregressive schemes included). These schemes may have the sum of squares of b_j - coefficients either greater than, equal to, or lower than unity. The influence of the dependence generating process on the distribution of surplus, deficit and range should be studied for the following two factors: (a) effect of change in the variance, obtained for the dependent variable by the generating process in comparison with the variance of independent variable; and (b) effect of the dependence parameter or parameters. The above case of the dependence generating model of eq. 8.41 is given here to show only the latter effect by eliminating the effect of the first factor, or by keeping the variances unities for both the independent and generated dependent variables while changing the dependence parameter, ρ.

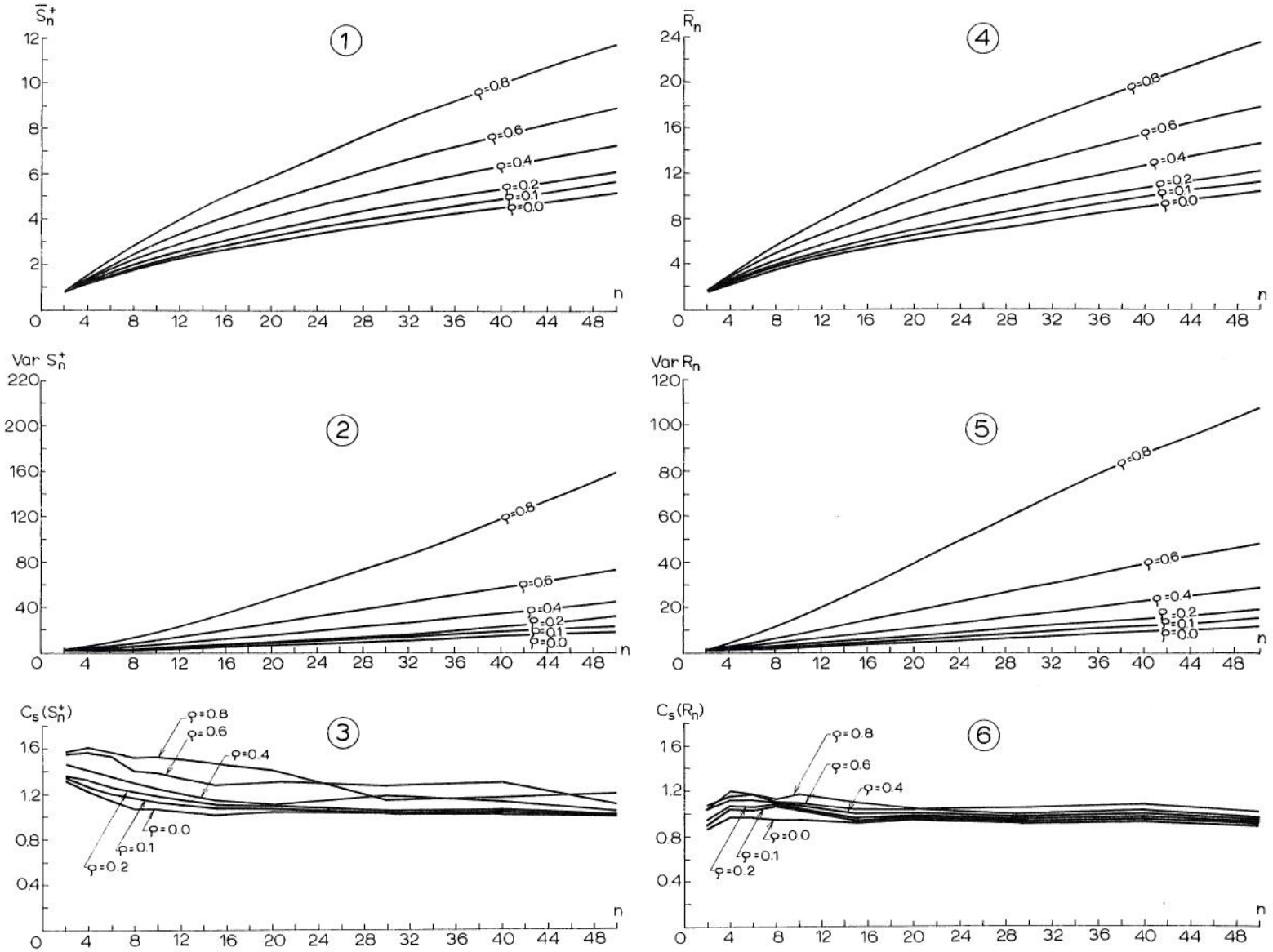


Fig. 8.21 Mean, variance and skewness coefficient of the surplus and range as they change with n (1 - 50) and with ρ (ρ = 0, 0.1, 0.2, 0.4, 0.6, and 0.8) for the dependent standard normal variables (0, 1, ρ), with the dependence generated by the model of eq. 8.41, and the results obtained by the data generation method (from 100,000 independent numbers): (1) Mean of surplus, $\overline{S_n^+}$; (2) Variance of surplus, var S_n^+; (3) Skewness coefficient of surplus, $C_s(S_n^+)$. (4) Mean of range, $\overline{R}_n$; (5) Variance of range, var R_n; and (6) Skewness coefficient of range, $C_s(R_n)$.

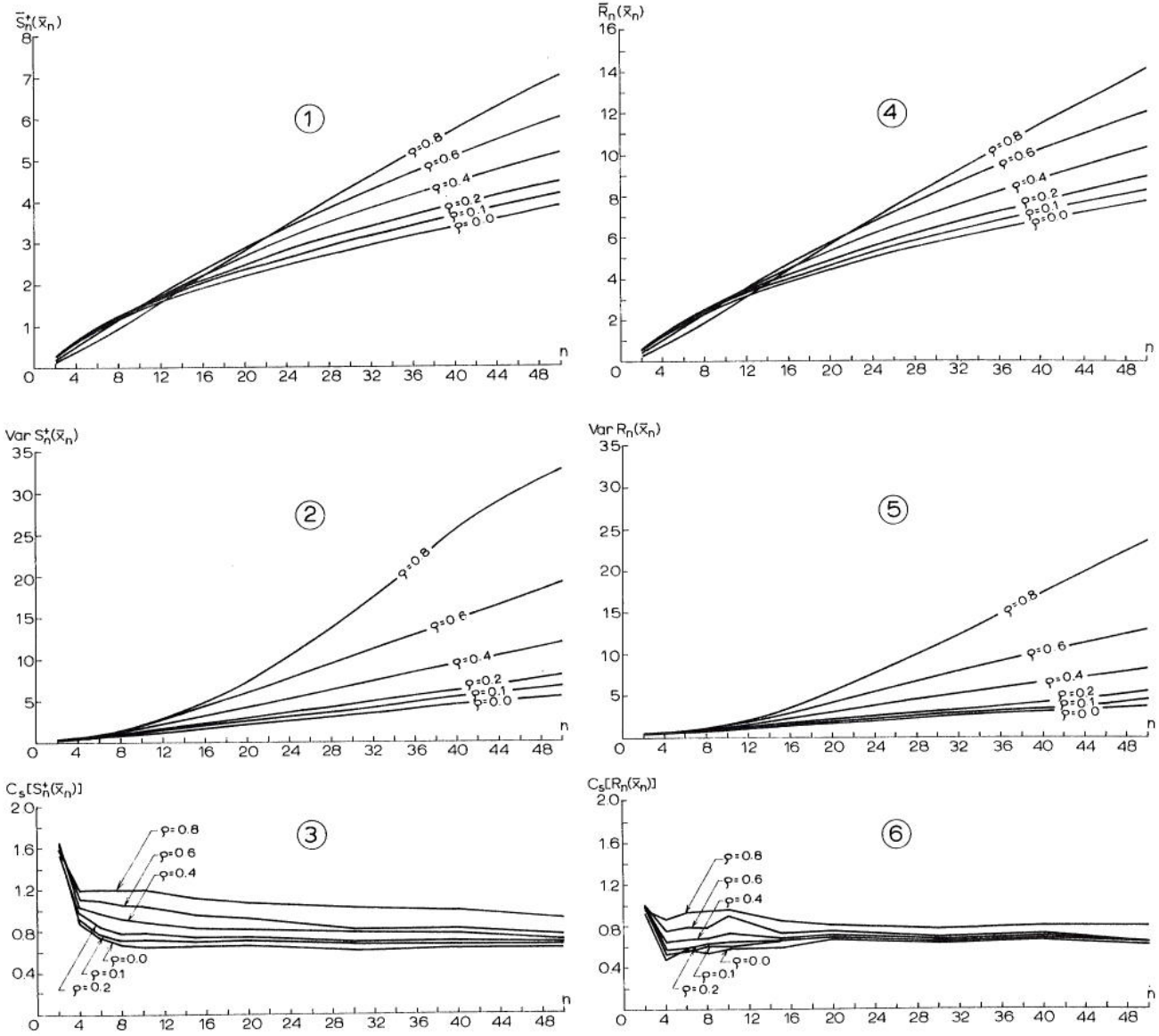


Fig. 8.22 Mean, variance and skewness coefficient of the adjusted surplus and adjusted range, as they change with n (2 - 50) and with ρ (ρ = 0, 0.1, 0.2, 0.4, 0.6 and 0.8) for the dependent standard normal variables (0, 1, ρ), with the dependence generated by the model of eq. 8.41, and the results obtained by the data generation method (from 100,000 independent numbers): (1) Mean of adjusted surplus, $\bar{S}_n^+(\bar{x}_n)$; (2) Variance of adjusted surplus, var $S_n^+(\bar{x}_n)$; (3) Skewness coefficient of adjusted surplus, $C_s[S_n^+(\bar{x}_n)]$; (4) Mean of adjusted range $R_n(\bar{x}_n)$; (5) Variance of adjusted range, var $R_n(\bar{x}_n)$; and (6) Skewness coefficient of adjusted range, $C_s[R_n(\bar{x}_n)]$.

CHAPTER IX

DISTRIBUTION OF SURPLUS, DEFICIT AND RANGE FOR INDEPENDENT GAMMA VARIABLES

1. Gamma variables. A large number of distributions of hydrologic variables may be approximated by gamma distributions with one, two or three parameters. The special form of this latter case is known as the Pearson Type III distribution.

The one-parameter gamma distribution has the form

$$p(x) = \frac{1}{\Gamma(a)} x^{a-1} e^{-x} \qquad 9.1$$

with the lower boundary zero, and unbounded function to the right; a is the shape parameter, and $\Gamma(a)$ is given by

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx . \qquad 9.2$$

In eq. 9.1, the mean $\mu = a$, the variance $\sigma^2 = a$, with the distribution being J-shaped for $0 < a \leq 1$, and bell-shaped for $a > 1$.

The two-parameter gamma function is

$$p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \qquad 9.3$$

where b is the scale parameter, and a is the shape parameter, the mean $\mu = a/b$, and the variance $\sigma^2 = a/b^2 = \mu/b$. Equation 9.3 may be transformed to a one-parameter gamma distribution by the transformation $y = bx$ and $p(y) = p(x)/b$. As a result,

$$p(y) = \frac{1}{\Gamma(a)} y^{a-1} e^{-y} \qquad 9.4$$

which is identical to eq. 9.1.

By using y and $p(y)$ instead of x and $p(x)$, the results of $p(y)$ may be transformed to those of $p(x)$ by $x = y/b$, and $p(x) = b\, p(y)$. This transformation enables the distribution computations of surplus, deficit, range, adjusted surplus, adjusted deficit and adjusted range of the two-parameter gamma variables from the corresponding distributions of these six variables of the corresponding one-parameter gamma variables. It is sufficient to multiply the surplus, deficit and range by b and divide their probability densities by b of the one-parameter gamma variable to obtain the distributions of surplus, deficit and range of the two-parameter gamma variable. The same is true for the adjusted surplus, adjusted deficit and adjusted range. Therefore, it suffices to investigate the distributions of the above six variables for the one-parameter gamma variable.

The three-parameter gamma distribution is

$$p(x) = \frac{b^a}{\Gamma(a)} (x+c)^{a-1} e^{-b(x+c)} \qquad 9.5$$

where a and b are the same parameters as in eq. 9.3 and the third parameter c is the lower boundary of x. Using eq. 9.5, the first transformation $x + c = z$ reduces eq. 9.5 to eq. 9.3, then the transformation $y = bz$ and $p(y) = p(z)/b$ reduces it to eq. 9.4, or the one-parameter gamma variable. The variable z does not change the distribution of surplus, deficit and range in comparison with the variable x, because c is the location parameter and means only the shifting of distribution along the x-axis. It follows from the given two transformations that it is sufficient to investigate the distribution of the six variables with the one-parameter gamma variable.

2. Generation of large samples of independent one-parameter gamma variables. The expressions for the data generation of independent gamma variables are given in Chapter VI, eqs. 6.15 and 6.16. The number m of independent normal standard variables used in the transformation of eq. 6.15 to obtain u-variables determines the skewness of gamma variables. The obtained samples of independent gamma variables have the expected mean $m/2$, the expected variance $m/2$ and the expected coefficient of skewness $(8/m)^{1/2}$. The parameter a of eq. 9.1 is $m/2$.

The following cases are investigated in order to show the influence of the skewness, C_s, on the properties of surplus, deficit, range, adjusted surplus, adjusted deficit and adjusted range of the independent one-parameter gamma variables:

m =	∞	32	16	8
$E(C_s)$ =	0.000	0.500	0.705	1.000

m =	4	2	1
$E(C_s)$ =	1.410	2.000	2.820

where $E(C_s)$ is the expected value of the skewness coefficient, with the expected values of all serial correlation coefficients zeroes, or $E(r_k) = \rho_n = 0$, for any $k \neq 0$. This approach gives only the effect of the independent gamma variables with a positive skewness $0 \leq C_s \leq 2.82$. The same approach was used when developing the dependent one-parameter gamma numbers used in the study of simultaneous effect of skewness and dependence on surplus, deficit,

range, adjusted surplus, adjusted deficit and adjusted range. However, the presentation and analysis of those results are not part of this paper.

The generated samples of independent gamma variables for various m (or C_s) had 100,000 numbers. To obtain samples of 100,000 independent numbers, the total number of independent standard normal numbers was 100,000 m. So, for the case m = 32, it was necessary to generate 3,200,000 independent standard normal numbers in order to obtain 100,000 independent one-parameter gamma numbers with the expected skewness coefficient 0.500.

3. Parameters of distributions of surplus, deficit and range. Figure 9.1 gives the properties of mean, variance and skewness coefficient of surplus, (S_n^+) for various values of skewness of the one-parameter gamma variable. The left three graphs of this figure show the change of mean $(\overline{S}_n^+)$, variance $(\text{var } S_n^+)$ and skewness coefficient $(C_s(S_n^+)$ or $_sC_s)$ of the surplus, both as a function of n, and of C_s of the independent gamma variables. The three right graphs give the differences of values of mean, variance and skewness coefficient of surplus, S_n^+, as given in the left graphs for $C_s \neq 0$, from the case $C_s = 0$ for an independent normal variable in percentage of the corresponding values of mean, variance and skewness coefficient of surplus, S_n^+, for this $C_s = 0$. The values of C_s, given in fig. 9.1, for independent gamma variables are those which are computed from the generaged large samples. The computed C_s values are usually close to the expected values of $E(C_s) = \sqrt{8/m}$ for every m. The deviations between them may be considered only as sampling errors.

The means of surplus for various C_s values of independent gamma variables, as given in graph (1) fig. 9.1, show a rapid convergence by an increase in n to the means of an independent standard normal variable with $C_s = 0$. This independent standard normal variable will be designated in further text as (0, 1, 0) - variable. The smaller C_s of a independent gamma variable, the faster is this convergence of means of surplus to those of the (0, 1, 0) - variable. Or, the greater C_s of an independent gamma variable and the smaller n, the greater is the difference of its means of surplus and the means of surplus of (0, 1, 0) - variable. These differences are negative.

Graphs (3) and (4) of fig. 9.1 show a slower convergence of the variance of surplus of independent gamma variables to that of the (0, 1, 0) - variable, than in the case of means. The greater C_s of a gamma independent variable for a given n, the larger becomes the variance of surplus.

Graphs (5) and (6) of fig. 9.2 show that the skewness coefficients, $_sC_s$, of the surplus for independent gamma variables increase with an increase of the C_s - value of these variables in comparison with $_sC_s$ - values of surplus of the (0, 1, 0) - variable.

It should be noted that the surplus is mainly affected by the long tail of gamma variables while the deficit is affected mainly by the short tail. The differences in shape of these two tails increases with an increase of skewness coefficient, C_s, of independent gamma variables. Therefore, it is expected that the differences in means, variances and skewness coefficients of the surplus and the deficit, both between themselves and between the independent gamma variables and (0, 1, 0) - variable, should increase with an increase of C_s of gamma variables.

Graphs (1) through (6) of fig. 9.2 show the characteristics of distribution parameters of deficit for independent gamma variables with various skewness coefficients, C_s. In general, the patterns of change with n of means, variances and skewness coefficients of deficit seem to be opposite those for the surplus. The means of deficit, $\overline{S}_n^-$, of independent gamma variables converge to those of (0, 1, 0) - variable by an increase in n, though in this case they converge to the values which are about 5 percent smaller than the means of deficit of the (0, 1, 0) - variable. The variances of deficit, $\text{var } S_n^-$, of independent gamma variables are smaller than those of the (0, 1, 0) - variable; and they increase for a given n with an increase of C_s of independent gamma variables as shown on graphs (3) and (4) of fig. 9.2. The skewness coefficients, $_sC_s$, of deficit of independent gamma variables are also smaller than those of the (0, 1, 0) - variable, and they also increase for a given n with an increase of C_s of independent gamma variables as shown on graphs (5) and (6) of fig. 9.2. However, they seem to converge for larger values of n to $C_s = 1.0$ of the deficit of the (0, 1, 0) - variable.

Graphs (1) through (6) of fig. 9.3, show the mean $(\overline{R}_n)$, the variance $(\text{var } R_n)$ and the skewness coefficient (C_s) of range as they change with n for various values of skewness, C_s, of independent gamma variables. The means of the range are close to those of the (0, 1, 0) - variable, as shown in graph (1). Graph (2) gives $\Delta\overline{R}_n$ between those of independent gamma variables and those of the (0, 1, 0) - variable in a percentage of this latter case. On the average, these differences are about 1 percent on the negative side. For large values of C_s (1.883 and 2.673) and small values of n the means of range for independent gamma variables are smaller than those of the (0, 1, 0) - variable. These means converge rapidly with an increase in n to the means of range of the (0, 1, 0) - variable.

The variances of range $(\text{var } R_n)$ of independent gamma variables increase with an increase of C_s as shown in graphs (3) and (4) of fig. 9.3. For small values of C_s, say $C_s = 0 - 1.00$, the differences between the variances of range for independent gamma variables and the (0, 1, 0) - variable are small. These differences, as graph (4) shows,

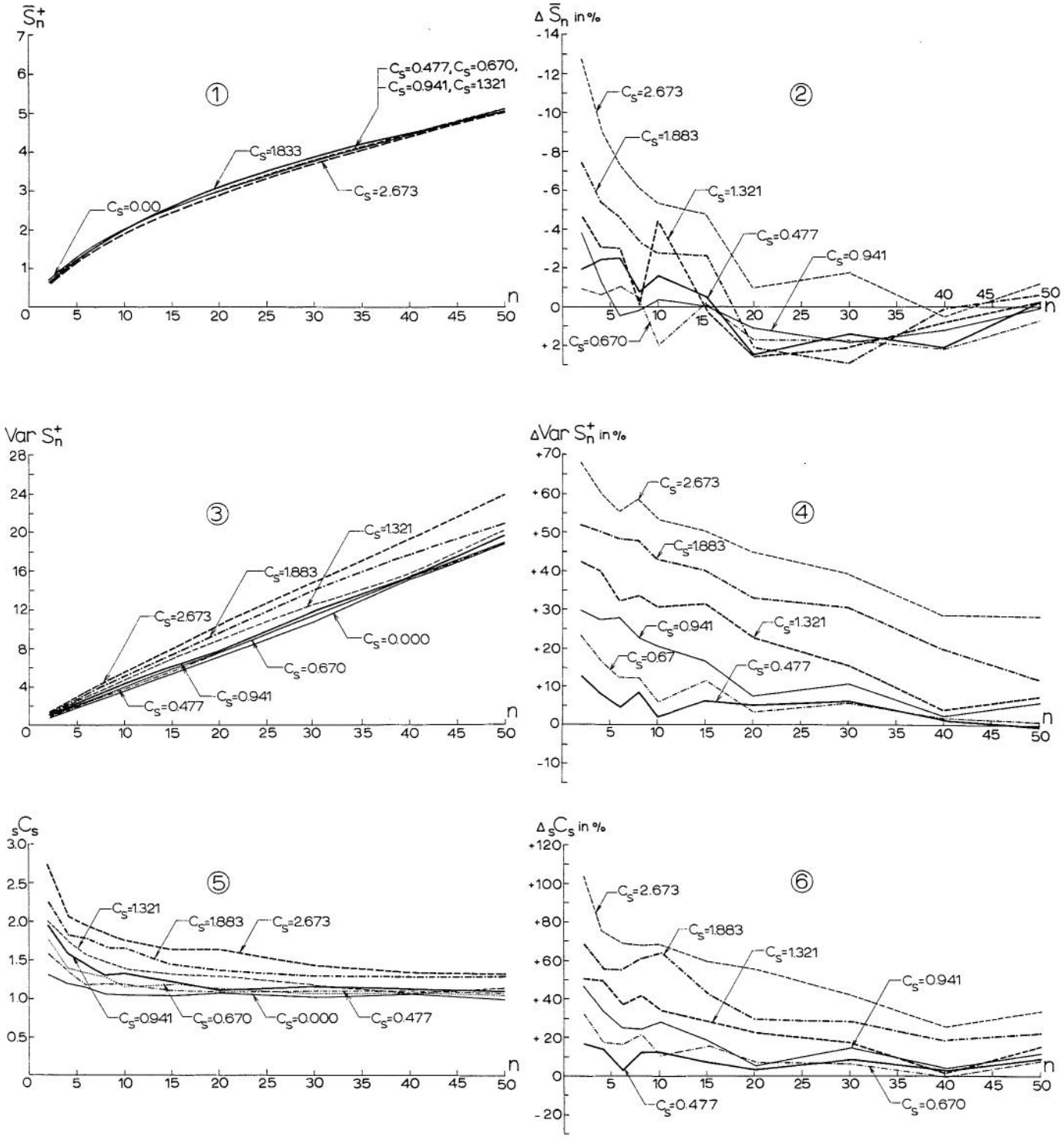


Fig. 9.1 Distribution parameters of the surplus for the independent gamma variables with various skewness coefficients, as they change with subseries length n: (1) Mean of surplus; (2) Differences of means of surplus for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (3) Variance of surplus; (4) Differences of variances of surplus for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (5) Skewness coefficient of surplus; and (6) Differences of skewness coefficients of surplus for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values.

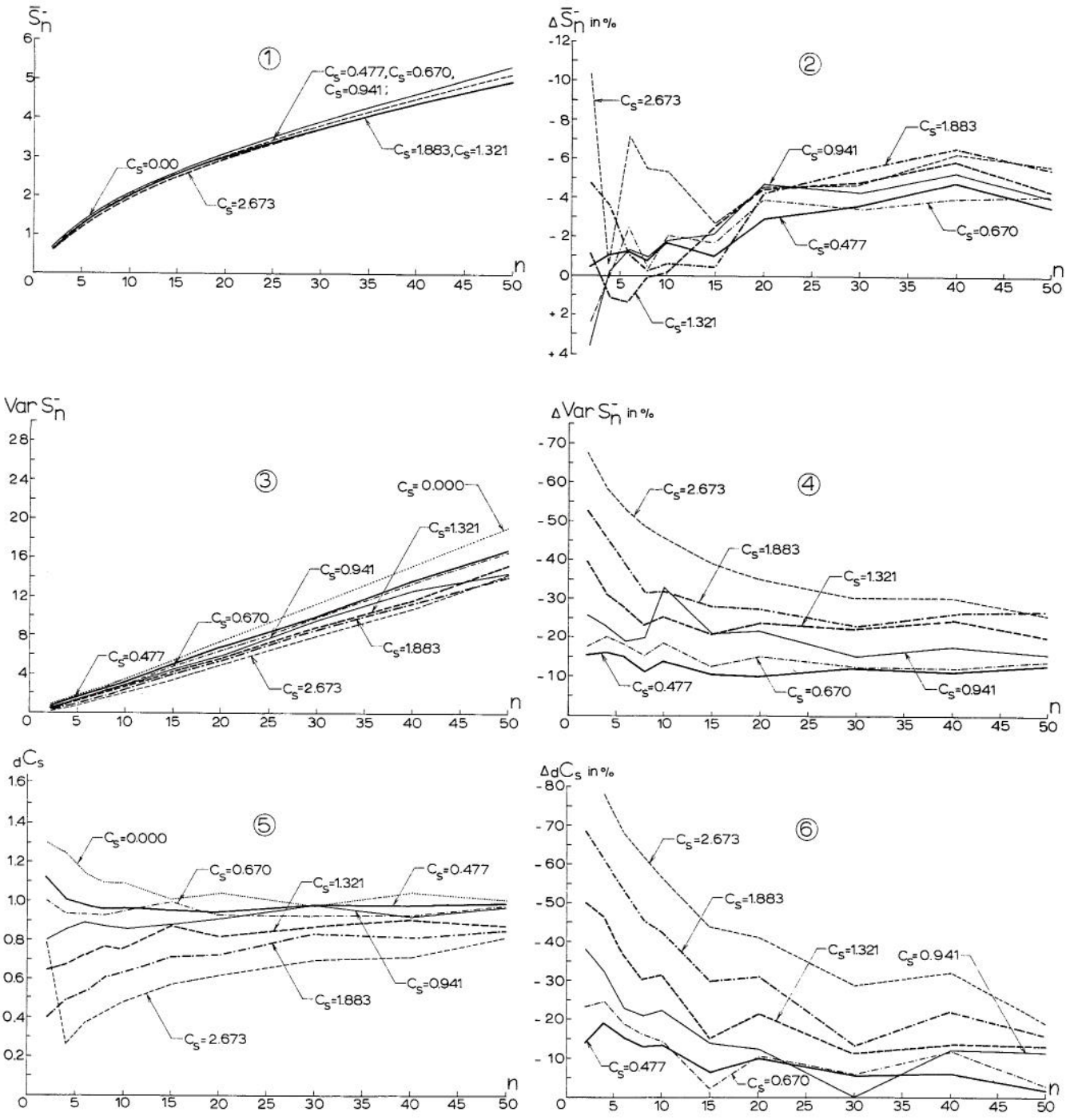


Fig. 9.2 Distribution parameters of the deficit for the independent gamma variables with various skewness coefficients, as they change with subseries length n: (1) Mean of deficit; (2) Differences of means of deficit for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (3) Variance of deficit; (4) Differences of variances of deficit for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (5) Skewness coefficient of deficit; and (6) Differences of skewness coefficients of deficit for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values.

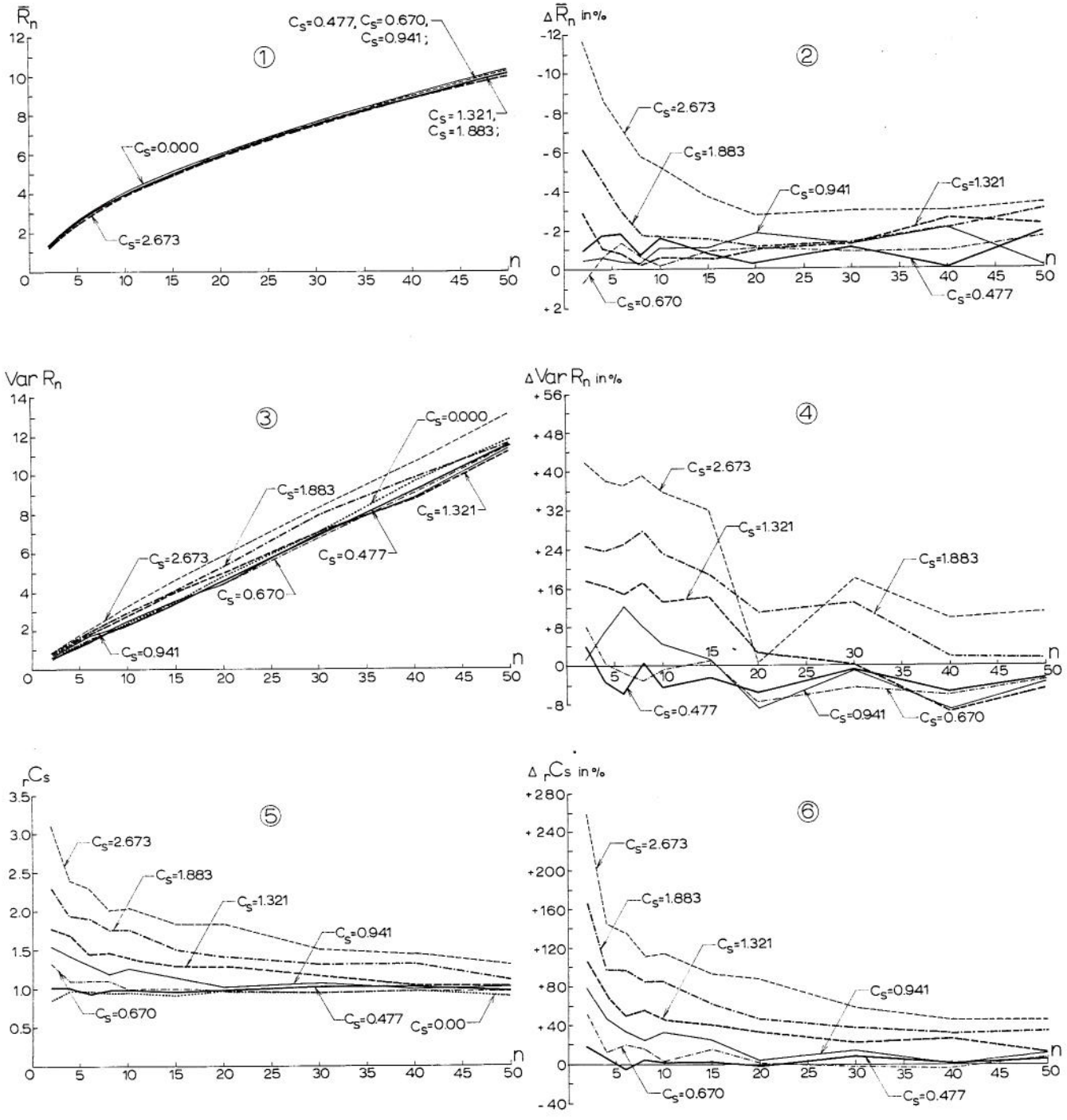


Fig. 9.3 Distribution parameters of the range for the independent gamma variables with various skewness coefficients, as they change with subseries length n: (1) Mean of range; (2) Differences of means of range for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (3) Variance of range; (4) Differences of variances of range for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (5) Skewness coefficient of range; and (6) Differences of skewness coefficient of range for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values.

are positive and increase with an increase of C_s, while they decrease with an increase in n.

The skewness coefficients ($_rC_s$) of range of independent gamma variables are greater than those of the (0, 1, 0) - variable. They increase with an increase of C_s and decrease asymptotically with an increase in n. The skewness coefficient of range is greatly affected by the skewness of independent gamma variables.

4. Parameters of distributions of adjusted surplus, adjusted deficit and adjusted range. Figures 9.4, 9.5 and 9.6 give the means, variances and skewness coefficients of adjusted surplus, adjusted deficit and adjusted range, respectively, for independent gamma variables of various skewness. The six graphs of each figure are analoguous case by case to the graphs of figs. 9.1, 9.2 and 9.3. The results in figs. 9.4 through 9.6 are similar to those of figs. 9.1 through 9.3 except that the absolute values of the three parameters and differences of parameters studied, as well as the rate of their convergence to parameters of the (0, 1, 0) - variable are somewhat different for the cases of adjusted surplus, adjusted deficit and adjusted range. The main exceptions in their patterns are that the variances and skewness coefficients of adjusted deficit of independent gamma variables (fig. 9.5) are greater than those of the (0, 1, 0) - variable, which is opposite from the case shown in fig. 9.2 for the variances and skewness coefficients of the deficit. In many cases, the differences given in graphs (2), (4) and (6) of figs. 9.4 through 9.6 show a small increase for a small n, and they decrease with an increase in n by converging to the zero-difference for very large values in n.

5. Conclusions. The above analysis points out the following conclusions:

(a) The skewness of an independent variable has a much greater effect on the distributions of surplus, deficit, adjusted surplus and adjusted deficit than on the distributions of range and adjusted range;

(b) The effect of skewness of an independent variable increases with an increase of the order of statistical moment used in computing the distribution parameters of surplus, deficit, range, adjusted surplus, adjusted deficit and adjusted range; and,

(c) The effect of skewness of an independent variable on the properties of distribution of these six variables, and particularly on the means, may be neglected only when simultaneously the skewness coefficients are small and the values of n are large.

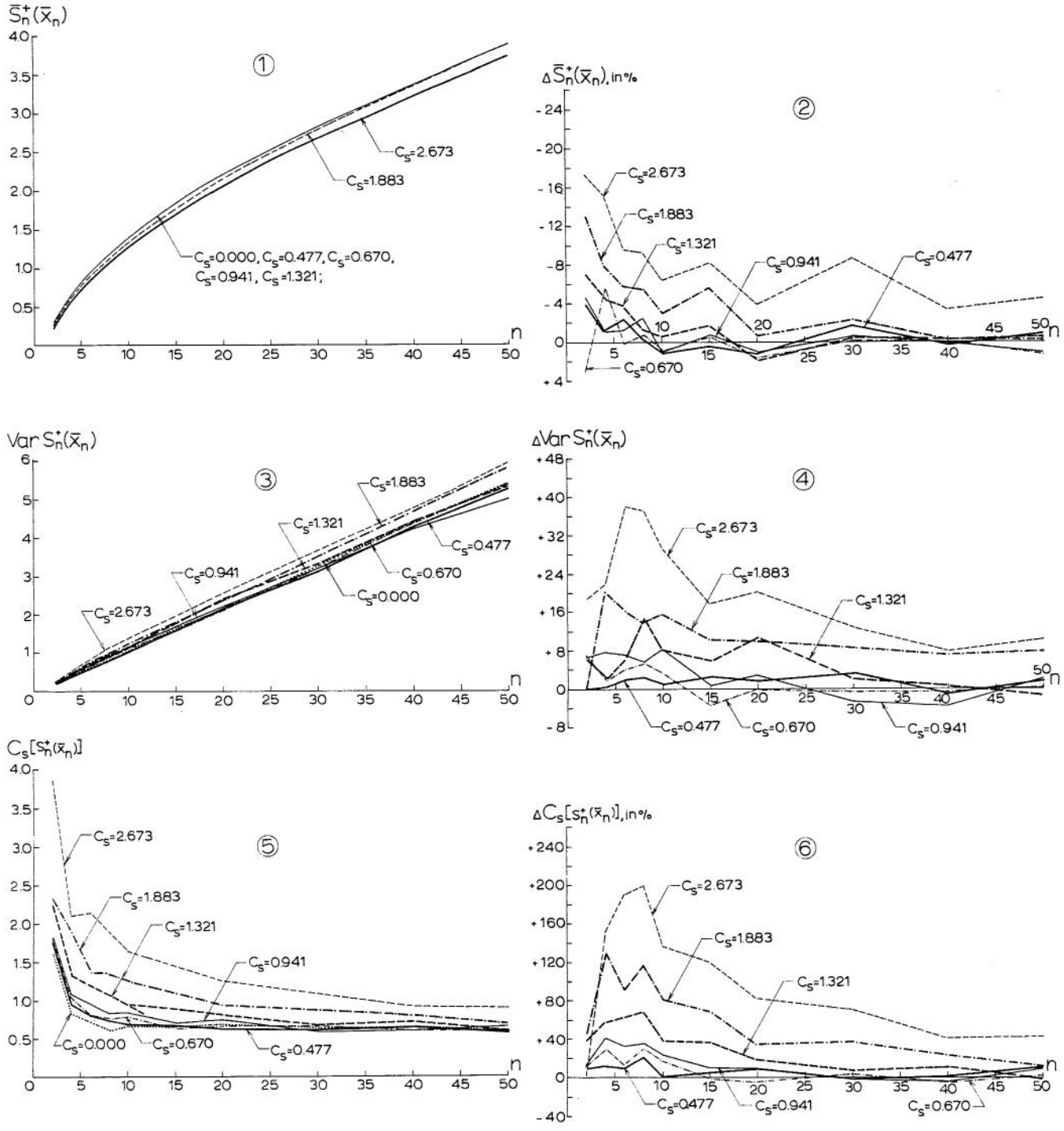


Fig. 9.4 Distribution parameters of the adjusted surplus for the independent gamma variables with various skewness coefficients, as they change with subseries length n: (1) Mean of adjusted surplus; (2) Differences of means of adjusted surplus for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (3) Variance of adjusted surplus; (4) Differences of variances of adjusted surplus for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (5) Skewness coefficient of adjusted surplus; and (6) Differences of skewness coefficients of adjusted surplus for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values.

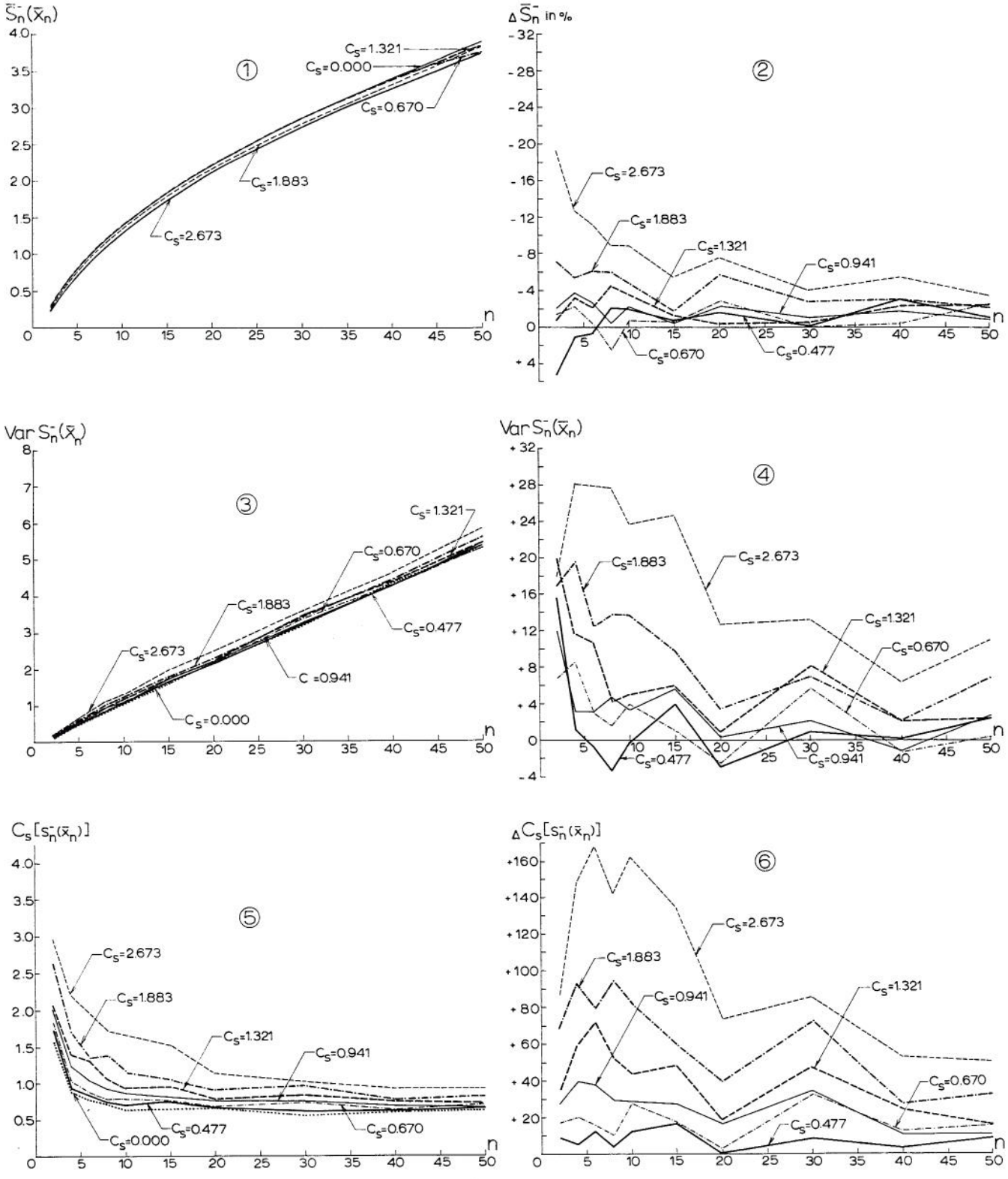


Fig. 9.5 Distribution parameters of the adjusted deficit for the independent gamma variables with various skewness coefficients, as they change with subseries length n: (1) Mean of adjusted deficit; (2) Differences of means of adjusted deficit for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (3) Variance of adjusted deficit; (4) Differences of variances of adjusted deficit for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (5) Skewness coefficient of adjusted deficit; and (6) Differences of skewness coefficients of adjusted deficit for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values.

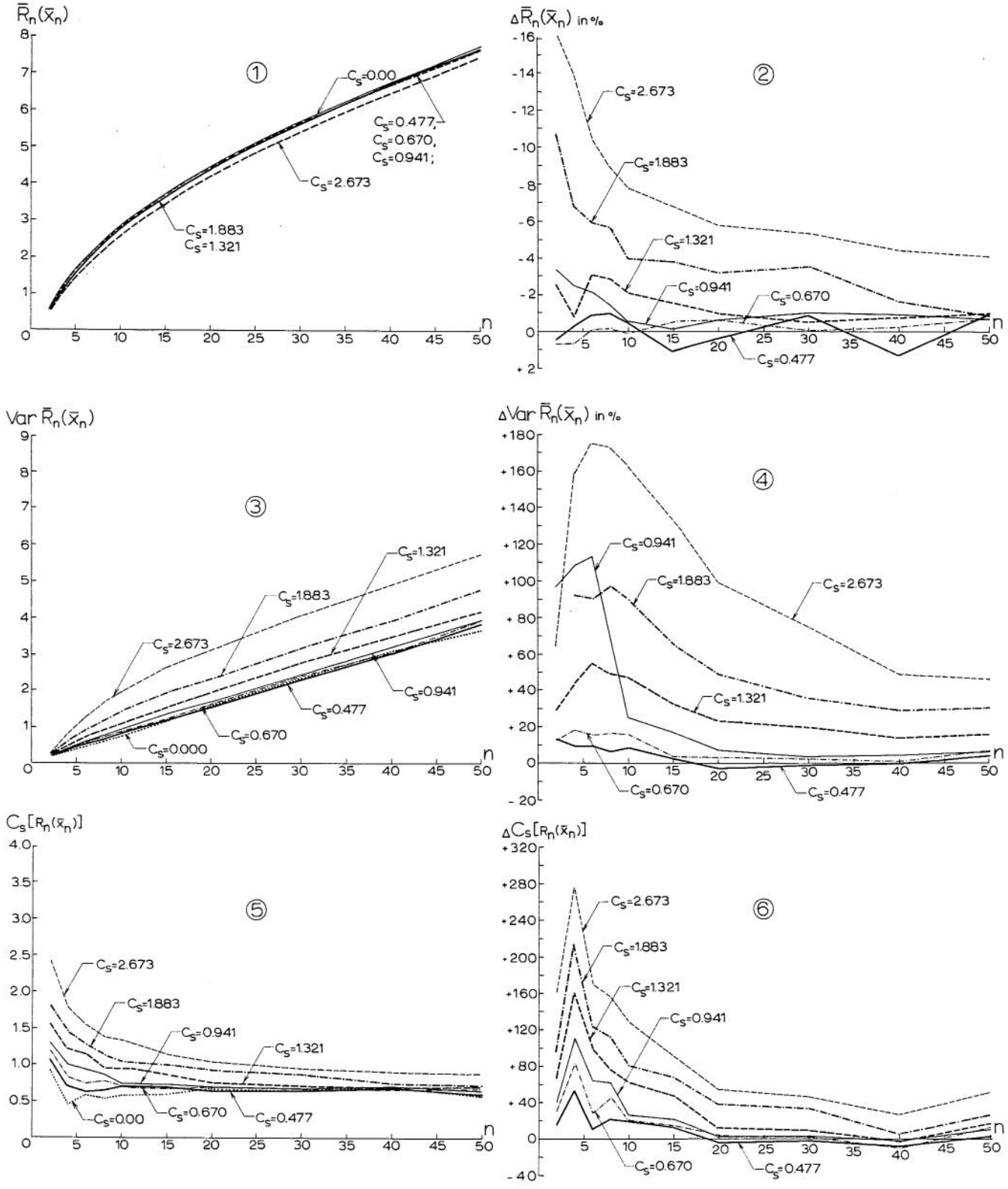


Fig. 9.6 Distribution parameters of the adjusted range for the independent gamma variables with various skewness coefficients, as they change with subseries length n: (1) Mean of adjusted range; (2) Differences of means of adjusted range for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (3) Variance of adjusted range; (4) Differences of variances of adjusted range for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values; (5) Skewness coefficient of adjusted range; and (6) Differences of skewness coefficients of adjusted range for the independent gamma variables and the (0, 1, 0) - variable, in percent of the latter values.

BIBLIOGRAPHY

1. Hurst, H. E., Long-term storage capacity of reservoirs, Am. Soc. Civil Engrs. Trans., 1951, Vol. 116, no. 2447, p. 776.

2. Hurst, H. E., Methods of using long-term storage in reservoirs, Institution of Civil Engrs. Proc., London, 1956, Pt. 1, Vol. 5, p. 519.

3. Hurst, H. E., The problem of long-term storage in reservoirs, Intern. Union Geophysics and Geodesy Information Bulletin, London, 1956, no. 15, p. 463.

4. Feller, W., The asymptotic distributions of the range of series of independent random variables, Ann. Math. Stat., 1951, Vol. 22, pp. 427-432.

5. Anis, A. A., and Lloyd, E. H., On the range of partial sums of a finite number of independent normal variables, Biometrika, 1953, Vol. 40, pp. 35-42.

6. Anis, A. A., The variance of the maximum of partial sums of a finite number of independent normal variates, Biometrika, 1955, Vol. 42, pp. 96-101.

7. Anis, A. A., On the moments of the maximum of partial sums of a finite number of independent normal variates, Biometrika, 1956, Vol. 43, pp. 79-84.

8. Moran, P. A. P., The theory of storage, Methnen and Company, Ltd., London, 1959.

9. Yevdjevich, V. M., Fluctuations of Wet and Dry Years, Part I, Research data assembly and mathematical models, Colorado State University Hydrology Papers No. 1, July 1963, 55 p.

10. Yevdjevich, V. M., Fluctuations of wet and dry years, Part II, Analysis by serial correlation, Colorado State University Hydrology Papers No. 4, June 1964, 50 p.

11. Melentijevich, M. J., Characteristics of storage when outflow is dependent upon reservoir volume, Ph. D. dissertation, Colorado State University, 1965.

Key Words: Range, Application of range in hydrology, Reservoir storage, Range of the cumulative deviations.

Abstract: Surplus is defined as the maximum positive sum, deficit as the minimum negative sum, and range as their difference (or the sum of their absolute values) on a curve of cumulative deviations from a base value x for a given subseries of length n. An attempt is made to systematize the types of storage equations. Application of these statistical parameters in hydrology is discussed. Storage problems and the use of surplus, deficit and range in analyzing these problems are viewed from the three approaches: empirical method, data generation method and analytical method. Properties of these three methods are also investigated in detail. Smoothness in the results of the latter two methods should not be mistaken for increased information. The three methods are compared on the bases of the Rhine River's annual flows. Surplus and range of normal independent and dependent variables are investigated by the analytically derived expressions and by the data generation method.

The effect of dependence in time series on surplus and range is studied for the Markov first order linear dependence model of a normal variable. The surplus and range change significantly with an increase of the dependence parameter.

The effect of skewness on the statistical parameters (mean, variance and skewness coefficient) or surplus, deficit, range, are investigated for independent gamma variables with skewness coefficients ranging from zero to $\sqrt{8}$. The effect of skewness is larger on the surplus and deficit than on the range, and it increases with an increase of the order of statistical moment.

Reference: Yevdjevich, V. M., Colorado State University, Hydrology Papers No. 10 (September 1965), "The Application of Surplus, Deficit and Range in Hydrology."

Key Words: Range, Application of range in hydrology, Reservoir storage, Range of the cumulative deviations.

Abstract: Surplus is defined as the maximum positive sum, deficit as the minimum negative sum, and range as their difference (or the sum of their absolute values) on a curve of cumulative deviations from a base value x for a given subseries of length n. An attempt is made to systematize the types of storage equations. Application of these statistical parameters in hydrology is discussed. Storage problems and the use of surplus, deficit and range in analyzing these problems are viewed from the three approaches: empirical method, data generation method and analytical method. Properties of these three methods are also investigated in detail. Smoothness in the results of the latter two methods should not be mistaken for increased information. The three methods are compared on the bases of the Rhine River's annual flows. Surplus and range of normal independent and dependent variables are investigated by the analytically derived expressions and by the data generation method.

The effect of dependence in time series on surplus and range is studied for the Markov first order linear dependence model of a normal variable. The surplus and range change significantly with an increase of the dependence parameter.

The effect of skewness on the statistical parameters (mean, variance and skewness coefficient) or surplus, deficit, range, are investigated for independent gamma variables with skewness coefficients ranging from zero to $\sqrt{8}$. The effect of skewness is larger on the surplus and deficit than on the range, and it increases with an increase of the order of statistical moment.

Reference: Yevdjevich, V. M., Colorado State University, Hydrology Papers No. 10 (September 1965), "The Application of Surplus, Deficit and Range in Hydrology."

Key Words: Range, Application of range in hydrology, Reservoir storage, Range of the cumulative deviations.

Abstract: Surplus is defined as the maximum positive sum, deficit as the minimum negative sum, and range as their difference (or the sum of their absolute values) on a curve of cumulative deviations from a base value x for a given subseries of length n. An attempt is made to systematize the types of storage equations. Application of these statistical parameters in hydrology is discussed. Storage problems and the use of surplus, deficit and range in analyzing these problems are viewed from the three approaches: empirical method, data generation method and analytical method. Properties of these three methods are also investigated in detail. Smoothness in the results of the latter two methods should not be mistaken for increased information. The three methods are compared on the bases of the Rhine River's annual flows. Surplus and range of normal independent and dependent variables are investigated by the analytically derived expressions and by the data generation method.

The effect of dependence in time series on surplus and range is studied for the Markov first order linear dependence model of a normal variable. The surplus and range change significantly with an increase of the dependence parameter.

The effect of skewness on the statistical parameters (mean, variance and skewness coefficient) or surplus, deficit, range, are investigated for independent gamma variables with skewness coefficients ranging from zero to $\sqrt{8}$. The effect of skewness is larger on the surplus and deficit than on the range, and it increases with an increase of the order of statistical moment.

Reference: Yevdjevich, V. M., Colorado State University, Hydrology Papers No. 10 (September 1965), "The Application of Surplus, Deficit and Range in Hydrology."

Key Words: Range, Application of range in hydrology, Reservoir storage, Range of the cumulative deviations.

Abstract: Surplus is defined as the maximum positive sum, deficit as the minimum negative sum, and range as their difference (or the sum of their absolute values) on a curve of cumulative deviations from a base value x for a given subseries of length n. An attempt is made to systematize the types of storage equations. Application of these statistical parameters in hydrology is discussed. Storage problems and the use of surplus, deficit and range in analyzing these problems are viewed from the three approaches: empirical method, data generation method and analytical method. Properties of these three methods are also investigated in detail. Smoothness in the results of the latter two methods should not be mistaken for increased information. The three methods are compared on the bases of the Rhine River's annual flows. Surplus and range of normal independent and dependent variables are investigated by the analytically derived expressions and by the data generation method.

The effect of dependence in time series on surplus and range is studied for the Markov first order linear dependence model of a normal variable. The surplus and range change significantly with an increase of the dependence parameter.

The effect of skewness on the statistical parameters (mean, variance and skewness coefficient) or surplus, deficit, range, are investigated for independent gamma variables with skewness coefficients ranging from zero to $\sqrt{8}$. The effect of skewness is larger on the surplus and deficit than on the range, and it increases with an increase of the order of statistical moment.

Reference: Yevdjevich, V. M., Colorado State University, Hydrology Papers No. 10 (September 1965), "The Application of Surplus, Deficit and Range in Hydrology."