J. Reimann

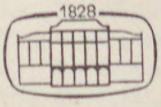
MATHEMATICAL STATISTICS WITH APPLICATION IN FLOOD HYDROLOGY

Akadémiai Kiadó, Budapest

MATHEMATICAL STATISTICS WITH APPLICATION IN FLOOD HYDROLOGY

J. REIMANN

MATHEMATICAL STATISTICS WITH APPLICATION IN FLOOD HYDROLOGY



AKADÉMIAI KIADÓ, BUDAPEST 1989

Translated by

ANDRÁS SZÖLLŐSI NAGY, MIKLÓS BOSZNAI and PÁL MAGYAR

ISBN 963 05 4832 1

Printed in Hungary by Szegedi Nyomda

CONTENTS

PREFACE

The basic objective of this book is to introduce the reader into the probabilistic and statistical model building techniques related to flood problems. The book was written in the hope that the relatively simple statistical techniques contained therein will help the hydrologist — or the hydrologist student — to identify information contained in hydrologic records relevant to flood protection activities.

In order to understand the techniques presented in this book nothing beyond the knowledge of elementary calculus is assumed.

In order to make easier the understanding of the statistical methods used in the analysis of flood waves the first few chapters of the book summarize the bases of probability theory which is illustrated through a number of hydrological, possibly flood related, examples.

Subsequent chapters deal with the rather known and conventional statistical methods but from the point of view of flood hydrology. Therefore, these methods are somewhat modified and improved in accordance with the particular features of flood computations.

The majority of examples is related to the hydrological problems of River Tisza, Hungary. This stems, on the one hand, from the particular feature that the catchment system thereof can be handled as an entity and, on the other hand, from her importance in the country's flood protection system. During floods, River Tisza and its tributaries endanger an area of 17600 sq.km that is roughly one fifth of the territory of Hungary, inhabited by nearly one fourth of our population. The levee system of River Tisza is one of the most developed flood protection systems in Europe even now. The techniques advocated in this book can, however, be applied, perhaps with some minor modifications, for any other river system.

The statistical analysis of flood waves does not require the application of just one or two particular chapters of probability theory. The analysis will be efficient only if the hydrologist possesses a rather broad statistical knowledge and the methods of statistics are combined according to the very nature of the problem. This book tries to provide for this task as far as the limits of its reasonable extent will allow.

The book will achieve its objectives if it inspires the reader to derive further efficient methods beyond those presented therein.

And now a few words about the structure of the book. Part I, i.e. Chapters 1 to 3,

contain the bases of probability theory. In hydrological applications exactly the fundamental bases of the theory play the most important role. Nevertheless, certain applications will be demonstrated through a few examples in these chapters, too, though they cover only a rather small fraction of the entire area of applications. In Chapter 1, when discussing the properties of the Poisson distribution, it is shown that the number of floods follow the Poisson rule. (This is justified through a x^2-test later in Chapter 6.) Whenever it was possible, a simple combinatorial approach was used. In this context I want to call the attention of the reader to the fact that combinatorial techniques play an important role in modern mathematical statistics. Part II, Chapters 4 to 6, cover the basic methods of mathematical statistics, illustrated possibly in connection with flood wave analysis problems. It is shown in these Chapters, among others by mathematical statistical techniques, that the number of exceedances in a given time interval generally follow Poisson distribution while maximum exceedances over a given time period are distributed according to a certain double exponential distribution.

Part III, Chapters 7 and 8, deal with the methods of analysing stochastic relations between random variables. This is an important field of practical hydrology and its scope is wider than the usual correlation and regression analyses. Some new methods which were established recently are introduced in these Chapters. Theory is still developing in this field and is far from being ready and complete.

At last I would say thanks to all those who helped to publish this book. First of all I wish to express my sincere gratitude to the readers of the book, *Dr. Zoltán Szigyártó* and *Dr. István Vincze,* for their valuable comments and advices which helped to improve the manuscript. Dr. Szigyártó was giving useful advices concerning the structure of the book and the applications. Dr. Vincze was providing an assistance with respect to the mathematical presentation and to the work as a whole, far beyond the duties of a publisher's reader.

I also extend my thanks to the Publishing House of the Hungarian Academy of Sciences, for its efficient help in publishing this book.

PART I

FUNDAMENTALS OF PROBABILITY THEORY

CHAPTER 1

1.1. THE ROLE OF PROBABILITY THEORY IN FLOOD HYDROLOGY

For the interest of flood control, the behaviour of rivers, as reflected by historical hydrological sequences, has been in the focal point of reasearch long ago, applying probability theory and statistical analysis to hydrologic records. Gauging stations had been installed along river banks and water level data were observed on a regular basis. Observation data had been, in turn, published in Hydrological Year Books. One of the objectives of these observations is to infer the future behaviour of the river in question. The water level (stage) of a river at a given location and in a given time epoch depends on several factors, which may be considered random events such as e.g. the volume of precipitation fallen over a catchment, runoff and temperature conditions, stages and discharges of tributaries, etc. These variables, which are affected by a large number of, sometimes not even quantifiable, effects are called random variables. (This notion will be discussed in detail in Chapter II.)

Probability theory deals with the analysis of random variables. Mathematical statistics, being a branch of probability calculus, is for the practical application of this theory.

Probability theory and mathematical statistics are jointly termed as stochastic methods including the theory of stochastic processes that is also part of the probability theory. Due to the fact that the measurement of water stages is simpler and more accurate than that of the discharges, the former will mainly be analyzed in the subsequent chapters of this book. Primarily, *high stage values contain important information in flood studies*. Earlier, observed annual maximum stages of a given river section were considered as basic data for flood protection planning.

This book adopts a different approach, viz. not only the highest, but also *additional high stage values will be included in the analysis*.

How can one 'read out' relevant information from a sequence of flood stage values? Probability calculus and mathematical statistics are the basic means for answering this question. The basic aim of this work is to justify this statement. However, before starting an in-depth analysis, first the essence of flood is to be defined along with a few other concepts which will be used frequently throughout this text.

The first, second and third levels of flood preparedness are determined for all major river sections in Hungary. Let C denote the first level of flood preparedness. If this level C is plotted along with the sequence of daily stage values, called hydrograph,

than the uninterrupted sequence of stages above this level C is considered as the mathematical model of flood events, see Figure 1.

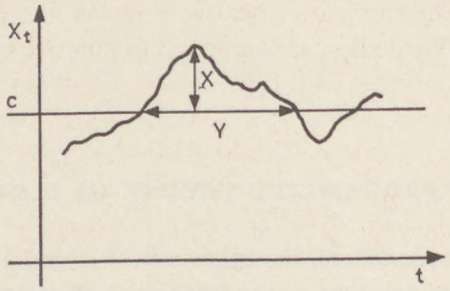


Figure 1

In other words: a flood event starts when the stage exceeds level C and ends when the stage has plummeted again below level C. The maximum of a flood wave is called flood crest, or flood peak. Subsequently, the difference between the peak and level C, i.e. the maximum value of exceedance, will simply be called exceedance. From now on, the value of the exceedance will be denoted by X.

The value of X depends on several random effects, i.e. the exceedance is a random variable. The duration of flood, denoted by Y in Fig. 1., depends also on random factors, such as precipitation, temperature, runoff conditions, etc. Therefore, Y is also a random variable.

The following chapters will deal with the investigation of statistical laws governing exceedances and flood durations. These are essential data, as it is important to know the frequency of floods, what is the expected exceedance and what will be the expected duration of a flood? Also, the relation between exceedances at the different gauges will be analyzed, together with the relation between the measure of exceedance and flood duration. Stochastic relations among exceedances observed at water gauges in confluencing river sections will also be investigated along with some theoretically and practically interesting problems.

To undertake probability analysis of floods the reader must have a basic understanding of probability calculus and mathematical statistics. For easy reference, the first few chapters will summarize the bases of the essential theory. No complete and detailed treatment of probability theory and mathematical statistics will be given at this place. Only the notions and relations will be outlined, in a presentation suitable for hydrologists, which are indispensable for the understanding of the subsequent chapters.

Probability calculus helps in identifying the statistical laws of random fluctuations inherent in hydrological phenomena. In order to promote the validity of computations, the mean (expected value), variance, and confidence interval of random variables or their average values are defined and calculated.

It is impossible to exclude the effects of randomness, as the inputs to these phenomena will continuously change. Natural and social processes will never repeat themselves exactly. It would lead to serious errors if someone excluded this variability and the random effects. Therefore, our objective here (as in any scientific investigation) is to apply stochastic methods in such a way as to improve our understanding of the processes with respect to the knowledge obtained without the application of these techniques. It is our firm belief that the understanding of probability theory and its methods together with some successful applications will certainly resolve the reservations against the use of stochastic methods. The so-called deterministic relations are only contours of given phenomena. By a closer and deeper analysis one should realize the randomness in those phenomena.

This discussion certainly has led to the realization of applicability of the probability theory to flood studies. Since randomness plays an important role in the forming of floods, the only tools to cope with this situation are those of probability theory and mathematical statistics.

1.1.1. RANDOM PHENOMENON

The aim of probability theory is to formulate mathematically, analyze and determine the objective laws of the so called random mass phenomena or experiment.

The term mass phenomenon is related to processes which can be observed under the same conditions (theoretically) any times. Now we clarify what is the meaning of a random phenomenon or random experiment.

There are phenomena or experiments the outcome of which are usually well determined by fixing certain number of factors. In such cases we can assume that all the conditions, circumstances and influencing factors can be enlisted. For instance, it is well-known that distilled water will boil at a temperature of 100 °C and at a pressure of 1013.2 mbar. Let A denote the event that the water starts boiling. If the conditions: chemically clean water, pressure of 1013.2 mbar and temperature of 100 °C respectively hold simultaneously, event A will necessarily occur.

Phenomena which are uniquely determined by the presence of certain circumstances are called *deterministic schemes*. On the other hand there are phenomena or experiments which do not have a unique outcome under given conditions. This means, that repeating the experiment many times the outcome will show in each repetition a certain change, so called random fluctuation. These phenomena will be described by *stochastic schemes* and are called random phenomena, or random experiments.

There is no antagonistic contradiction between these two cases. A stochastic scheme, i.e. a random phenomenon may become deterministic if *all* the causing factors are determined (and measured). The number of floods of a river during a certain time period of the year can be described by a stochastic scheme only. It is a random phenomenon, because not all the causing factors and their complicated interactions can be taken into consideration in all observation or realization. For instance,

if we wish to determine whether there will be any flood in the first three months of the year at Szeged on River Tisza, we must know the snow conditions over the catchment, the rainfall pattern, the speed of snowmelt as a function of temperature, the distribution of precipitation over the tributary watersheds, etc. Moreover, we must know whether the floods of River Tisza and of its tributaries will coincide in time, i.e. we must know exactly the runoff process. If these flood producing factors were known for given years it would not be a help still, as these condition change from year to year. Consequently, floods can be described by stochastic schemes only, at the time being.

This means that event A related to a phenomenon described by a stochastic scheme will not necessarily occur though it may occur. Therefore, those are said to be random phenomena which are not uniquely determined by the conditions or causes considered.

Probability theory deals with the investigation of possible outcomes, i.e. events, the occurrence or non-occurrence of which can be observed under the same conditions. There are statistical regularities in such random mass phenomena. The mathematical description of these laws is the task of probability theory. The discovery of these laws enables to forecast the outcome of random phenomena in the case of a large number of future observations. Probability theory and mathematical statistics are important tools when coping with randomness in long-term water resources planning. Probability theory does not deal with single phenomena as the outcome of a chess-party. In the following, the mathematical description of random events will be presented.

In the forthcoming discussions the random phenomenon considered will be called random experiment.*

The outcome of an experiment has, in general, a fairly complicated structure, therefore one or more characteristic values related to the experiment (e.g. peak, duration, discharge values in flood) are selected and investigated, they are called random variables. A random variable (e.g. the peak of the River Tisza at Szeged in a given time period of the year) is an abstract quantity containing all its possible values. For instance, if the experiment reduces to the water level at a given time and site then the outcome of the observation of the experiment is one record, viz. the observed stage. This experiment has as many possible outcomes as many different water levels may occur between zero and a feasible upper bound of k. In spite of the fact that in the practice stages are measured with an accuracy of 1 cm, the set of all possible outcomes is in fact the $(0, k)$ interval.

Therefore, water level observation may have infinite number of outcomes. Any point in the $(0, k)$ interval is called a sample point. The set of all sample points is called sample space. The sample space is a simplified model or projection of reality with respect to the investigated phenomenon. The essential thing here is that the

* The adjective 'random' will be omitted in the following as this book deals only with random experiments.

model should reflect reality from the point of view of practical interest and of the problem to be solved.

If, for example the experiment consists of observing and registrating the rainy days in March at a particular location (a day is defined rainy, if say, at least 5 mm precipitation has occurred). The possible outcomes of this experiment are 0, 1, 2, ...31. These are the sample points the set of which will form the sample space.

Consider now another example for sample space. During a long observation period it was found that stages at Bratislava, River Danube, fluctuated between 1 and 10 meters. In analyzing the changes usually not the occurrence of a single value, say of 637 cm is of interest but how frequently was the water level greater than 6 m, or how often did it fall between 8 and 10 meters?

In this experiment the sample space is some $(0, k)$ interval, where k is again a feasible upper bound. The sample space is denoted by $\mathfrak{X}$, which in this special case is an interval defined by the set

$$\mathfrak{X} = \{x: x \in (0; k)\}.$$

Any element x of set $\mathfrak{X}$ is a sample point. *Any sub-set of the sample space $\mathfrak{X}$ is called an event.*

Events will be denoted by latin capital letters, A, B, C etc. Set $\mathfrak{X}$ is called the certain (sure) event, and is denoted by I, because at a particular experiment some x sample point will certainly occur. If an observed stage X is in the range (8 m, 10 m) then the event

$$A = \{x: x \in (8\,\text{m};\ 10\,\text{m})\}$$

occurred. As the mathematical model of an event is a set, the sum or product of events and its complementary event can also be defined. These notions are all related to those of the Boolean algebra.

If, e.g.

$$A = \{x: x \in (4\,\text{m};\ 6\,\text{m})\}, \quad B = \{x: x \in (5\,\text{m};\ 7\,\text{m})\}$$

then event $A+B$ will occur whenever stage X is found in interval (4 m, 7 m), i.e.

$$A+B = \{x: x \in (4\,\text{m};\ 7\,\text{m})\} = \{x: x \in (4\,\text{m};\ 6\,\text{m})\} \cup \{x: x \in (5\,\text{m};\ 7\,\text{m})\}$$

and

$$A \cdot B = \{x: x \in (4\,\text{m};\ 6\,\text{m})\} \cap \{x: x \in (5\,\text{m};\ 7\,\text{m})\} = \{x: x \in (5\,\text{m};\ 6\,\text{m})\}.$$

The product of two events means the simultaneous occurrence of the two events, this is the intersection of set A and B.

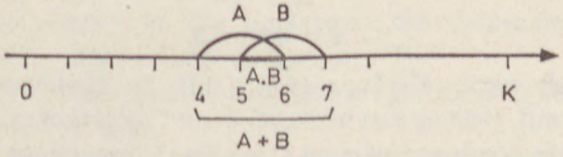


Figure 2

The complementary event $\bar{A}$ of event A consists of those elements x of the sample space which are not in A, i.e.,:

$$\bar{A} = \{x: x \notin A\} = I - A.$$

With reference to the previous example, if

$$A = \{x: x \in (4\,\text{m}; 6\,\text{m})\}$$

then

$$\bar{A} = \{x: x \in (0; 4\,\text{m}) \cup (6\,\text{m}; k)\}.$$

Obviously, $\bar{\bar{A}} = A$, that is the complementary event of the complement of event A is event A itself.

The complement of the certain event I is an empty set, called impossible event, denoted by $\emptyset$:

$$\bar{I} = \emptyset; \quad \bar{\emptyset} = I.$$

If A has occurred but B has not then the event

$$C = A - B = A\bar{B}$$

has occurred. It is easy to show that

$$A + B = B + A\bar{B} = B + C.$$

E.g., if

$$A = \{x: x \in (4\,\text{m}; 6\,\text{m})\}$$

and

$$B = \{x: x \in (5\,\text{m}; 7\,\text{m})\}$$

then

$$C = A - B = A\bar{B} = \{x: x \in (4\,\text{m}; 5\,\text{m})\}.$$

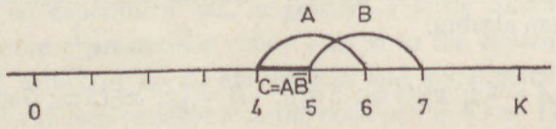


Figure 3

Events B and C are mutually exclusive since if one experiment yields B then C cannot occur, and vice versa. Mutually exclusive events correspond to disjoint sets. The sum of two events can always be generated as the sum of two mutually exclusive events. The fact that B and C are mutually exclusive events is denoted by $BC = \emptyset$.

This example also shows that

$$A + B = A + \bar{A}B$$

and

$$B = AB + \bar{A}B.$$

In our example event $C = \{x: x \in (4\,\text{m}; 5\,\text{m})\}$ is contained in event $A = \{x: x \in (4\,\text{m}; 6\,\text{m})\}$ yielding that whenever event C occurs event A must also occur. In other words, the occurrence of event C implies the occurrence of event A which is symbolically denoted by $C \subset A$.

Obviously, if $C \subset A$ then $AC = C$, and $A + C = A$. Let the interval $(0; k)$ or the base-set I of the above example subdivided by the division points $0 = X_0 < X_1 < \ldots < X_n = K$ into n parts and let A_i denote the event $[X_{i-1}, X_i)$, see Figure 4. Then events $A_1, A_2, \ldots, A_n$ are mutually exclusive:

$$A_i \cdot A_j = \emptyset \quad \text{if} \quad i \neq j$$

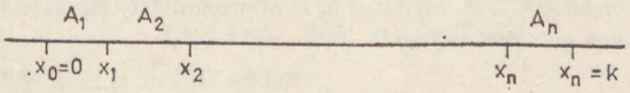


Figure 4

and during one experiment one and only one will certainly occur:

$$A_1 + A_2 + \ldots + A_n = I.$$

The system of events $A_1, A_2, \ldots, A_n$ is called a complete system of events. The complete system of events is a partition of the sample space into disjoint events.

1.1.2. THE NOTION OF PROBABILITY. AXIOMS

The reader has certainly met previously the notion of probability. Moreover, in many cases it can be said without any difficulty whether a random event occurs with a high probability or not. In an experiment we wish to express numerically the chance of the occurrence of some events of interest with respect to some measuring scale, similarly to the measurement of temperature, weight, etc. This desire is motivated by the requirement that we wish to know how many times a particular event will occur in a sequence of observations.

The notion of the probability of a given event will be approached by means of another notion, the relative frequency of a given event in a sequence of observation.

Consider an experiment that can be repeated many times under the same conditions; what we are interested in is how many times a given event A will occur among n repetitions. Assume, that by repeating the experiment n times event A will occur k_n times. It is also assumed that the outcome of an observation will have no effect on the outcome of any other experiments. The quantity k_n is said to be the frequency of event A, while the ratio k_n/n is called the relative frequency of event A in the sequence of experiments. It is obvious that repeating the sequence of n experiments under the same conditions will result in a frequency k_n' which, in general, differs from k_n; this is a consequence of randomness. The usual term is the random fluctuation of quantity k_n while k_n itself is what we called random variable.

The sequence of relative frequencies will fluctuate and for large n the deviation from a certain constant will be small.

The law which states that if an experiment is repeated many times under the same conditions, then the relative frequencies of a given event A will be stable, is called

the *law of large numbers* (for more precise formulation and details see Chapter 2.3). This stability involves that the relative frequencies of event A computed from a long record of experiments are practically close to a constant number which is *called the probability of event A* and denoted from now on by $P(A)$.

The probability of a given event is an objective measure that can be measured like physical quantities, such as e.g. temperature, weight, etc. The fact that the relative frequency k_n/n of an event A, for large n, is approximately equal to $P(A)$, the probability of the event A, is denoted by

$$g(A) = \frac{k_n}{n} \approx P(A).$$

The "measurement" of the numerical value of a probability by relative frequencies has to be carried out in such a way that the individual experiments should not have any influence on each other, or in other words: the experiments are "independent", and performed under the same conditions.

Since, in case of a large number of experiments relative frequency does not deviate too much from probability, the basic properties of probability may be derived from those of the relative frequency.

The following relations are always valid for relative frequencies, $g(\,.\,)$:

(1) For any event A, its relative frequency $g_n(A)$ in a sequence of n observations satisfies the following inequality:

$$0 \leqq g_n(A) \leqq 1.$$

(2) The relative frequency of the sure event I (i.e. the one that occurs in each experiment) is

$$g_n(I) = 1,$$

and of an impossible event

$$g_n(\emptyset) = 0.$$

(3) If events A and B are mutually exclusive then

$$g_n(A+B) = g_n(A)+g_n(B).$$

This property holds for the union of finite number of mutually exclusive events, too.

Based on the above mentioned properties of the relative frequency the probability $P(A)$ of an event A is defined as a measure that satisfies the following *axioms:*

Axiom I: For the probability $P(A)$ of any event A it is true that

$$0 \leqq P(A) \leqq 1.$$

Axiom II: The probability of the sure event is $P(I)=1$.

Axiom III: If $A_1, A_2, \ldots, A_n, \ldots$ are mutually exclusive events, i.e. if for $i \neq j$

$$A_i \cdot A_j = \emptyset$$

then

$$P(A_1+A_2+\ldots+A_n+\ldots) = P(A_1)+P(A_2)+\ldots+P(A_n)+\ldots.$$

As direct consequences of the axioms some useful relations will be discussed below that can be applied to calculate the probability of somewhat more composite cases.

Theorem 1. If the probability of event A is $P(A)$ then the probability of the complementary event $\bar{A}$ is

$$P(\bar{A}) = 1 - P(A).$$

To prove this Theorem one should consider that

$$A + \bar{A} = I, \quad \text{and} \quad A \cdot \bar{A} = \emptyset.$$

On the basis of Axioms II and III we have

$$P(I) = P(A + \bar{A}) = P(A) + P(\bar{A}) = 1,$$

from which one obtains Theorem 1. It follows from Theorem 1 that the probability of an impossible event $\emptyset$ is 0*, and

$$P(\emptyset) = P(\bar{I}) = 1 - P(I) = 1 - 1 = 0.$$

Theorem 2. For events A and B the probability that at least one of them will occur is

$$P(A+B) = P(A) + P(B) - P(AB).$$

To prove this statement both $(A+B)$ and B are expressed as the sum of two disjunctive events. It follows from Fig. 5. that

$$A + B = A + \bar{A}B \quad \text{and} \quad B = AB + \bar{A}B.$$

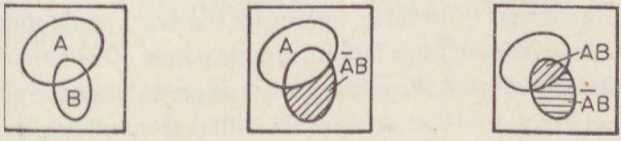


Figure 5

According to Axiom II:

$$P(A+B) = P(A+\bar{A}B) = P(A) + P(\bar{A}B)$$
$$P(B) \quad = P(AB+\bar{A}B) = P(AB) + P(\bar{A}B).$$

By substraction of the two equations

$$P(A+B) - P(B) = P(A) - P(AB)$$

from which the Theorem follows. It also follows from Theorem 2 that

$$P(A+B) \leqq P(A) + P(B),$$

and if

$$P(A) + P(B) > 1,$$

* Note: The fact that the probability of an event A is 0 does not mean that A is an impossible event!

then

$$P(AB) \geqq P(A)+P(B)-1.$$

Theorem 3. If $A_1, A_2, \ldots, A_n$ are arbitrary events then

$$P(A_1+A_2+\ldots+A_n) = \sum_i P(A_i) - \sum_{i_1<i_2} P(A_{i_1}A_{i_2}) + \sum_{i_1<i_2<i_3} P(A_{i_1}A_{i_2}A_{i_3}) - \ldots +$$

$$+(-1)^{n+1} \sum_{i_1<i_2<\ldots<i_n} P(A_{i_1}A_{i_2}\ldots A_{i_n}).$$

This relation can be proved by induction. It follows from Theorem 2 that for $n=2$ the statement is true. Assume that for $n-1$ the statement is also true, i.e.

$$P(A_2+A_3+\ldots+A_n) = \sum_{i=2}^{n} P(A_i) - \sum_{2\leq i_1<i_2} P(A_{i_1}A_{i_2}) + \sum_{2\leq i_1<i_2<i_3} P(A_{i_1}A_{i_2}A_{i_3}) - \ldots.$$

Moreover

$$P(A_1A_2+A_1A_3+\ldots+A_1A_n) = \sum_{i=2}^{n} + P(A_1A_i) -$$

$$- \sum_{2\leq i_1<i_2} P(A_1A_{i_1}A_{i_2}) + \sum_{2\leq i_1<i_2<i_3} P(A_1A_{i_1}A_{i_2}A_{i_3}) - \ldots.$$

Applying Theorem 2:

$$P(A_1+A_2+\ldots+A_n) = P(A_1)+P(A_2+\ldots+A_n) - P(A_1A_2+A_1A_3+\ldots+A_1A_n) =$$

$$= \sum_{i=1}^{n} P(A_i) - \sum_{1\leq i_1<i_2} P(A_{i_1}A_{i_2}) + \sum_{1\leq i_1<i_2<i_3} P(A_{i_1}A_{i_2}A_{i_3}) - \ldots.$$

Before discussing the next theorem an important remark is made concerning sample spaces containing finite or countable infinite sample points. This remark is the following: the probability of any event A equals the sum of probabilities of all sample points contained in A. Let event A be consisting of the different sample points $e_1, e_2, \ldots, e_k$. Obviously, these sample points are mutually exclusive because during one experiment one, and only one sample point may occur, see Figure 6.

$$P(A) = P(e_1+e_2+\ldots+e_k) = \sum_{i=1}^{k} P(e_i).$$

Consider now an experiment that has a finite number n of outcomes. Let these events be $e_1, e_2, \ldots, e_n$, each with the same probability of occurrence:

$$P(e_i) = \frac{1}{n} \quad (i = 1, 2, \ldots, n).$$

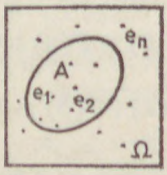


Figure 6

If B is an event related to the experiment and consists of k sample points, then, in accordance with our previous remarks,

$$P(B) = \sum_{e_i \in B} P(e_i) = \frac{k}{n}.$$

By this the next theorem can be obtained:

Theorem 4. If in an experiment one of n sample points may occur and any sample point has the same probability of occurrence, the probability of event B consisting of k sample points is

$$P(B) = \frac{k}{n}.$$

It is to be mentioned in this context that in the early days of probability calculus events of numerous finite outcomes and of equal probabilities were mainly investigated.

The next Theorem expresses the monotonic property of the probability, i.e. the probability of a given set is not less than that of any subset of it.

Theorem 5. If $A \supset B$, then $P(A) \geqq P(B)$. To prove the theorem let us consider that if $A \supset B$, then $A = B + \bar{A}B$, i.e.

$$P(A) = P(B + \bar{A}B) = P(B) + P(\bar{A}B) \geqq P(B).$$

1.1.3. EXAMPLES FOR COMBINATORIAL CALCULATION OF PROBABILITIES

a) Let the experiment be of tossing a coin n times. What is the probability of tossing a head exactly k times?

Assume that n tossing are performed and we assign the number 1 to heads and zero to tails. Because every tossing may have two different outcomes, the number of all possible cases is 2^n (variation with repetition). To determine the number of successful cases i.e. those sequences which contain k1's and $(n-k)$ 0's, respectively, the corresponding quantity has to be counted. The number of such sequences is $\binom{n}{k}$ as altogether this is the number of options one might have in selecting k cells out of n. The probability then is:

$$P_k = \frac{\binom{n}{k}}{2^n}.$$

One can see that the dependence of the probability on k is expressed by the nominator. The sequence of binomial coefficients increases at the beginning with k but later decreases. Its maximum is obtained at $k = \frac{n}{2}$. Consequently, it is most probable that the half of the tossings will be tails (or heads).

b) Random walk along the line

Assume that a particle walks over the integer points along the x-axis starting at the origin and taking one step to the left or one step to the right, equally with a probability of $\frac{1}{2}$. The coordinate of the object will change by $+1$, in case of a step to the right, and by -1 by a step to the left. The path of the object can be illustrated by plotting the number of steps on the horizontal axis, versus displacement on the vertical axis. In case of a step to the right a vector upwards by 45°, while in case of a step to the left a vector downwards by 45° is to be drawn, see Figure 7. What is the probability that after n steps the particle will be at point $x=l$, i.e., in the height $x=l$ in our figure? Assume that during n steps the particle moves a-times to the right and b-times to the left.

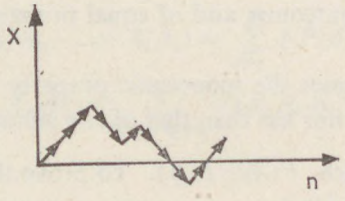


Figure 7

Then:

$$\begin{array}{l} a+b=n \\ a-b=l \\ \hline a=\dfrac{n+l}{2}; \quad b=\dfrac{n-l}{2}. \end{array}$$

That is, the particle is found at point $x=l$ if it has stepped $\frac{n+l}{2}$ times to the right, and $\frac{n-l}{2}$ times to the left. This case may occur in $\binom{n}{\frac{n+l}{2}}$ different ways.

Then the probability in question is:

$$P_l=\frac{\binom{n}{\frac{n+l}{2}}}{2^n}.$$

Since any step to the right or left can be only an integer and $\frac{n+l}{2}$ is integer if n and l are both even or both odd numbers, it follows, that after even number of steps the particle can stand only on a height of an odd number. Particular is the case when the particle returns to its starting point — the origo — that may occur only after even number of steps.

The probability that the moving point will return to the origin after $2n$ steps is

$$P_0 = \frac{\binom{2n}{n}}{2^{2n}} \approx \frac{1}{2^{2n}} \frac{\left(\frac{2n}{e}\right)^{2n} \sqrt{2\pi \cdot 2n}}{\left(\frac{n}{e}\right)^{2n} 2\pi n} = \frac{1}{\sqrt{\pi \cdot n}}.$$

The next example is of particular importance from the point of view of mathematical statistics therefore it is specially recommended to the reader's attention.

c) Returning to the previous example assume that the particle returns to the origin after $2n$ steps. What is the probability that along its path the particle did not reach the point $x=k$? This question can be reformulated as: what is the probability that a sequence of $2n$ length, consisting of $n(+1)$s and $n(-1)$s, has no partial sums greater or equal than k? According to our assumptions all possible sequences of n $(+1)$s and n (-1)s have the same probability.

To answer this question let the following interpretation be introduced. Assume that in case of $2n=8$ the following steps were experienced: $+1, +1, -1, +1, -1, -1, +1, -1$.

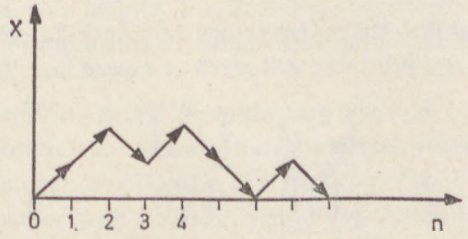


Figure 8

Assign now the vectorial sequence to the sequence of steps as it was done in a previous example. In such a way a path or a trajectory is obtained. The number of possible trajectories for n steps to the right $(+1)$ and to the left (-1) is $\binom{2n}{n}$.

The question now is: how many trajectories, each consisting of $2n$ vectors, can be plotted out of point $(0, 0)$ to point $(2n, 0)$ that will not reach the $X=k$ line?

It is easier to determine the number of trajectories which will reach or intersect line $X=k$. If this number is known then by subtracting it from $\binom{2n}{n}$, from the total number of trajectories, the number of trajectories not reaching line $X=k$ is obtained. The number of trajectories reaching (or intersecting) line $X=k$ can be determined quite easily.

If that part of a trajectory, that reaches line $X=k$, is reflected with respect to the height $X=k$ which begins from the first reaching point then this transformed trajectory will start at $(0, 0)$ and ends at point $(2n, 2k)$. This is true for all trajectories reaching or intersecting line $X=k$. These transformed trajectories correspond to the

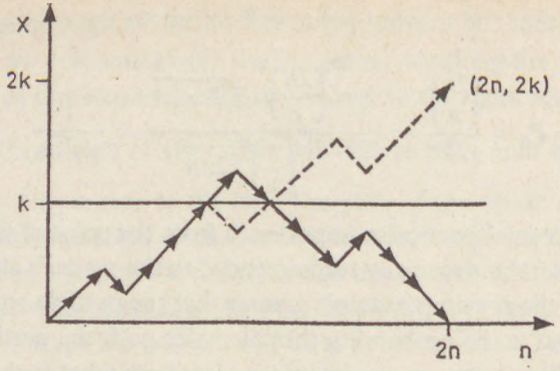


Figure 9

random walk of a particle starting at (0, 0) and found at height $2k$ after $2n$ steps. This is possible in

$$\binom{2n}{n+k} = \binom{2n}{n-k}$$

different ways.

Therefore, the probability that a trajectory of length $2n$ consisting of n steps to the right $(+1)$ and n to the left (-1), will reach or exceed line $X=k$ is:

$$\frac{\binom{2n}{n-k}}{\binom{2n}{n}} = \frac{\binom{2n}{n+k}}{\binom{2n}{n}}.$$

Consequently, the probability that the path will not reach the height $X=k$ is:

$$\frac{\binom{2n}{n} - \binom{2n}{n+k}}{\binom{2n}{n}} = 1 - \frac{\binom{2n}{n+k}}{\binom{2n}{n}}. \tag{1.3}$$

It is to be mentioned here that Eq. (1.3) can hardly be used for practical purposes as the calculation of the binomial coefficients, on the right-hand side of Eq. (1.3), is rather cumbersome. On the other hand, a fairly good asymptotic approximation can be achieved if n is large, i.e.

$$\frac{\binom{2n}{n+k}}{\binom{2n}{n}} = \frac{(2n)!}{(n+k)!(n-k)!} \cdot \frac{(n!)^2}{(2n)!} \approx \frac{\left(\frac{n}{e}\right)^{2n} 2\pi n}{\left(\frac{n+k}{e}\right)^{n+k}\left(\frac{n-k}{e}\right)^{n-k}\sqrt{2\pi(n+k)2\pi(n-k)}} =$$

$$= \frac{1}{\left(1+\frac{k}{n}\right)^{n+k}\left(1-\frac{k}{n}\right)^{n-k}\left[\left(1+\frac{k}{n}\right)\left(1-\frac{k}{n}\right)\right]^{1/2}}.$$

Using the Taylor expansion $\ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\ldots$ it is easy to show that

$$\frac{\binom{2n}{n+k}}{\binom{2n}{n}} \approx e^{-k^2/n}. \tag{1.4}$$

This is a favourable form for statistical applications and will be used extensively in Chapter 6, Section 6.4.3.

d) Runs

A number of interesting and practical statistical analyses can be performed with the aid of the previous random walk problem (See: Section 6.4). One of them is the investigation of 'runs'. A run is defined as the uninterrupted sequence of the same numbers. For example, in sequence

$$+1+1-1-1+1-1-1+1+1+1-1+1-1$$

there are four $(+1)$ runs, with lengths 2, 1, 3, and 1, respectively, and four (-1) runs (lengths 2, 2, 1, 1). (If a random walk is illustrated by a trajectory then a run is an unbroken straight line.)

Assume, that a sequence of length N contains n piece of $(+1)$s and m piece of (-1)s. What is the probability that the total number of runs is exactly $R=2k$.

Let R_{+1} denote the number of $(+1)$ runs and R_{-1} that of the (-1) runs.

The sequence starts either with $(+1)$ or with (-1). Assume, it has started with $(+1)$. Then the last value must be (-1) otherwise condition $R=2k$ will not hold. As there are n pieces of $(+1)$s and because from the point of view of $(+1)$ runs the (-1) runs are just separators, $R_{+1}=k$ may occur as many times as the $(+1)$ sequence consisting of m members can be disaggregated into k parts. Disaggregation can be performed by $k-1$ vertical lines, such as e.g. for $n=8$ and $k=3$ one possibility is:

$$1, 1, 1/1, 1, 1/1, 1.$$

Since there may be $(n-1)$ vertical lines in a $(+1)$ sequence consisting of n members the event of $\{R_{+1}=k\}$ may occur in $\binom{n-1}{k-1}$ different ways. Similarly, event $\{R_{-1}=k\}$ may occur in $\binom{m-1}{k-1}$ different ways. If the sequence consisting of $n(+1)$s is divided by $(k-1)$ lines into k sections, any k decomposition into k parts of the $m(-1)$ sequence can be inserted at the places of the lines. Therefore, if the sequence starts with $(+1)$ and ends up with (-1), event $R=2k$ may occur in $\binom{n-1}{k-1}\binom{m-1}{k-1}$ different ways. Obviously, event $\{R=2k\}$ will occur similarly if the sequence started with (-1) and ended up with $(+1)$. As in N experiments the $n(+1)$s and the

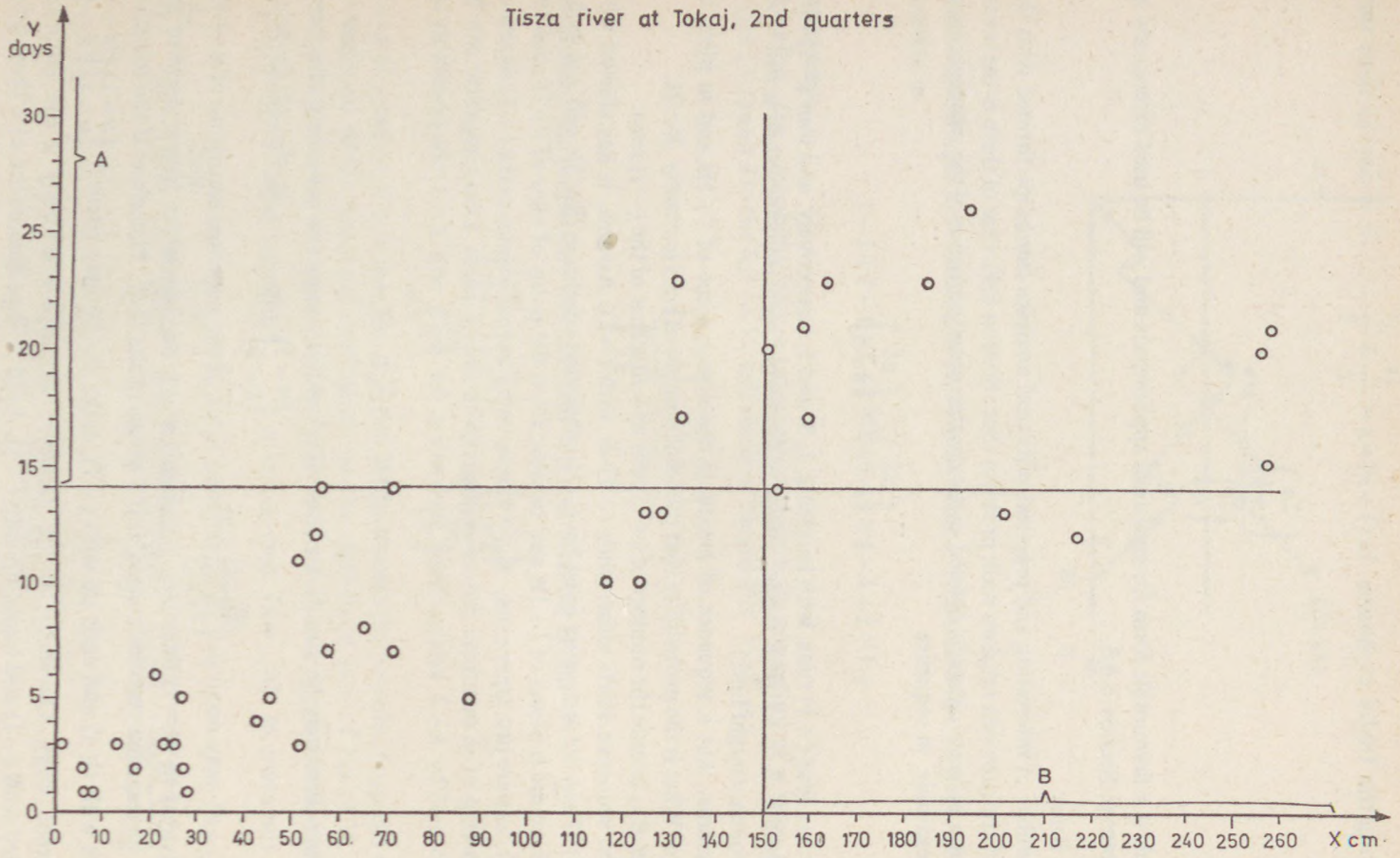

Tisza river at Tokaj, 2nd quarters
Y
days
30
25
20
15
10
5
0
A
B
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260
X cm

Figure 10

$m(-1)$s may occur in $\binom{N}{n}$ different ways, the probability is:

$$P(R=2k)=\frac{2\binom{n-1}{k-1}\binom{m-1}{k-1}}{\binom{N}{n}}.$$

Using similar arguments one can also see that the probability of event $\{R=2k+1\}$ is:

$$P(R=2k+1)=\frac{\binom{n-1}{k}\binom{m-1}{k-1}+\binom{n-1}{k-1}\binom{m-1}{k}}{\binom{N}{n}}.$$

If n and m are large, the computation of the above described probabilities given above is fairly difficult. (For approximations see Section 6.5.1.)

1.1.4. CONDITIONAL PROBABILITY AND INDEPENDENCE

The notion of conditional probability is very important and fundamental both from the point of view of theory and practice. The importance of this notion is further amplified by the fact that different phenomena occurring in Nature, and also in hydrology, are usually not independent from one another. The occurrence of certain phenomena implies that of other phenomena. For instance, heavy precipitation usually implies the occurrence of high flows or stages. Therefore, the probability of the occurrence of certain events is influenced by the occurrence of other events.

As in the discussion of the probability of a given event the notion of relative frequency was used, here the notion of conditional relative frequency will by used to approach the concept of conditional probability.

In *Table T1.* flood data of the Tisza River, Section Tokaj are presented above

Table T.1

TOKAJ
First quarter (From 1st January to 31st March)
c level: 600 cm

Year	X (cm)	Y (days)	$Z(t)$ (Max)
1907	009	003	009
1908	058	010	058
1909	069	004	069
	026	006	
1912	078	005	078
	028	005	
	003	002	

Year	X (cm)	Y (Days)	$Z(t)$ (Max)
1913	027	003	027
1914	024	001	100
	100	020	
1915	036	005	036
1916	088	013	088
	035	014	
	058	006	
1919	086	014	086
1920	112	010	136
	136	016	
1922	092	010	170
	170	009	
1923	004	002	004
1924	168	005	168
1926	173	010	173
1932	028	003	028
1933	020	003	020
1937	122	021	122
1940	173	004	173
1941	008	003	200
	200	013	
	014	003	
1942	030	006	030
1945	047	008	047
1947	040	002	040
1948	101	007	181
	181	020	
1950	030	004	030
1953	147	012	147
1955	080	008	093
	093	008	
	088	004	
1956	000	001	000
1957	059	008	059
	040	006	
1958	156	021	156
1960	000	001	000
1962	074	007	074
1963	071	002	129
	129	007	
1964	010	001	053
	053	002	
1965	055	002	055
1966	153	021	153
1967	229	021	229
	025	001	
1968	135	008	135
1970	128	005	128

SZOLNOK

First quarter (From 1st January to 31st March)

c level: 600 cm

Year	*X* (cm)	*Y* (days)	*Z*(*t*) (Max)
1903	004	001	004
1908	029	007	029
1909	000	001	000
1912	002	003	002
1914	100	015	100
1915	038	007	038
1916	178	024	178
	088	021	
1919	085	012	085
1920	104	016	116
	116	016	
1922	134	016	134
1926	178	023	178
1931	000	001	000
1935	018	004	018
1937	150	024	150
1940	004	002	066
	066	004	
1941	222	034	222
1942	128	018	128
1945	001	002	001
1946	017	001	017
1947	024	010	033
	033	009	
1948	184	034	184
1953	201	023	201
1955	020	006	045
	045	010	
1957	045	012	045
	013	005	
1958	108	031	108
1962	035	007	035
1963	091	016	091
1964	003	001	003
1965	000	003	039
	017	004	
	039	011	
1966	255	039	255
1967	281	028	281
1968	063	011	063
1969	059	008	059
1970	019	003	065
	056	008	
	065	004	

TOKAJ

Second quarter (From 1st April to 30th June)

c level: 600 cm

Year	*X* (cm)	*Y* (days)	*Z*(*t*) (Max)
1907	159	021	159
	056	014	
1909	021	006	021
1912	125	013	125
1913	027	002	027
1914	072	007	072
1915	058	007	058
1916	052	003	052
1922	160	017	160
	008	001	
	024	003	
1924	202	013	202
	128	013	
1932	256	020	256
1933	028	001	028
1935	006	001	006
1937	017	002	017
1940	217	012	217
	042	004	
1941	133	017	185
	185	023	
	116	010	
1942	024	003	024
1944	051	011	051
1945	065	008	065
	026	005	
1951	013	003	013
1952	164	023	164
1955	088	005	088
1956	071	014	071
1958	045	005	045
	006	002	
	002	003	
1962	194	026	194
1964	257	015	257
1965	123	010	123
1967	132	023	132
1968	055	012	055
1970	151	020	258
	258	021	
	153	014	

SZOLNOK

Second quarter (From 1st April to 30th June)

c level: 600 cm

Year	X (cm)	Y (days)	$Z(t)$ (Max)
1907	088	019	088
	006	009	
1912	063	012	063
1914	065	016	065
1915	032	012	032
1916	052	011	052
1919	023	002	023
1920	006	001	006
1922	134	032	134
1924	196	049	196
1932	244	025	244
1937	100	013	100
1940	230	038	230
1941	206	065	206
1942	020	008	020
1944	012	010	012
1952	083	025	083
1956	028	013	028
1958	006	005	007
	007	009	
1962	185	030	185
1964	203	022	203
1965	143	017	143
1967	174	035	174
1968	023	006	023
1970	259	091	259

SZEGED

First quarter (From 1st January to 31st March)

c level: 600 cm

Year	X (cm)	Y (days)	$Z(t)$ (Max)
1901	159	018	059
1902	040	015	040
1906	018	006	018
1908	059	021	059
1909	084	012	084
1912	102	034	102
1913	033	005	033
1914	248	020	248
1915	052	016	180
	180	012	

Year	X (cm)	Y (days)	Y (days)
1916	271	030	271
	204	026	
1917	090	006	090
1920	188	028	188
	183	020	
1922	184	021	184
1923	000	001	117
	117	017	
1924	043	002	043
1926	232	026	232
	010	006	
1931	083	012	083
1932	012	004	147
1933	147	009	
	020	004	020
1934	006	004	006
1935	054	009	054
1937	181	026	181
1940	173	013	173
1941	010	004	269
	104	011	
	269	041	
1942	260	030	260
1945	032	006	032
1946	010	004	010
1947	087	022	087
1948	199	037	199
1953	186	026	186
1955	079	011	137
	137	016	
	020	004	
	026	002	
1956	036	007	036
	032	007	
1957	043	014	043
	005	004	
1958	210	030	210
1960	059	008	059
1962	002	002	002
1963	066	015	066
1965	014	003	045
	045	008	
	042	009	
1966	278	042	278
1967	270	027	270
1968	054	006	058
	058	012	
1969	106	013	106
	004	002	
1970	172	015	172
	147	023	

SZEGED
Second quarter (From 1st April to 30th June)
c level: 600 cm

Year	X (cm)	Y (days)	$Z(t)$ (Max)
1901	029	005	029
1902	014	003	014
1907	108	042	108
1912	072	010	072
	034	010	
1914	128	022	128
1915	110	035	110
1916	073	013	073
1919	266	049	266
1920	016	002	016
1922	124	036	124
1924	220	051	220
1932	273	042	273
1937	053	011	053
1940	197	038	197
	040	008	
	028	005	
1941	204	068	204
1942	038	007	060
	051	011	
	060	014	
1944	004	003	004
1952	002	005	002
1956	039	010	039
1958	037	007	066
	066	025	
1962	170	033	170
1964	114	019	114
1965	098	015	098
1967	134	041	134
1970	309	091	309

a given level of 600 cm together with the duration (in days) of these events. These data are for the 2nd quarter (1 April—30 June) of the year for the period 1901—1970.

Flood exceedance X and duration Y are plotted on the plane by co-ordinates $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. See Fig. 10.

Let event A be defined as the flood duration exceeding 14 days. Figure 13 indicates that A has occurred 11 times during 41 flood events. Thus, the relative frequency of event A is:

$$g(A) = \frac{k(A)}{n} = \frac{11}{41} \approx \frac{1}{4}.$$

Let now event B defined as the flood exceedance above 150 cm. In this example, as it can be seen from the figure, the relative frequency of event B is $k(B)=12$. Consider now the floods larger than 150 cm and lasting longer than 14 days, i.e. consider the occurrence of event AB. This is:

$$k(AB) = 9.$$

The relative frequency of A, under the condition that event B occurred, is a conditional relative frequency and is defined as

$$g(A|B) = \frac{k(AB)}{k(B)}.$$

In our example:

$$g(A|B) = \frac{9}{12} = \frac{3}{4}.$$

The conditional relative frequency of event A is about three times greater than the relative frequency $g(A)$ of event A, in this case. That is, three quarters of the floods exceeding 150 cm lasted more than two weeks. As the flood peak usually appears at the half time of the duration in a given flood situation one can predict the whole duration of a flood event by knowing its conditional relative frequency.

Consequently, conditional relative frequency signifies that, considering only those events when B has occurred, what is the percentage of event A occurring simultaneously with event B?

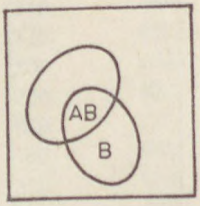


Figure 11

Many experiences justified that if the number of experiments increased then the conditional relative frequency showed a similar stability as the relative frequency itself. The number around which conditional relative frequency is oscillating is called conditional probability and is denoted by $P(A|B)$.

As

$$g(A|B) = \frac{k(AB)}{k(B)} = \frac{\frac{k(AB)}{n}}{\frac{k(B)}{n}} \approx \frac{P(AB)}{P(B)},$$

the conditional probability $P(A|B)$ is defined by the following expression

$$P(A|B) = \frac{P(AB)}{P(B)}, \quad P(B) > 0.$$

From this definition

$$P(AB) = P(A|B)P(B). \tag{1.5}$$

Equation (1.5) is called the rule of multiplication of probabilities.

Event A is said to be independent of event B if

$$P(A|B) = P(A).$$

Then

$$P(A) = \frac{P(AB)}{P(B)}$$

that is

$$P(AB) = P(A)P(B). \tag{1.6}$$

The probability of the product of *independent* events is equal to the product of the corresponding probabilities.

The conditional probability of event B with respect to event A can be defined in the same way:

$$P(B|A) = \frac{P(AB)}{P(A)}, \quad P(A) > 0. \tag{1.7}$$

By comparing Eqs. (1.5) and (1.7) one obtains

$$P(A|B)P(B) = P(B|A)P(A)$$

yielding

$$\frac{P(A|B)}{P(B|A)} = \frac{P(A)}{P(B)}.$$

It is easy to show that if event A is independent of event B then, in turn, B is also independent of A. If follows from

$$P(A|B) = P(A)$$

that

$$P(B|A) = P(B)$$

as, based on Eq. (1.7),

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B).$$

It is easy to see that if A is independent of B then A is independent of $\bar{B}$, $\bar{A}$ is independent of B, and also $\bar{A}$ is independent of $\bar{B}$.

Therefore, Eq. (1.6) is to be considered as the definition of independence between events A and B. Independence of more than two events may be defined similarly. It is necessary, however, to be careful if independence of more than two events, e.g., of A, B and C is to be defined. Let now three events, A, B and C be represented in the following diagrams:

As it can be seen

$$P(A) = P(B) = P(C) = \frac{1}{2}.$$

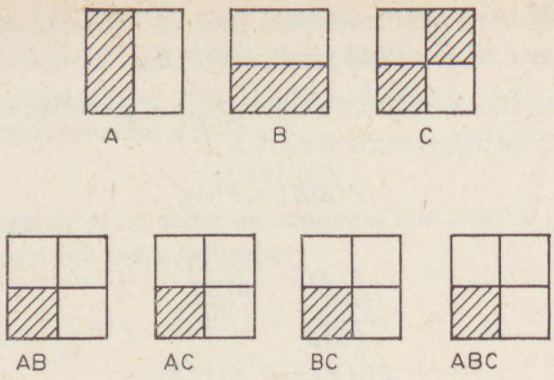


Figure 12

It is easy to understand that

$$P(AB) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A)P(A)$$

$$P(AC) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A)P(C)$$

$$P(BC) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(B)P(C).$$

As a conclusion A, B and C are pairwise mutually *independent*. However,

$$P(ABC) = \frac{1}{4} \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(A)P(B)P(C)$$

that is, the three events considered jointly are not independent. Therefore, if $n>2$ the independence of events $A_1, A_2, \ldots A_n$ is defined as follows:

$A_1, A_2, \ldots A_n$ are *completely independent*, or briefly independent, if the following relations are valid:

$$P(A_i A_j) = P(A_i)P(A_j) \qquad i<j$$

$$P(A_i A_j A_k) = P(A_i)P(A_j)P(A_k) \qquad i<j<k$$

$$\ldots$$

$$P(A_1 A_2 \ldots A_n) = P(A_1)P(A_2)\ldots P(A_n).$$

These relations require the validity of conditions of number

$$\binom{n}{2} + \binom{n}{3} + \ldots + \binom{n}{n} = 2^n - n - 1,$$

the validity of which is due to the binomial law, namely,

$$\sum_0^n \binom{n}{k} = 2^n; \quad \binom{n}{0} = 1; \quad \binom{n}{1} = n.$$

It can be shown that if events $A_1, A_2, \ldots A_n$ are independent, then if any or all of these are replaced by their complementary event(s) $\bar{A}_i$ the system of events obtained is still independent. The rule of multiplication of probabilities expressed by Eq. (1.5), can be generalized to n events. This general rule is as follows:

If $A_1, A_2, \ldots, A_n$ are arbitrary events then

$$P(A_1 A_2 \ldots A_n) = P(A_n|A_1 A_2 \ldots A_{n-1}) P(A_{n-1}|A_1 A_2 \ldots A_{n-2}) \ldots P(A_2|A_1) P(A_1).$$

It is assumed here that the probabilities of the conditional events are all positive. The proof of the theorem is trivial:

$$P(A_1 A_2 \ldots A_{n-1} A_n) = P(A_n|A_1 A_2 \ldots A_{n-1}) P(A_1 A_2 \ldots A_{n-1}),$$
$$P(A_1 A_2 \ldots A_{n-1}) = P(A_{n-1}|A_1 A_2 \ldots A_{n-2}) P(A_1 A_2 \ldots A_{n-2}),$$
$$\vdots$$
$$P(A_1 A_2) = P(A_2|A_1) P(A_1).$$

Beyond the notion of independence of events also the notion of independence of experiments will frequently be used. This latter is somewhat more general than the previous one. Two experiments are regarded as being independent if their outcomes are not influencing each other. This means, that any event in the frame of the first experiment is independent of any event in connection with the second experiment. Textbooks on probability theory usually present the following simple examples of repeated tossing of dice or coins to represent sequences of independent experiments. Similarly, if balls marked with numbers are put in an urn, one is selected randomly, its value is preserved, then replaced in that urn, and the whole set of balls is shaken up before the next trial; this will also form an independent sequence of experiments. This model can be used, for instance, to select elements from a demographic (industrial, etc.) population in a random way, as, e.g., drawing from a lot.

In the hydrological practice usually those observations are considered independent experiments which are fairly far away from each other on the time scale. Examples are: river stages of a particular day of the year, annual stage maxima, annual mean flows, number of floods in the individual years, etc.

An important relationship will be here derived, called the total probability rule which will be referred to frequently in the following.

Let $B_1, B_2, \ldots, B_n$ be a *complete system of events*, i.e.

$$B_1 + B_2 + \ldots + B_n = I$$

and

$$B_i B_j = \emptyset, \quad \text{if} \quad i \neq j,$$

which means that during our experiment at least one but only one B_i will occur. Let A be an arbitrary event, then

$$P(A) = P(A \cdot I) = P\{A(B_1 + B_2 + \ldots + B_n)\} = P(AB_1 + AB_2 + \ldots + AB_n).$$

The events within the brackets are mutually exclusive thus

$$P(A) = P(AB_1)+P(AB_2)+\ldots+P(AB_n)$$

since B_i and B_j are disjoint sets therefore subsets AB_i and AB_j are also disjoint and mutually exclusive, see Figure 13.

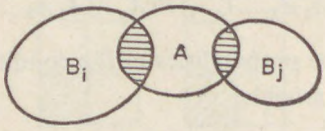


Figure 13

Based on Eq. (1.5)

$$P(AB_i) = P(A|B_i)P(B_i)$$

and the relation

$$P(A) = P(A|B_1)P(B_1)+\ldots+P(A|B_n)P(B_n) = \sum_{j=1}^{n} P(A|B_j)P(B_j) \tag{1.8}$$

will be obtained.

Expression (1.8) is called the theorem of total probability. If the conditional probabilities of event A with respect to all B_j events are known together with the probabilities of the B_j events, $j=1, 2, \ldots, n$, then the conditional probabilities of events B_j with respect to A may be calculated.

Using Eq. (1.7) one would obtain

$$P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{P(A)} . \tag{1.9}$$

Substituting Eq. (1.8) into the denominator of Eq. (1.9)

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)} . \tag{1.10}$$

Expression (1.10) is called the *theorem of Bayes*. This formula given by Bayes plays an important role in statistics, particularly in decision theory.

At the end of this Chapter a remark is made concerning complete system of events. A complete system of events $B_1, B_2, \ldots, B_n$ is a decomposition of a basic set I into disjoint subsets in such a way that all the elements of I are in one of the subsets B_j, and nowhere else. The question arises then: given a basis set I, how many complete systems of events may be formulated? If I consists of an infinite number of elements (points) then the number of complete system of events, or in other words, the number of set partitions is, obviously, also infinite. If I is a finite set, say $I_n=\{1, 2, \ldots, n\}$, then the number of complete system of events, that can be generated from I_n, may

be determined by recursion. Denote T_n the number of complete system of events generated from set I_n. Let $T_0=1$. Then the following recursive relation holds:

$$T_n \sum_{k=0}^{n} \binom{n}{k} T_{n-k} = \sum_{k=0}^{n} \binom{n}{k} T_k.$$

To derive this relation we consider all partitions of I_{n+1} and classify them first according to the location of the element $(n+1)$. Suppose that this element is added to a subset of I_n consisting of k elements. The number of n possible partitions of this kind is $\binom{n}{k} T_{n-k}$, i.e., this subset can be taken from the set I_n in $\binom{n}{k}$ different ways and to each subset belong all possible partitions of the rest of $(n-k)$ elements which is T_{n-k}.

CHAPTER 2

2.1. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

2.1.1. THE NOTION OF RANDOM VARIABLE AND PROBABILITY DISTRIBUTION

In engineering practice the outcome of experiments is usually expressed in terms of numerical forms. In the majority of cases, the result of an experiment is numerical itself. For instance, if stage measurements are at stake, the outcome of the experiment is a number; at the same time if the measurement is repeated at different time instants, these numbers will display random fluctuation since there are many causative factors which may influence a particular water stage. The stages of a river measured at the same section, but at different time instants; the number of rainy days in a particular month (observed in different years) are all numbers displaying random fluctuation.

Quantities which depend upon randomness are said to be random quantities or random variables. The numbers which are results of experiments (all possible numerical outcomes of experiments) form the sample space of a random variable.

If random variable X denotes the stage of a river at a given cross-section then the value of X may be any point in a feasible interval.

If random variable Y denotes the number of rainy days, say in May, then its possible values, the sample space, are the numbers 0, 1, 2, ..., 31.

These two examples show that there are several types of random variables. Random variables whose values may be anywhere along the line (or in an interval) are said to be continuous random variables. Measured data are usually continuous random variables. A random variable that may have a finite number or countable infinite number of values is a discrete random variable. Discrete random variables occurring in practice usually have non-negative integer values. There are random variables which are neither continuous nor discrete. These variables belong to the class of mixed random variables.

From a mathematical point of view random variable X is a function defined on a space Ω of elementary events ω. The value of this function depends on the occurrence of a particular elementary event ω: $X=X(\omega)$.

The space of elementary events Ω is essentially the mathematical model of an experiment.

Let the experiment be the tossing of a coin n times. Assume, we assign 1 to the tails and 0 to the heads. Then the possible outcomes of this experiment are the ele-

mentary events:

$$\begin{aligned}
\omega_1 &= 000\ldots00 \\
\omega_2 &= 000\ldots01 \quad \Omega = (\omega_1, \omega_2, \ldots, \omega_{2^n}\} \\
\omega_3 &= 000\ldots10 \\
&\vdots \\
\omega_n &= 100\ldots00 \\
&\vdots \\
\omega_{2^n} &= 111\ldots11.
\end{aligned}$$

Let random variable $X(\omega)$ denote the number of tails through n tossings. Then $X(\omega_1)=0$, $X(\omega_2)=1$, $X(\omega_3)=1$, ..., $X(\omega_{2^n})=n$. If ω_i is an elementary event containing $k(+1)$s, and $n-k$ 0's then $X(\omega_i)=k$. The possible values of random variable X are then the numbers 0, 1, 2, ..., n. The set $\mathfrak{X}=\{0, 1, 2, \ldots, n\}$ of the possible values of the random variable X is called *sample space*. Let A_k denote the set of elementary events ω_i which consists of exactly $k(+1)$s—obviously there are $\binom{n}{k}$ of this type. Then

$$P(X = k) = P(A_k) \quad (k = 0, 1, 2, \ldots, n).$$

Assume, the tossings are performed by a homogeneous, regular coin. Then both, tails and heads have the same probability of occurrence, i.e. $\frac{1}{2}$. As the outcomes of the tossings are independent, the probability of any elementary event ω_i is

$$P(\omega_i) = \frac{1}{2^n} \quad (i = 1, 2, \ldots, 2^n).$$

Thus

$$P(X = k) = P(A_k) = \sum_{\omega_i \in A_k} P(\omega_i) = \frac{\binom{n}{k}}{2^n}.$$

It can be seen from our example that random variable $X(\omega)$ defined as a function over the space of elementary events does not necessarily assign different values to the different elementary events. Function $X(\omega_i)$ assigns the same value k to points i of the $\binom{n}{k}$ elementary events of the above mentioned event A_k. The probability of taking a given k value for random variable $X(\omega)$ is equal to the probability measure $P(A)$ of set A of the elementary events ω for which

$$X(\omega) = k.$$

Random variable X assigned to the tossing of a coin can have a finite number of values. Possible values for X are the numbers 0, 1, 2, ..., n. If we know the proba-

bility of all possible values of random variable X, then also the *distribution* of the random variable is known. In our example

$$P(X = k) = \binom{n}{k} \frac{1}{2^n} \quad (k = 0, 1, \ldots, n).$$

If, for instance, the experiment is the tossing of a fair coin five times, then

$$P(X = 0) = \binom{5}{0} \frac{1}{2^5} = \frac{1}{32}$$

$$P(X = 1) = \binom{5}{1} \frac{1}{2^5} = \frac{5}{32}$$

$$P(X = 2) = \binom{5}{2} \frac{1}{2^5} = \frac{10}{32}$$

$$P(X = 3) = \binom{5}{3} \frac{1}{2^5} = \frac{10}{32}$$

$$P(X = 4) = \binom{5}{4} \frac{1}{2^5} = \frac{5}{32}$$

$$P(X = 5) = \binom{5}{5} \frac{1}{2^5} = \frac{1}{32}.$$

Obviously,

$$\sum_{k=0}^{5} P(X = k) = 1.$$

This probability distribution can also be illustrated graphically by aid of a probability diagram, see Figure 14.

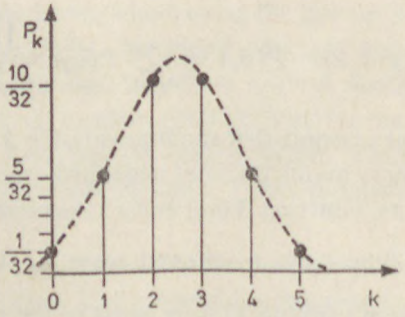


Figure 14

Let the experiment be throwing of a die until the number six is obtained.

Let the random variable Y denote the number of tossings until the six is obtained. Possible values of Y are then $1, 2, \ldots$. Random variable Y has a countable number of

values. If $P(1)=p$, $P(0)=1-p=q$ $\left(\text{for a fair die } p=\frac{1}{6},\ q=\frac{5}{6}\right)$ then the probability that Y will have the value of k is

$$P(Y=k)=q^{k-1}\cdot p \quad (k=1,2,\ldots).$$

Naturally,

$$\sum_{k=1}^{\infty} P(Y=k)=\sum_{1}^{\infty} q^{k-1}\cdot p=p(1+q+q^2+\ldots)=\frac{p}{1-q}=1.$$

The probability distribution of the random variable Y is shown in Figure 15.

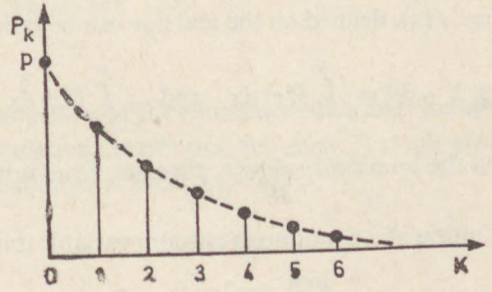


Figure 15

A random variable X having finite or a countable number of values is said to have a discrete distribution, or simply a discrete random variable. The distribution of a discrete random variable is specified by defining probabilities to all its possible values.

Discrete random variables occur in hydrology often. Examples are: the number of rainy days in a month, the number of exceeding a threshold level c in water stages during a given time interval, etc.

The discrete random variables discussed so far in the previous examples had all non-negativ integer numbers for their possible values. In practice, however, particularly in the statistical analysis of observation data one may encounter discrete random variables which are not necessarily integer values.

Let X be a discrete random variable, with possible values $x_1, x_2, \ldots, x_n\ldots$ (finite or countable infinite) and these values have the probabilities

$$P(X=x_1)=p_1;\ P(X=x_2)=p_2,\ \ldots,\ P(X=x_n)=p_n,\ \ldots.$$

If the probability of the event that the value of X lies between limits a and b is of interest, then the probabilities of all x_i-s for which $a\leqq x_i<b$ holds are to be summed up:

$$P(a\leqq X<b)=\sum_{x_i\in(a,b)} p_i.$$

In hydrological practice the analysis is often confined to random variables, such as, e.g. the stage of a river in given space and time, river flow, or the value of exceedance

above a given threshold level during floods, etc. The value of random variables like these can fall in any interval of the real line having a continuum of possible values, and therefore — as mentioned earlier — these are called continuous random variables. If, for example, random variable X is the stage at a given section of a river then usually that probability is sought that the stage will be between 600 and 700 cm or it is less than 850 cm. The distribution of random variable X is known if for any interval (a, b) the probability of event $\{a \leqq X < b\}$

$$P(a \leqq X < b)$$

can be specified.

The distribution of a random variable X will be said to be continuous if a non-negative, integrable function $f(x)$, defined on the real line can be assigned to it, for which:

$$P(a \leqq X < b) = \int_a^b f(x)\,dx \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Function $f(x)$ is called the *probability density function,* from now on pdf, of the random variable X.

From the above definition of a continuous random variable it follows that

$$P(X = x) = 0.$$

Moreover, the probability that the value of X will be in the small interval Δx around x is:

$$P\left(x - \frac{\Delta x}{2} \leqq X < x + \frac{\Delta x}{2}\right) \approx f(x)\,dx.$$

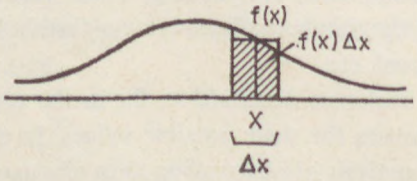


Figure 16

To calculate the probability $P(a \leqq X < b)$ it is sufficient to know what the probability is that X is less than any x.

Probability $P(X < x)$ is a function of variable x. Let this function be denoted by $F(x)$. Obviously,

$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

Function $F(x)$ is called the *cumulative distribution function* (cdf) of the random variable X. An ordinate of the cumulative distribution function $F(x)$ specifies, for any real x, the probability of finding X less than x.

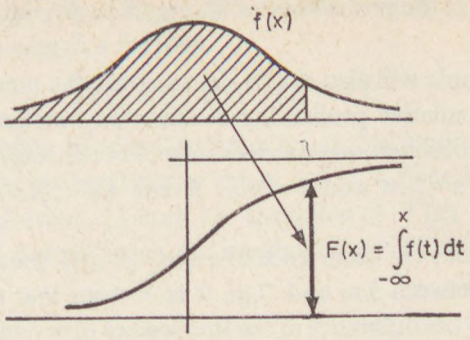


Figure 17

Consider now the properties of the cumulative distribution function $F(x)$.

(1) Values of a cumulative distribution function $F(x)$ are always between 0 and 1 as $F(x)$ expresses probability (of event $\{X<x\}$):

$$0 \leqq F(x) \leqq 1.$$

(2) The cumulative distribution function $F(x)$ is a monotonous non-decreasing function of x, i.e. if $b>a$, $F(b) \geqq F(a)$.

It is easy to show the validity of this property as for $b>a$ the event $\{X<b\}$ will always occur if $\{X<a\}$ but also if $a \leqq X<b$. E.g.

$$\{x < b\} = \{X < a\} \cup \{a \leqq X < b\}.$$

There are mutually exclusive events on the right-hand-side. Therefore, based on Axiom III.

$$P(X < b) = P(X < a) + P(a \leqq X < b).$$

That is

$$F(b) = F(a) + P(a \leqq X < b).$$

From this last relationship one obtains

$$P(a \leqq X < b) = F(b) - F(a).$$

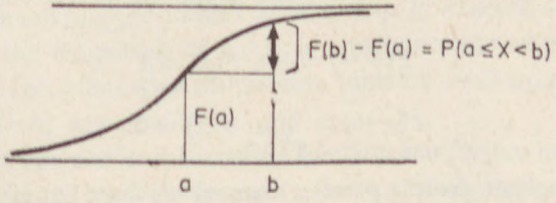


Figure 18

So, by knowing $F(x)$ the probability of finding X in any interval (a, b) can also be specified.

The following example will also clearly demonstrate the monotonous non-decreasing nature of the cumulative probability distribution function $F(x)$. Assume, that our experiment is the observation of a river stage. Let the stage be defined by the random variable X. Let $a=5$ m and $b=7$ m. Event $\{X<5\}$ will occur if the stage is less than 5 m.

Event $\{X<5\}$ implies the occurrence of event $\{X<7\}$ though this latter will also occur if the stage is between 5 m and 7 m. It is obvious that the frequency of event $\{5\leqq X<7\}$ is equal to the difference of the frequencies for events $\{X<7\}$ and $\{X<5\}$. Naturally, the same is true for the relative frequencies of the same events. According

Table 2.1.

Frequency of the yearly maximal water-level of the River Duna at Bratislava 1892—1961

Maximal water-level between	k_i (frequency)
450—499	1
500—549	1
550—599	7
600—649	11
650—699	15
700—749	12
750—799	15
800—849	2
850—899	3
900—949	1
950—999	2
	$k_i=70$

to Table 2.1 the relative frequency of event $\{X<5\}$ is 1/70, of event $\{5\leqq X<7\}$ it is 34/70 and that of event $\{X<7\}$ is 35/70. As relations valid for relative frequencies coming from a given sample are equally valid for probabilities, one has

$$P(X<7)=P(X<5)+P(5\leqq X<7).$$

(3) Random variables may have, with probability 1, finite values only; event $\{X<+\infty\}$ is said to be a sure event, while event $\{X<-\infty\}$ an impossible event. Thus

$$F(+\infty)=\lim_{x\to\infty}F(x)=1$$

$$F(-\infty)=\lim_{x\to-\infty}F(x)=0.$$

Random variables playing role in practice are not only finite but usually also bounded. This implies that there exist numbers like k and K: $-\infty<k<K<+\infty$ for which

$F(k)=0$ and $F(K)=1$. If, for example, X is a random variable denoting the number of points gained in throwing of a die then

$$F(1) = 0 \quad \text{and} \quad F(7) = 1.$$

The cumulative probability distribution function of this random variable X may be easily constructed. If the probabilities of values 1, 2, 3, 4, 5 and 6 are all equal to 1/6 then the probability of event $\{X<x\}$ i.e. the value of $F(x)$ will be equal to 1/6 times 1, 2, 3, 4, 5 or 6 depending on the number of events that are to the left of x, see Figure 19.

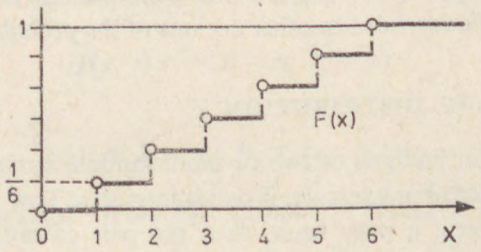


Figure 19

The cumulative distribution function then looks like:

If a is fixed and b is approaching a in expression $P(a \leqq X<b)=F(b)-F(a)$, then:

$$F(a+0)-F(a) = P(X = a).$$

If, however, b is fixed and a is approaching b, then:

$$F(b)-F(b-0) = 0.$$

(4) This latter relationship indicates that the cumulative distribution function is continuous from the left for all values of x. It is, however, not continuous from the right in those $x=a$ points where $P(x=a)\neq 0$.

At this point the cumulative distribution function has a jump of a magnitude $P(X=a)$. In case of dice-throwing the probability is $P(x=a)=\frac{1}{6}$ at points $a=$ $=1, 2, 3, 4, 5, 6$ and, as shown in Fig. 23., the cumulative distribution function has a jump of 1/6.

Similarly, a discontinuous cumulative distribution function is obtained if the random variable X representing the number of rainy days in a month or a year is considered. In this case, the magnitude of jumps is usually not the same. The magnitude of discontinuity of the cumulative distribution function $F(x)$ would equal the probability of event $\{X=k\}$ at points $x=k$.

The meaning of a cumulative distribution function can further be explained by the following example. Let us consider the axis x as a bar of infinitesimal small diameter, on which a mass of total amount 1 is distributed. The amount of mass belonging to

the different sections of the bar may be very different. There might be some parts without any mass (on the other hand certain isolated points may be the carriers of positive mass quantities but let us take it out of consideration for a moment). If $F(x)$ denotes the quantity of mass lying on the left side of x then $F(x)$ possesses the properties of a cumulative distribution function. The quantity of mass falling in interval (a, b) is $F(b)-F(a)$; it is equivalent to the probability that the value of the random variable X is found in interval (a, b), if the mass is replaced by probability.

In case of discrete distributions mass is placed only to points x_i for which $P(x=x_i)=p_i>0$ and the magnitude of the mass is exactly p_i. This is how one can talk about probability mass. It will be shown later that this terminology borrowed from mechanics may be extended to other notions of the probability theory.

2.1.2. MULTIVARIATE DISTRIBUTION

In many cases the joint analysis of two or more random variables related to some phenomenon is required. For example, if one is interested how river stages influence groundwater levels along a river bank, then the pair of random variable (X, Y) of river stage X and groundwater level Y, or the random vector $\vec{Z}=(X, Y)$ is to be analyzed to find the distribution that serves as a basis for identifying any relationship between the variables.

If river flow, precipitation and groundwater level are considered jointly, the distribution of a random vector

$$\vec{X} = (X_1, X_2, X_3)$$

is to be determined or in other words: the joint distribution of random variables X_1, X_2 and X_3 must be analyzed.

If n measurements $X_1, X_2, \ldots, X_n$ are performed with respect to quantity X then the joint distribution of these measurements, or the distribution of random vector

$$\vec{X} = (X_1, X_2, \ldots, X_n)$$

is sought. More often, instead of the distribution of random vector $(X_1, X_2, \ldots, X_n)$ the distribution of random variable $Y=f(X_1, X_2, \ldots, X_n)$ is of interest, where f is a function with n variables. In the most important areas of mathematical statistics, such as decision theory and testing of hypotheses such functions of random argument are of primary importance.

Let us discuss first the distribution of two-dimensional random vectors. Let X and Y be random variables and consider the distribution of random points (X, Y) over the plane.

Let $P(a_1 \leqq X < a_2, b_1 \leqq Y < b_2)$ denote the probability that the random point, i.e. random vector $\vec{Z}=(X, Y)$ will fall into the rectangular area defined by a_1, a_2, b_1 and b_2, see Figure 20.

If this probability is known for every a_1, a_2, b_1 and b_2, then, one might say that the joint probability distribution of random variables X and Y is known.

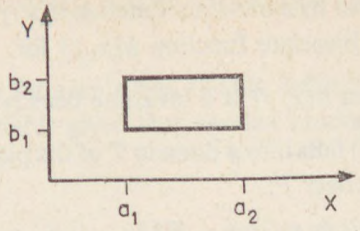


Figure 20

The joint distribution of variables X and Y is defined by the joint cumulative distribution function:

$$H(x, y) = P(X < x,\ Y < y).$$

It is easy to see that

(2.1) $$P(a_1 \leqq X < a_2,\ b_1 \leqq Y < b_2) =$$

$$= H(a_2, b_2) - H(a_1, b_2) - H(a_2, b_1) + H(a_1, b_1) \geqq 0,$$

which shows that if the bivariate cumulative distribution function $H(x, y)$ is known then the probability measure of any rectangular area on the plane can be calculated.

It is mentioned here without proof that $H(x, y)$ is a monotonous non-decreasing function of both variables X and Y and

(2.2) $$H(+\infty, +\infty) = 1, \quad H(x, -\infty) = H(-\infty, y) = 0.$$

The mathematical treatment of two-dimensional probability distributions is essentially analog with the analysis of unit mass distributed over a plane (x, y). The quantity of the mass belonging to a rectangle corresponds to the probability that a point (x, y) falls into this particular area. The value of $H(x, y)$ in a given point (x_0, y_0) represents now the quantity of mass lying in the quadrant defined by $x < x_0$, $y < y_0$.

If the joint cumulative distribution $H(x, y)$ of (X, Y) is known then the distributions $F(x)$ and $G(y)$ of variable X and Y can be easily derived:

(2.3) $$H(x, +\infty) = P(X < x,\ Y < +\infty) = P(X < x) = F(x),$$

$$H(+\infty, y) = P(X < +\infty,\ Y < y) = P(Y < y) = G(y).$$

$F(x)$ and $G(y)$ are called the marginal distributions of $H(x, y)$.

If $a_1 = x_1$, $b_1 = y_1$, $a_2 = +\infty$ and $b_2 = +\infty$ are substituted in Eq. (2.1) then, by using Eqs. (2.2) and (2.3) one would have

(2.4) $$P(X > x,\ Y > y) = 1 - F(x) - G(y) + H(x, y).$$

From the point of view of applications the most important case is when both X and Y are continuously distributed with density functions $f(x)$ and $g(y)$, respectively, and the random vector $\vec{Z} = (X, Y)$ is also continuously distributed over the plane. In this case, similarly to the analogy of continuous mass distribution, the probability

density in point (x, y) is given by a bivariate function $h(x, y)$, being a two-dimensional surface. If there exists a bivariate function $h(x, y)$ for which $h(x, y) \geqq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)\,dx\,dy = 1$ then $h(x, y)$ is a bivariate density function. The probability that a random vector (X, Y) falls into a domain T of the plane is given by the (double) integral of function $h(x, y)$ over T:

$$P((X, Y) \in T) = \iint_T h(x, y)\,dx\,dy.$$

If this domain is the quadrant defined by $(X < x, Y < y)$ then

$$P(X < x, Y < y) = \int_{-\infty}^{x} \int_{-\infty}^{y} h(u, v)\,du\,dv = H(x, y), \tag{2.5}$$

which is a bivariate cumulative distribution function.

The following relation holds between the bivariate cumulative distribution function and the bivariate density function:

$$h(x, y) = \frac{\partial^2 H(x, y)}{\partial x \partial y}.$$

That is, the bivariate density function, if it exists, is the second mixed partial derivate of the bivariate cumulative distribution function.

If the bivariate density function $h(x, y)$ is known then the marginal distributions of variables X and Y may also be determined:

$$F(x) = H(x, +\infty) = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} h(u, v)\,du\,dv = \int_{-\infty}^{x} \left[\int_{-\infty}^{+\infty} h(u, v)\,dv \right] du$$

$$G(y) = H(+\infty, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y} h(u, v)\,du\,dv = \int_{-\infty}^{y} \left[\int_{-\infty}^{+\infty} h(u, v)\,du \right] dv.$$

After differentiation the probability density functions of variables X and Y are obtained

$$f(x) = F'(x) = \int_{-\infty}^{+\infty} h(x, v)\,dv \tag{2.6}$$

$$g(y) = G'(y) = \int_{-\infty}^{+\infty} h(u, y)\,du. \tag{2.6'}$$

2.1.3. CONDITIONAL CUMULATIVE DISTRIBUTION AND DENSITY FUNCTIONS

Conditional probability for events was already defined. Similarly, the distribution function of a random variable given that another random variable has a particular value can also determined:

It is easy to see that

$$P(X < x|Y < y) = \frac{P(X < x, Y < y)}{P(Y < y)} = \frac{H(x, y)}{G(y)}$$

$$P(Y < y|x_1 \leqq X < x_2) = \frac{P(x_1 \leqq X < x_2, Y < y)}{P(x_1 \leqq X < x_2)} = \frac{H(x_2, y) - H(x_1, y)}{F(x_2) - F(x_1)}$$

$$P(Y < y|X \geqq x) = \frac{P(X \geqq x_1 Y < y)}{P(X \geqq x)} = \frac{G(y) - H(x, y)}{1 - F(x)}.$$

It can be seen that in the latter formula event $\{X \geqq x, Y < y\}$ is the difference of events $\{Y < y\}$ and $\{X < x, Y < y\}$ and the second event is a part of the first. Therefore the difference of the respective probabilities could be calculated.

It has been always assumed that the denominator differs from zero.

The probability of event $\{X = x\}$ is zero for continuous variables, yet there is a need to determine the distribution of Y, given this condition. Under certain conditions one can obtain this from the second formula. Let x_1 and x_2 be substituted by x and $x + \Delta x$, respectively, and let both the numerator and the denominator be divided by Δx:

$$P(Y < y|x \leqq X < x + \Delta x) = \frac{\frac{H(x + \Delta x, y) - H(x, y)}{\Delta x}}{\frac{F(x + \Delta x) - F(x)}{\Delta x}}.$$

If $\Delta x \to 0$ then the expression in the denominator approaches the limit $f(x) \neq 0$ and the conditional distribution is obtained as

$$G(y|x) = P(Y < y|X = x) = \frac{\frac{\partial H(x, y)}{\partial x}}{f(x)} = \frac{\int_{-\infty}^{y} h(x, v)\, dv}{\int_{-\infty}^{+\infty} h(x, v)\, dv}. \tag{2.7}$$

This indeed is a cumulative distribution function as $G(y|x)$ is a monotonous non-decreasing function with zero at $-\infty$ and one at $+\infty$. The conditional density function is obtained by differentiation:

$$g(y|x) = \frac{\partial G(y|x)}{\partial y} = \frac{h(x, y)}{f(x)}. \tag{2.8}$$

Similarly, the cumulative distribution and density function of X given $Y=y$ is

$$F(x|y) = \frac{\int_{-\infty}^{x} h(u, y)\, du}{\int_{-\infty}^{+\infty} h(u, y)\, du} \tag{2.7'}$$

and

$$f(x|y) = \frac{h(x, y)}{\int_{-\infty}^{+\infty} h(u, y)\, du} = \frac{h(x, y)}{g(y)}. \tag{2.8'}$$

The denominator in both cases is the density function $g(y)$ of variable Y.

This was the case of continuous variables. The situation is somewhat simpler if both variables are discrete. There is no unique system for notation in this case and it is usually selected as a function of the particular problem.

Let X and Y be random variables having discrete distributions with possible values of non-negative integers. A usual notation for such probabilities is:

$$P(X = k,\ Y = l) = r_{kl}, \quad k, l = 0, 1, 2, \ldots$$

$$\sum_k \sum_l r_{kl} \equiv 1.$$

Probabilities used in practice can be determined from these quantities. In the following the term "cumulative" will, in general, be omitted when dealing with distribution functions. The cumulative distribution and marginal distribution functions are as follows:

$$P(X = k) = p_k = \sum_l r_{kl} \quad (\sum p_k = 1) \tag{2.9}$$

$$P(Y = l) = q_l = \sum_k r_{kl} \quad (\sum q_l = 1) \tag{2.10}$$

$$H(k, l) = P(X < k,\ Y < l) = \sum_{i=0}^{k-1} \sum_{j=0}^{l-1} r_{ij} \quad (k, l = 1, 2, \ldots). \tag{2.11}$$

Some conditional probabilities:

$$P(Y = l|X = k) = \frac{r_{kl}}{p_k}; \quad P(X = k|Y = l) = \frac{r_{kl}}{q_l}$$

$$P(X < k|l_1 \leqq Y < l_2) = \frac{\sum_{i=0}^{k-1} \sum_{j=0}^{l-1} r_{ij}}{\sum_{j=l_1}^{l_2} q_j}.$$

In mathematical statistics the analysis of more than 2 variables is often required. If $X_1, X_2, \ldots, X_n$ are the results of n observations with respect to some random

quantities then these variables may be considered as elements of an n-dimensional random vector

$$\vec{X} = (X_1, X_2, \ldots, X_n).$$

Vector $\vec{X}$ is a random point in an n-dimensional space. The distribution of vector $\vec{X}$ is the joint distribution of variables $X_1, X_2, \ldots, X_n$. Consider now the case when the random variables are continuously distributed. The joint distribution of random variables $X_1, X_2, \ldots, X_n$ is an n-dimensional distribution function:

$$H(x_1, x_2, \ldots, x_n) = P(X_1 < x_1,\ X_2 < x_2, \ldots, X_n < x_n).$$

The n-dimensional density function, if it exists, is then

$$h(x_1, x_2, \ldots, x_n) = \frac{\partial^n H(x_1, x_2, \ldots, x_n)}{\partial x_1, \partial x_2, \ldots, \partial x_n}.$$

The joint distribution function of n random variables, therefore, gives the probability of the joint simultaneous occurrence of n events i.e. $\{X_i < x_i\}$ $(i = 1, 2, \ldots, n)$.

The n-dimensional distribution function $H(x_1, x_2, \ldots, x_n)$ is a monotonous non-decreasing function and continuous from the left for all n. If $-\infty$ has been substituted for any of the variables, then $H(x_1, x_2, \ldots, x_n) = 0$ (as putting $-\infty$ for X_i would imply event $\{X_i < -\infty\}$ which is impossible). On the other hand, $H(+\infty, +\infty, \ldots, +\infty) = 1$ which is the case if all variables are equal to $+\infty$.

The probability that an n-dimensional random vector $\vec{X}$ is in a prescribed rectangle of the n-dimensional space, i.e.

$$P(a_1 \leqq X_1 < b_1,\ a_2 \leqq X_2 < b_2, \ldots, a_n \leqq X_n < b_n)$$

can be calculated by the distribution function $H(x_1, x_2, \ldots, x_n)$. This must certainly mean that these probabilities expressed by the values of H should be non-negative, see Figure 21.

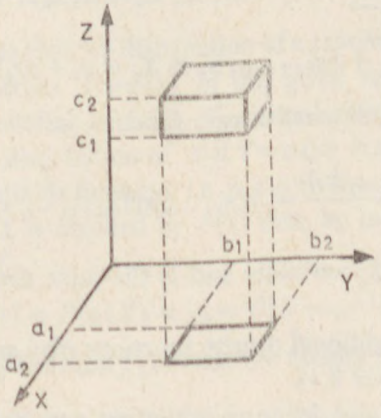


Figure 21

If there exists any n-dimensional function $h(x_1, x_2, \ldots, x_n) \geqq 0$ for any domain E of the n-dimensional space, for which

$$P(\vec{X} \in E) = \iint_E \ldots \int h(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n$$

then $\vec{X}$ is distributed continuously and relations (2.2) and (2.3) are valid for functions $H(x_1, x_2, \ldots, x_n)$ and $h(x_1, x_2, \ldots, x_n)$. Equivalently:

$$\int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_n} h(t_1, t_2, \ldots, t_n)\, dt_1 dt_2 \ldots dt_n = H(x_1, x_2, \ldots, x_n), \tag{2.11}$$

and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n = 1. \tag{2.12}$$

Equations (2.7) and (2.8) can be easily generalized for multi-dimensional distributions. The only thing to do is to replace X and Y by random vectors and obviously the functions $f(x)$, $g(y)$, $f(x|y)$, $g(y|x)$ should have the same number of arguments like the dimensions of X and Y.

For instance, if the joint distribution of $X_1, X_2, \ldots X_n$ is continuous then the joint conditional distribution function of $X_1, X_2, \ldots, X_n$ given conditions $X_{k+1} = x_{k+1}, \ldots, X_n = x_n$ is:

$$H(x_1, x_2, \ldots, x_k | x_{k+1}, \ldots, x_n) =$$

$$= P(X_1 < x_1, \ldots, X_k < x_k | X_{k+1} = x_{k+1}, \ldots, X_n = x_n)$$

and may be defined by the following limit:

$$H(x_1, x_2, \ldots, x_k | x_{k+1}, \ldots, x_n) =$$

$$= \lim_{\Delta x_{k+1} \to 0} \ldots \lim_{\Delta x_n \to 0} P(X_1 < x_1, \ldots, X_k < x_k | x_{k+1} \leqq X_{k+1} < x_{k+1} +$$

$$+ \Delta x_{k+1}, \ldots, x_n \leqq X_n < x_n + \Delta x_n).$$

The corresponding conditional density function is:

$$\frac{\partial H(x_1, \ldots, x_k | x_{k+1}, \ldots, x_n)}{\partial x_1, \ldots, \partial x_k} = h(x_1, \ldots, x_k | x_{k+1}, \ldots, x_n) = \frac{h(x_1, \ldots, x_n)}{g(x_{k+1}, \ldots, x_n)}$$

where function has $(n-k)$ variables and is the joint density function of variables $X_{k+1}, X_{k+2}, \ldots, X_n$.

Multi-dimensional conditional density functions play an important role in regression analysis (see: Chapter 8.1).

Let us now investigate and define an important new notion by knowing the joint distribution function of more variables: the independence of random variables.

Random variables $X_1, X_2, \ldots, X_n$ are said to be completely independent if for any

$$a_i < b_i \quad (i = 1, 2, \ldots, n)$$

$$P(a_1 \leqq X_1 < b_1,\ a_2 \leqq X_2 < b_2, \ldots, a_n \leqq X_n < b_n) =$$

$$= P(a_1 \leqq X_1 < b_1) P(a_2 \leqq X_2 < b_2) \ldots P(a_n \leqq X_n < b_n).$$

Specifically, if $a_i = -\infty$ and $b_i = x$ $(i = 1, 2, \ldots, n)$

$$P(X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n) = P(X_1 < x_1) P(X_2 < x_2) \ldots P(X_n < x_n).$$

That is

$$H(x_1, x_2, \ldots, x_n) = F_1(x_1) F_2(x_2) \ldots F_n(x_n). \tag{2.13}$$

In other words: n random variables are completely independent if their joint distribution function is equal to the product of the distribution functions of the individual variables. If the variables are continuous then their joint distribution function is also continuous. It also follows from the above defined independence that the joint density function must be the product of the particular density functions:

$$h(x_1, x_2, \ldots, x_n) = f_1(x_1) f(x_2) \ldots f_n(x_n). \tag{2.14}$$

The notion of independence is of special importance since the well-established methods of mathematical statistics are almost exclusively based on independent random variables. The further development of statistical methods for non-independent random variables is one of the most urgent tasks of mathematical statistics.

Intensive research is being conducted nowadays in this direction.

To decide whether random variables $X_1, X_2, \ldots, X_n$ are independent or not is a task of mathematical statistics. Appropriate methods will be discussed in Chapter 6.

2.1.4. THE DISTRIBUTION OF THE MONOTONOUS FUNCTION OF A RANDOM VARIABLE

In practice it often happens that the distribution of a random variable X is known and the distribution of a function $Y = \varphi(x)$ of this given variable is sought. Variable Y is, of course, another random variable. The distribution of Y can be easily determined on the basis of the distribution of X if $Y = \varphi(x)$ is a monotonously increasing or decreasing and differentiable function, i.e. it can be inverted. If this is true, and the distribution function of X is denoted by $F(x)$ then by using notation $G(y)$ for the distribution function of Y:

$$G(y) = P(Y < y) = P[\varphi(X) < y] = P[X < \varphi^{-1}(y)] = F[\varphi^{-1}(y)].$$

If variable X has a density function $f(x)$, then the density function of variable Y is:

$$g(y) = \frac{dG(y)}{dy} = \frac{dF[\varphi^{-1}(y)]}{dy} = f[\varphi^{-1}(y)] \frac{d\varphi^{-1}(y)}{dy}. \tag{2.15}$$

If, for example, $Y=aX+b$, or $\varphi(x)=ax+b$ and the density function of X is $f(x)$, then

$$x = \varphi^{-1}(y) = \frac{y-b}{a},$$

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right). \tag{2.16}$$

If

$$Y = e^X$$

and

$$y = \varphi(x) = e^x; \quad x = \varphi^{-1}(y) = \ln y$$

then

$$g(y) = \frac{1}{|y|} f(\ln y). \tag{2.17}$$

2.1.5. THE DISTRIBUTION OF THE SUM OF TWO RANDOM VARIABLES

Let first discuss the case of continuously distributed variables. Let the joint density function of continuous random variables X and Y be $h(x, y)$. In order to find the distribution function of $Z=X+Y$, the probability of event

$$\{Z < z\} = \{X+Y < z\}$$

must be determined for all z. Consequently, the density function $h(x, y)$ is to be integrated over a region T_z^+ on the plane, for which the condition $x+y<z$ holds. If the distribution function of z is denoted by $K(z)$, then

$$K(z) = P(Z < z) = \iint_{T_z^+} h(x, y)\, dx\, dy.$$

This integration can be performed in the following way: first x is fixed and we integrate according to y in interval $(-\infty, z-x)$, then according to x in interval $(-\infty, +\infty)$. In other words, the double integral may be converted into two successive integrations:

$$K(z) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-x} h(x, y)\, dy \right) dx. \tag{2.18}$$

If X and Y are independent random variables, i.e.

$$h(x, y) = f(x) \cdot g(y)$$

then

$$K(z) = \int_{-\infty}^{\infty} f(x) \left(\int_{-\infty}^{z-x} g(y)\, dy \right) dx. \tag{2.19}$$

The internal integral is nothing else than the value of the distribution function of X at $z-x$, and therefore

$$K(z) = \int_{-\infty}^{\infty} G(z-x)f(x)\,dx. \tag{2.20}$$

The density function is then obtained by derivation according to z:

$$K'(z) = k(z) = \int_{-\infty}^{\infty} g(z-x)f(x)\,dx. \tag{2.21}$$

In case of identically distributed, independent variables, in other words, if $f(x)=g(x)$, the density function of $X+Y$ is

$$k(z) = \int_{-\infty}^{\infty} f(z-x)f(x)\,dx. \tag{2.22}$$

The distribution of sums in a discrete case will be discussed only for independent variables:

Let be the possible values of X, $x_1, x_2, \ldots$, and of Y (independent of X) $y_1, y_2, \ldots$. Let denote the probability of their values by:

$$P(X = x_i) = p_i \quad (i = 1, 2, \ldots)$$

$$P(Y = y_j) = q_j \quad (j = 1, 2, \ldots).$$

Then for the random variable $Z=X+Y$ it is true that

$$P(Z < z) = \sum_{x_i+y_j=z} p_i q_j \tag{2.23}$$

which means that summation should be extended for every pair of the (i, j) indices for which $x_i+y_j=z$. If z cannot be calculated like this above described sum then its probability is zero.

Let allow for X and Y to be non-negative integers. Then $Z=X+Y$ will be also a non-negative integer and the probabilities are the following:

$$P(Z = k) = \sum_{i=0}^{k} p_i q_{k-i} \tag{2.23'}$$

where $p_i=P(X=i)$, $q_j=P(Y=j)$.

2.1.6. THE DISTRIBUTION OF THE PRODUCT AND QUOTIENT OF TWO INDEPENDENT RANDOM VARIABLES

Let the density function of X be $f(x)$ and of Y (independent of X) $g(y)$. Let denote by T_z^+ the region on the plane (x, y) for which $xy<z$ and by T_z^* for which $y<z$. The following relationships are obvious:

$$R(z) = P(XY < z) = \iint\limits_{T_z^+} f(x)g(y)\,dx\,dy$$

$$S(z) = P\left(\frac{X}{Y} < z\right) = \iint\limits_{T_z^*} f(x)g(y)\,dx\,dy.$$

If Y may take only non-negative values: $Y \geqq 0$, and $g(y)=0$ if $y \leqq 0$, then the following expressions may be obtained for the distribution and the density functions:

$$R(z) = \int_0^\infty g(y)\left[\int_{-\infty}^{z/y} f(x)\,dx\right]dy = \int_0^\infty F\left(\frac{z}{y}\right)g(y)\,dy \tag{2.24}$$

$$S(z) = \int_0^\infty g(y)\left[\int_{-\infty}^{xy} f(x)\,dx\right]dy = \int_0^\infty F(zy)g(y)\,dy \tag{2.25}$$

$$r(z) = R'(z) = \int_0^\infty f\left(\frac{z}{y}\right)g(y)\cdot\frac{1}{y}\,dy \tag{2.26}$$

$$s(z) = S'(z) = \int_0^\infty yf(zy)g(y)\,dy. \tag{2.27}$$

If Y may take negative values, then the density functions will take the form:

$$r(z) = \int_{-\infty}^\infty f\left(\frac{z}{y}\right)g(y)\frac{1}{|y|}\,dy \tag{2.28}$$

$$s(z) = \int_{-\infty}^\infty |y|f(zy)g(y)\,dy. \tag{2.29}$$

2.1.7. THE PARAMETERS OF DISTRIBUTION FUNCTIONS

The distribution of a random variable is described by help of the distribution function or (if available) of the density function. In practice, it is often sufficient (and necessary) that an overall picture be at hand about the distribution by aid of some parameters. The analogy between the distribution of a random variable and of mass has been earlier mentioned. In case of the distribution of mass the main question is, where can be found the center of gravity, or the centrum and what is the density of mass around this point? The measure of this latter is called moment of inertia. Similarly, it is of prime importance to know also in the distribution of random variables where the center of gravity of the distribution might be which we call "expectation".

It is also interesting how close these random numbers will scatter around this value — or centrum — and which interval (short or longer) will contain a given percentage (say 70—90 percent) of the variables. In probability theory the moment of inertia is replaced by variance, by a squared average deviation of the variable from the expected value.

Beyond expectation, and variance also other numerical characteristics will be discussed. Limits will be set up containing 25, 50, 75 percent of the values of the random variables, and a number that is characteristics to the symmetry of the distribution. Parameters representing stochastic relationships among more random variables will be also discussed.

a) *The expectation and its characteristic features*

From now on the term "mean" is used in this book mostly instead of "expectation" or "expected value" which are synonymous expressions.

The calculation of the mean or the centrum of a distribution is analogous to the calculation of mass central gravity. Let X be a discrete random variable with possible values of $x_1, x_2, \ldots, x_n, \ldots$ and let

$$P(X = x_k) = p_k \quad (k = 1, 2, \ldots).$$

In this case the mean of random variable X, denoted by $E(X)$, is obtained as

$$E(X) = \sum_k x_k p_k, \tag{2.30}$$

if $\sum_k |x_k| p_k < \infty$. Because $\sum p_k = 1$, it is evident that the mean of X is a weighted arithmetic average of the possible values of X, where the weights are composed of the probabilities attached to each variable.

If X is a continuous variable with a density of $f(x)$ than the mean is obtained by expression

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx \tag{2.31}$$

similarly to the definition of the mass central gravity point from a continuous distribution of mass, supposed that the improprious integral is absolutely convergent, i.e.,

$$\int_{-\infty}^{\infty} |x| f(x)\, dx < +\infty.$$

Some important properties of the mean are expressed by the following rules.

Theorem 1: The mean of a constant is the constant itself:

$$E(c) = c.$$

Proof: Constant c may be considered a random variable that will take its only value c with a probability of 1. So, according to the definition of the mean:

$$E(c) = c \cdot 1 = c.$$

Theorem 2: If X is a bounded random variable, i.e.

$$a \leqq X \leqq b$$

then its mean exists, and

$$a \leqq E(X) \leqq b.$$

Proof: If X is a discrete random variable with possible representations of $x_1, x_2, \ldots, x_n, \ldots$ and with attached probabilities of $p_1, p_2, \ldots, p_n, \ldots$ then because every $x_1, x_2, \ldots, x_n, \ldots$ is in between the limits a and b:

$$a = ap_1 + ap_2 + \ldots \leqq x_1p_1 + x_2p_2 + \ldots \leqq b = bp_1 + bp_2 + \ldots$$

and

$$a = a \sum p_i \leqq \sum x_i p_i \leqq b \sum p_i = b.$$

If X is a continuous random variable, then

$$a = a \int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} af(x)\,dx \leqq \int_{-\infty}^{\infty} xf(x)\,dx \leqq$$

$$\leqq \int_{-\infty}^{\infty} bf(x)\,dx = b \int_{-\infty}^{\infty} f(x)\,dx = b.$$

It should be mentioned that in case of $a \leqq X \leqq b$

$$\int_a^b f(x)\,dx = 1.$$

Theorem 3: If $Y = aX + b$, i.e. random variable Y is a linear function of random variable X, then

(2.32) $$E(Y) = E(aX+b) = aE(X) + b.$$

Proof: If X is a discrete random variable with possible values of $x_1, x_2, \ldots$ and respective probabilities of $p_1, p_2, \ldots$ then the possible values of Y will be $ax_1 + b$, $ax_2 + b + \ldots$ with probabilities $p_1, p_2, \ldots$, so

$$E(Y) = \sum_k (ax_k + b)p_k = a \sum_k x_k p_k + b \sum_k p_k = aE(X) + b$$

supposed of course that the series of $\sum_k |x_k| p_k$ is converging, say $|X|$ must have a mean.

If X is continuously distributed with a density function of $f(x)$, then if $X = x$ so $Y = ax + b$ and the density function of Y on the basis of formula (2.16) is:

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right)$$

and

$$E(Y) = \int_{-\infty}^{\infty} yg(y)\,dy = \frac{1}{|a|}\int_{-\infty}^{\infty} yf\left(\frac{y-b}{a}\right)dy$$

and by substitution of $\frac{y-b}{a} = u$.

$$E(Y) = \frac{a}{|a|}\int_{-\infty}^{\infty}(au+b)f(u)\,du = a\int_{-\infty}^{\infty} uf(u)\,du + b\int_{-\infty}^{\infty} f(u)\,du = aE(X)+b.$$

It should be mentioned as a special case that if $Y=X-E(X)$, then

$$E(Y) = E(X)-EE(X) = E(X)-E(X) = 0.$$

Definition: Let random variable Y be a continuous function of random variable X: $Y=\varphi(X)$. The mean of Y is then defined as

(2.33) $$E(Y) = \sum_k \varphi(x_k)p_k$$

in case of a discrete distribution, and

(2.33′) $$E(Y) = \int_{-\infty}^{\infty}\varphi(x)f(x)\,dx$$

in case of a continuous distribution supposed that the sum, or the integral is absolutely convergent.

The mean of functions of random—vector variables is defined in an analogous way. E.g. if continuous random variables are at hand, and the joint density function of random variables $X_1, X_2, \ldots, X_n$ is $h(x_1, x_2, \ldots, x_n)$ and $Y=\varphi(X_1, X_2, \ldots, X_n)$, then

(2.34) $$E(Y) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\varphi(x_1, \ldots, x_n)h(x_1, \ldots, x_n)dx_1, \ldots, dx_n$$

if the improprious integral is absolutely convergent.

Theorem 4: Let X and Y be random variables having any kind of distribution with existing means, then

(2.35) $$E(X+Y) = E(X)+E(Y).$$

Proof is first presented on discrete variables. Let the possible values of X numbers $x_1, x_2, \ldots, x_n, \ldots$ and of Y numbers $y_1, y_2, \ldots, y_n, \ldots$. And let define

$$P(X = x_i) = p_i; \quad P(Y = y_j) = q_j;$$

$$P(X = x_i, Y = y_j) = r_{ij} \quad (i, j = 1, 2, \ldots)$$

where the probabilities r_{ij} stand for a joint distribution of the pair of variables (X, Y). On the basis of the definition of the mean:

$$E(X+Y) = \sum_i \sum_j (x_i+y_j)r_{ij} = \sum_i \sum_j x_i r_{ij} + \sum_i \sum_j y_j r_{ij} =$$

$$= \sum_i x_i \left(\sum_j r_{ij}\right) + \sum_j y_j \left(\sum_i r_{ij}\right).$$

According to (2.9) and (2.10)

$$\sum_j r_{ij} = p_i, \quad \sum_i r_{ij} = q_j.$$

Consequently,

$$E(X+Y) = \sum_i x_i p_i + \sum_j y_j q_j = E(X)+E(Y).$$

In case of continuous variables if the joint density function of X and Y is $h(x, y)$, then according to (2.34) and by substituting $\varphi(x, y)=x+y$

$$E(X+Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y)h(x, y)\,dx\,dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xh(x, y)\,dx\,dy +$$

$$+ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yh(x, y)\,dx\,dy = \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} h(x, y)\,dy\right)dx + \int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} h(x, y)\,dx\right)dy =$$

$$= \int_{-\infty}^{\infty} xf(x)\,dx + \int_{-\infty}^{\infty} yg(y)\,dy = E(X)+E(Y).$$

The theorem dealing with the mean of the sum of random variables may be generalized for any finite number of addends by induction:

(2.35′) $$E(X_1+X_2+\ldots+X_n) = E(X_1)+E(X_2)+\ldots+E(X_n).$$

If, namely, $X=X_1+X_2+\ldots+X_{n-1}$; $Y=X_n$ and suppose that the theorem is valid for $n-1$ addends, or in other words, if

$$E(X) = E(X_1+\ldots+X_{n-1}) = E(X_1)+E(X_2)+\ldots+E(X_{n-1})$$

then it follows from the proved theorem that

$$E(X_1+\ldots+X_{n-1}+X_n) = E(X)+E(X_n) = E(X)+E(Y) =$$

$$= \sum_{i+1}^{n-1} E(X_i)+E(X_n) = \sum_{i=1}^{n} E(X_i).$$

Theorem 5. The mean of the product of independent random variables is equal to the product of their respective means. So, if X and Y are random variables, then

(2.36) $$E(XY) = E(X)\cdot E(Y).$$

Proof is given first for the continuous case. Because of their independence, the joint density function of random variable X and Y is equal to the product of the den-

sity functions of each separate variable:

$$h(x, y) = f(x)\, g(y).$$

Based on formula (2.34) and with a substitution of $\varphi(x, y) = x \cdot y$:

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x)\, g(y)\, dx\, dy = \int_{-\infty}^{\infty} xf(x) \Big(\int_{-\infty}^{\infty} yg(y)\, dy \Big) dx =$$

$$= \int_{-\infty}^{\infty} xf(x)\, dx \cdot \int_{-\infty}^{\infty} yg(y)\, dy = E(X)E(Y).$$

If X and Y are discrete variables and $P(X=x_i)=p_i$, $P(Y=y_j)=q_j$ and their joint distribution is:

$$P(X = x_i,\ Y = y_j) = r_{ij}$$

then because X and Y are independent

$$r_{ij} = p_i q_j \quad (i, j = 1, 2, \ldots).$$

Consequently,

$$E(XY) = \sum_i \sum_j x_i y_j r_{ij} = \sum_i \sum_j x_i y_j p_i q_j =$$

$$= \sum_i x_i p_i \Big(\sum_j y_j q_j \Big) = \sum_i x_i p_i \sum_j y_j q_j = E(X)E(Y).$$

This last theorem is easily generalized for the calculation of the mean of products of independent random variables in any desired finite number. Proof to this may be obtained by the use of induction.

It is true, that in case of independence

(2.36′) $$E(X_1 X_2 \ldots X_n) = E(X_1)E(X_2) \ldots E(X_n)$$

assuming the existence of the means of the individual variables.

b) *Conditional expectation*

Let have X and Y discrete random variables with the following distribution:

$$X: \begin{pmatrix} x_1, x_2, \ldots, x_n, \ldots \\ p_1, p_2, \ldots, p_n, \ldots \end{pmatrix}$$

$$Y: \begin{pmatrix} y_1, y_2, \ldots, y_n, \ldots \\ q_1, q_2, \ldots, q_n, \ldots \end{pmatrix}.$$

Let be

$$P(X = x_i,\ Y = y_j) = r_{ij} \quad (i, j = 1, 2, \ldots).$$

One may define the conditional mean of random variable Y by the condition that X has taken a value x_i:

(2.37) $$E(Y|X = x_i) = \sum_j y_j P(Y = y_j | X = x_i).$$

By aid of this equation the conditional mean of Y may be attained for every actual value of variable X as a condition. The value of this conditional mean depends on a value x_i of X, which is a random event. $E(Y|X)$ is therefore a random variable with possible values of

$$E(Y|X=x_1), E(Y|X=x_2), \ldots, E(Y|X=x_n), \ldots.$$

It is easy to see that $E(Y|X)$ takes a value $E(Y|X=x_i)$ exactly with a probability of p_i.

Let us calculate the expectation of random variable $E(Y|X)$:

$$E[E(Y|X)] = \sum_i E(Y|X=x_i)p_i.$$

If relationship (2.37) is considered the following interesting result may be obtained:

$$E[E(Y|X)] = \sum_i \left(\sum_j y_j P(Y=y_j|X=x_i)p_i\right) = \tag{2.38}$$

$$= \sum_i \sum_j y_j r_{ij} = \sum_j y_j \sum_i r_{ij} = \sum_j y_j q_j = E(Y).$$

Here, the following relationship was also used:

$$P(Y=y_j|X=x_i)P(X=x_i) = P(Y=y_j,\ X=x_i) = r_{ij}.$$

Eq. (2.38) indicates that the mean of the conditional mean of Y related to X is equal to the unconditional mean of variable Y.

If X and Y are continuous random variables with density functions of $f(x)$ and $g(y)$ and a joint density function of $h(x, y)$ then the conditional mean of random variable Y related to X is defined by

$$E(Y|X=x) = \int_{-\infty}^{\infty} y\frac{h(x,y)}{f(x)}dy = \int_{-\infty}^{\infty} yg(y|x)\,dy.$$

The following relationship is again valid:

$$E[E(Y|X)] = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} y\frac{h(x,y)}{f(x)}dy\right)f(x)dx = \tag{2.38'}$$

$$= \int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} h(x,y)\,dx\right]dy = \int_{-\infty}^{\infty} yg(y)\,dy = E(Y).$$

(With the assumption — of course — that the above defined improper integral is convergent).

Function $\bar{y}(x)=E(Y|X=x)$ which depends on x is called the regression curve of variable Y related to X.

In an analogous way the conditional mean of random variable X related to Y may be defined.

A conditional mean defined as $E(Y|X_1X_2\ldots, X_n)$ may be also interpreted. This will be a function of the variables $X_1, X_2, \ldots, X_n$, and is called a function with n

variables. This function has an interesting minimum-property: if $g(X_1, X_2, ..., X_n)$ is a function of variables $X_1, X_2, ..., X_n$, then

$$E\{[Y-g(X_1, X_2, ..., X_n)]^2\} \geqq E\{[Y-E(YX_1, X_2, ..., X_n)]^2\}. \tag{2.39}$$

c) *The variance and its characteristics*

As it was mentioned earlier the mean is only one of the numerical characteritics of a distribution. It is for the definition of a central point around which the values of the random variable will fluctuate. About the *measure* of this fluctuation, however, it does not say anything. The scattering of the values of the random variable around this central point is usually described by

$$D(X) = +\sqrt{E[X-E(X)]^2},$$

and is called standard deviation.

Below the square root the squared average deviations of our random variable and of its mean are found. We may talk about standard deviation only in the case, if the mean $E(X)$ and another mean $E[X-E(X)]^2$ are existing.

The variance of random variable X is defined as:

$$D^2(X) = E[X-E(X)]^2. \tag{2.40}$$

Calculation of variance may be made easier by the use of the following expression:

$$D^2(X) = E[X^2-2XE(X)+E^2(X)] = \tag{2.41}$$

$$= E(X)^2-2E(X)E(X)+E^2(X) = E(X^2)-E^2(X).$$

Because $D^2(X) \geqq 0$, it is evident that

$$E(X^2) \geqq E^2(X).$$

The calculation of variance is the following in case of discrete variables:

$$D^2(X) = \sum_i [x_i-E(X)]^2 p_i = \sum_i x_i^2 p_i - E^2(X).$$

In case of continuous variables:

$$D^2(X) = \int_{-\infty}^{\infty} [x-E(X)]^2 f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - E^2(X).$$

By aid of the notion of variance another important characteristic of the mean denoted by $E(X)=m$ may be expressed.

If a is a real number, then

$$E[(X-a)^2] \geqq E[(X-m)^2]. \tag{2.42}$$

Namely,

$$E[(X-a)^2] = E[(X-m+m-a)^2] = E[(X-m)^2]+2(m-a)E(X-m)+(m-a)^2.$$

Because $E(X-m)=E(X)-m=0$, therefore

$$E[(X-a)^2] = E[(X-m)^2]+(m-a)^2 = D^2(X)+(m-a)^2 \geqq D^2(X).$$

The mean $E(X)=m$ is a point on the line from which the squared average deviation of our random variable is less than its squared average deviation from any other point.

Relationship (2.42) is called the Steiner-formula.

Theorems pertaining to the notion of variance.

Theorem 1: If random variable Y is a linear function of random variable X: $Y=aX+b$, then

$$D^2(Y) = a^2 D^2(X). \tag{2.43}$$

Proof: From the definition of variance

$$D^2(Y) = E\{[(aX+b)-E(aX+b)]^2\} = E\{[aX-E(aX)^2]\} =$$
$$= a^2 E\{[X-E(X)]^2\} = a^2 D^2(X).$$

Theorem 2: If X and Y are independent random variables and if $Z=X+Y$, then

$$D^2(Z) = D^2(X)+D^2(Y). \tag{2.44}$$

Proof: If X and Y are independent, then also $X-E(X)$ and $Y-E(Y)$ must be independent. With consideration to the fact that the mean of the product of independent variables is equal to the standard deviation of their mean values, and due to $E[X-E(X)]=0$ and $E[Y-E(Y)]=0$, on the basis of the definition of variance we may find that:

$$D^2(Z) = D^2(X+Y) = E\{[(X+Y)-E(X+Y)]^2\} = E\{[X-E(X)+Y-E(Y)]^2\} =$$
$$= E\{[X-E(X)]^2\}+2E[X-E(X)]\cdot E[Y-E(Y)]+E\{[Y-E(Y)]^2\} =$$
$$= D^2(X)+D^2(Y).$$

Theorem 2 may be extended to a finite number of independent random variables by induction:

If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$D^2(X_1+X_2+\ldots+X_n) = D^2(X_1)+D^2(X_2)+\ldots+D^2(X_n). \tag{2.44'}$$

By knowing the mean and standard deviation of a random variable the fluctuation of the latter around the mean may be well characterized. This is expressed by the so-called Chebyshev-inequality.

Theorem 3: The Chebyshev-inequality. If the mean and standard deviation of a random variable X exist, then

$$P[|X-E(X)| \geqq \lambda D(X)] \leqq \frac{1}{\lambda^2}. \tag{2.45}$$

Chebishev's inequality has the following verbal meaning: the probability that the value of a random variable will deviate from its mean in absolute terms more than $2D(X)$ is less than $\frac{1}{4}$, or the probability that the observed value of X will fall outside the interval $[E(X)-3D(X),\ E(X)+3D(X)]$ is less than $\frac{1}{9}$, etc., see Figure 22.

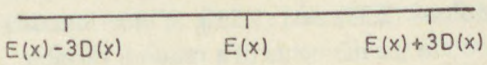


Figure 22

If this interval is depicted on the line of numbers, and a large number of observations has been performed on X, then, about 90 percent of the observed values will be found in this interval supposed that the mean and the standard deviation of X exist. If in Eq. (2.45) the substitution $\lambda D(X)=\varepsilon$ is introduced then $\frac{1}{\lambda^2}=\frac{D^2(X)}{\varepsilon^2}$ and Eq. (2.45) will be equivalent with the following statement:

$$P(|X-E(X)|>\varepsilon)=\frac{D^2(X)}{\varepsilon^2}. \tag{2.46}$$

Proof: Introduce the symbols $E(X)=m$ and $D(X)=\sigma$. If X is a discrete random variable with possible values of $x_1, x_2, \ldots$ and with parallel probabilities of $p_1, p_2, \ldots$, then

$$\sigma^2=\sum_i (x_i-m)^2 p_i \geqq \sum_{i:\,(x_i-m)>\varepsilon} (x_i-m)^2 p_i \geqq \varepsilon^2 \sum_{i:\,(x_i-m)>\varepsilon} p_i = \varepsilon^2 P(|X-m|>\varepsilon).$$

From this:

$$P(|X-m|>\varepsilon) \leqq \frac{\sigma^2}{\varepsilon^2}.$$

If X is continuously distributed with a density function of $f(x)$, then

$$\sigma^2=\int_{-\infty}^{\infty}(x-m)^2 f(x)\,dx \geqq \int_{-\infty}^{m-\varepsilon}(x-m)^2 f(x)\,dx+$$

$$+\int_{m+\varepsilon}^{\infty}(x-m)^2 f(x)\,dx \geqq \varepsilon^2\left(\int_{-\infty}^{m-\varepsilon} f(x)\,dx+\int_{m+\varepsilon}^{\infty} f(x)\,dx\right)=\varepsilon^2 P(|X-m|>\varepsilon).$$

Remark: Chebyshev's inequality is of general validity. It is a measure of the fluctuation of any random variable which has a mean and a standard deviation. Due to its general validity, however, no accurate calculation may be expected by its use for the probability of a given deviation. We will see that for example in case of the normal distribution, a very important distribution in practice, the values of the random variable are much closer around the mean than it might be expected from Chebyshev's

theorem. The probability of deviation $|X-m|>2\sigma$ in this case is less than 0.05. The theorem's importance, however, is in this informative role. It will be working even in cases if nothing else but the mean and standard deviation of a distribution are known.

d) *The moments of a random variable*

In the previous section the mean of some function $\varphi(X)$ of a random variable X has been defined by expressions (2.33) and (2.33'). If this function $\varphi(X)$ is selected in a special way we may obtain the moments of a random variable.

Let be $\varphi(X)=X^k$, then

$$\alpha_k = E(X^k). \tag{2.47}$$

This expectation* is called the kth moment of random variable X. The mean is the first moment of X.

If $\varphi(X)=[X-E(X)]^k$, then

$$\mu_k = E\{[X-E(X)]^k\} \tag{2.48}$$

is called the kth central moment of random variable X. Variance is nothing else than the second central moment of the random variable:

$$D^2(X) = \mu_2 = E\{[X-E(X)]^2\} \neq E(X^2)-E^2(X) = \alpha_2-\alpha_1^2.$$

It is obvious that higher order central moments can also be expressed by non-central moments. E.g.

$$\mu_3 = E[(X-\alpha_1)^3] = E(X^3)-3\alpha_1 E(X^2)+3\alpha_1^2 E(X)-\alpha_1^3 = \tag{2.49}$$

$$= \alpha_3-3\alpha_1\alpha_2+2\alpha_1^3.$$

$$\mu_4 = E(X-\alpha_1)^4 = \alpha_4-4\alpha_1\alpha_3+6\alpha_1^2\alpha_2-3\alpha_1^4. \tag{2.50}$$

If the variable is discrete, calculation of the moments is performed by aid of the following formulae**:

$$\alpha_k = \sum_i x_i^k p_k$$

and

$$\mu_k = \sum_i (x_i-\alpha_i)^k p_i.$$

If the variable is continuous, and the density function of X is denoted by $f(x)$ the following expressions stand for the same purpose:

$$\alpha_k = \int_{-\infty}^{\infty} x^k f(x)\,dx$$

and

$$\mu_k = \int_{-\infty}^{\infty} (x-\alpha_1)^k f(x)\,dx.$$

* If this expectation exists!

** Let assume that the sums and integrals used for this purpose are absolutely convergent.

It should be mentioned that if the distribution of a random variable X is symmetric with respect to its mean, its odd-order central moments are zero. It may happen, however, that the distribution of X with respect to its mean is asymmetric, or skewed. The measure of skewness is:

$$\gamma_1 = \frac{\mu_3}{\sigma^3}. \tag{2.51}$$

(This expression is usually used in case of continuous distributions).

For a measure of the excess of a density function the following equation was introduced:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3. \tag{2.52}$$

In case of a normal distribution $\gamma_2 = 0$. If for a continuous distribution $\gamma_2 > 0$ then its density function will reach higher — it will be peakier — than the density function of a normal distribution (see Section 5.6). If $\gamma_2 < 0$, it will be flatter than the density function of the normal distribution.

Another useful descriptor to characterize continuous distributions is the median denoted by M_e. It is a value which will be exceeded by the random variable with a probability of 1/2. In mathematical form:

$$F(M_e) = \frac{1}{2}.$$

If equation $F(X) = \frac{1}{2}$ has more solutions, so for example, if the distribution function $F(x)$ reaches a value of 1/2 in point x_0 and remains constant up to any point x_1, from which it will increase again, then

$$M_e = \frac{x_0 + x_1}{2}.$$

Solution to equation $F(x) = p$ (let assume there is only one solution) is called the pth quantile of a distribution denoted by q_p. It is obvious that $q_{\frac{1}{2}} = M_e$.

Quantiles $q_{\frac{1}{4}}$ and $q_{\frac{3}{4}}$ are called upper and lower quartiles, see Figure 23.

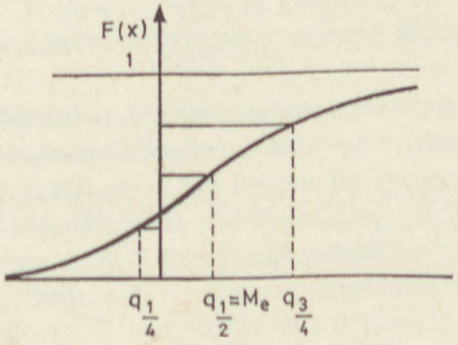


Figure 23

e) *The correlation coefficient*

There are some descriptors used with respect to the joint distributions of two or more random variables. Two of them will be here discussed in detail. If two random variables X and Y are taken simultaneously into consideration in an experiment we are usually faced with the following problem: are they independent of each other or is there any functional relationship between the respective values? It is often the case that a certain tendency may be discovered between X and Y, e.g. if X is large Y is also large or vice versa.

If, for instance, X is a stage value in a given point of a river and Y is groundwater depth in a nearby well then one may observe any kind of stochastic dependence — a tendency — between the two series of measurements even if there is absolutely no functional relationship at hand. Therefore, this stochastic dependence is poured in a numerical form by aid of the covariance and the correlation coefficient.

Let X and Y be random variables. If they are independent of each other then

$$E\{[X-E(X)][Y-E(Y)]\} = E[X-E(X)].E[Y-E(Y)] = 0,$$

based on Eq. (2.36).

If X and Y are not independent of each other then the above presented expectation will have a value C which is, in general, different from zero. The quantity expressed by

$$C = E\{[X-E(X)][Y-E(Y)]\} \tag{2.53}$$

is called the covariance of random variables X and Y. If $Y=X$, then $C=E[X-E(X)]^2$ which is identical to the variance of random variable X. The value C of the covariance may appear in the form of any real number depending on the distribution of the variables in question. This means that it would be difficult to conclude on the closeness of a stochastic dependence from the value of C. It appeared to be more rational to settle with another parameter which would have an upper and lower limit, moreover it would carry information about the closeness or looseness of the relationship between the two random variables. This parameter is called correlation coefficient, having the form:

$$\varrho = \frac{E\{[X-E(X)][Y-E(Y)]\}}{D(X)D(Y)}. \tag{2.54}$$

The correlation coefficient is the covariance of random variables X and Y divided by the product of their respective standard deviations. The value of the correlation coefficient ϱ may fluctuate between -1 and $+1$ due to the fact that the covariance in the numerator cannot be larger than the product of the standard deviations. This statement can be verified by the following theorem: if X and Y are random variables with existing expectations of their respective squared values, then

$$|E(XY)| \leqq \sqrt{E(X^2)E(Y^2)}. \tag{2.55}$$

Namely, for any optional, real λ

$$E(\lambda X-Y)^2 \geqq 0$$

and

$$\lambda^2 E(X^2)-2\lambda E(XY)+E(Y^2) \geqq 0.$$

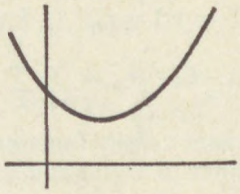


Figure 24

This expression is a second-order polynom for λ which may at best touch the X-axis, or in other words, its discriminant is not positive. So,

$$4[E(XY)]^2-4E(X^2)E(Y^2) \leqq 0$$

from which our earlier statement can be read out.

If now, in Eq. (2.55) $X-E(X)$ is substituted instead of X, and $Y-E(Y)$ instead of Y, then we obtain

$$E\{[X-E(X)][Y-E(Y)]\} \leqq \sqrt{E[X-E(X)]^2 E[Y-E(Y)]^2} = D(X)D(Y)$$

from which it follows that

$$|\varrho| \leqq 1.$$

In case if random variable Y is a linear function of variable X:

$$Y = aX+b$$

then $\varrho=1$, or $\varrho=-1$ according to being $a>0$ or $a<0$.
Its explanation is simple:

$$\varrho = \frac{E\{[X-E(X)][aX+b-aE(X)-b]\}}{D^2(X)D^2(aX+B)} = \frac{aE[X-E(X)]^2}{|a|D^2(X)} = \frac{a}{|a|} = \pm 1.$$

The statement is true in its reverse. If $|\varrho|=1$, then the first variable is a linear function of the second.

For the sake of practical applications expressions of both, covariance and correlation coefficient may be simplified somewhat:

$$C = E\{[X-E(X)][Y-E(Y)]\} = E[(XY-E(X)Y-E(Y)X+E(X)E(Y))] =$$

$$= E(XY)-E(X)E(Y).$$

And from this:

(2.56)
$$\varrho = \frac{E(XY)-E(X)E(Y)}{D(X)D(Y)}.$$

In order to calculate covariance and correlation between random variables X and Y their joint distribution must be known.

If X and Y are discrete, then:

$$P(X = x_i) = p_i, \quad P(Y = y_k) = p_k$$

and

$$P(X = x_i, \; Y = y_k) = r_{ik} \quad (i, k = 1, 2, \ldots)$$

and

$$c = \sum_i \sum_k [x_i - E(X)][y_k - E(Y) r_{ik} = \sum_i \sum_k x_i y_k \cdot r_{ik} - E(X)E(Y).$$

If X and Y are continuous with a joint density function of $h(x, y)$, then:

$$C = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyh(x, y)\, dx\, dy - E(X)E(Y).$$

Here, relationship (2.34) was used.

If X and Y are independent from each other, the covariance of the two is zero. Because the mean of the product of independent random variables is equal to the product of their respective mean values, therefore:

$$C = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0.$$

So, if X and Y are independent

$$\varrho = \frac{C}{D(X)D(Y)} = 0.$$

The statement is not valid for its reverse. If the correlation coefficient between X and Y is zero, it is not necessary, in general, that the two variables are independent of each other. The correlation coefficient is better a measure of the linearity of a stochastic relationship than of closeness! (An example: if X is uniformly distributed in interval $[-1, +1]$ and $Y = 5X^3 - 3X$ then $\varrho = 0$, although there is a functional relationship between X and Y.)

If the correlation coefficient between two random variables, say X and Y, is zero then we speak about uncorrelated Xs and Ys. As this was earlier mentioned, uncorrelated relationship does not mean independence, in general. If, however, the joint distribution of X and Y is a bivariate normal distribution then uncorrelated status is at the same time independent status. Independence is also predetermined by an uncorrelated condition if X and Y are the indicator variables of two different events (see Section 7.1.3).

2.1.8. GENERATING FUNCTION AND CHARACTERISTIC FUNCTION

a) *The generating function*

In the following, two useful analytical tools will be presented: the generating function and the characteristic function. By their help the moments of random variables and the distribution of the sums of independent random variables may be easily calcu-

lated. Also they are used and seem to be indispensable for the determination of the different marginal distributions.

The generating function is used for the investigation of the distribution of non-negative integer random variables.

Let X be a random variable with possible values of integers $0, 1, 2, \ldots, n, \ldots$ and with joint probabilities of $p_0, p_1, p_2, \ldots, p_n, \ldots$

Let formulate the function

$$G(X) = \sum_k p_k x^k \tag{2.57}$$

which will be called the generating function of random variable X. Due to the fact that the p_k numbers are probabilities, and $\sum_k p_k = 1$, the power-series representing this generating function (if k will take countable or finite values) is convergent if $|x| < 1$, and $G(1) = 1$. It follows that if $|x| < 1$, function $G(x)$ is derivable any times, and

$$G'(x) = \sum_k k p_k x^{k-1}.$$

If this power-series is convergent also at $X = 1$, then

$$G'(1) = \sum_k k p_k = E(X)$$

$$G''(x) = \sum_k k(k-1) p_k x^{k-2}$$

$$G''(1) = \sum_k k^2 p_k - \sum_k k p_k = E(X^2) - E(X).$$

It can be seen that the mean of the integer random variable X is:

$$E(X) = G'(1). \tag{2.58}$$

And its variance is:

$$D^2(X) = G''(1) + G'(1) - [G'(1)]^2. \tag{2.59}$$

It should be noted that the generating function

$$G(x) = \sum_k x^k p_k$$

itself is an expectation, the mean of random variable x^X. Namely (according to the definition of the mean):

$$E(x^X) = \sum_k x^k p_k = G(x). \tag{2.60}$$

From this observation an important theorem is derived concerning the generating function of the sum of independent random variables. If $X_1, X_2, \ldots, X_n$ are independent integer random variables and if $Y = X_1 + X_2 + \ldots + X_n$, then the generating

function of this random variable is:

$$G_Y(x) = E(x^Y) = E(x^{X_1+X_2+\ldots+X_n}) = E(x^{X_1} \cdot x^{X_2} \ldots x^{X_n}) =$$
$$= E(x^{X_2}) \ldots E(x^{X_n}) = G_{X_1}(x) G_{X_2}(x) \ldots G_{X_n}(x).$$

Due to the independence of variables $X_1, X_2, \ldots, X_n$ the variables $x^{X_1}, x^{X_2}, \ldots x^{X_n}$ are also independent and the mean of the product of independent random variables is equal to the product of their respective means.

It is now clear that the generating function of the sum of independent random variables is equal to the product of the generating functions of the individual variables. This theorem will be often used for the determination of the distribution of the sums of independent random variables along with the discussion of the most important distribution functions. The generating function is much more suitable for the determination of the distribution of sums than the rule of convolution which is mostly of theoretical importance.

It may often happen that the distribution of the sum of independent random variables—identically distributed—must be determined. In this case function $G_Y(x)$ will take the following form (if n members are at hand):

$$G_X(x) = [G_{X_i}(x)]^n. \tag{2.61}$$

In the explicit form of the generating function $G_Y(x)$ the coefficient of x^k will give the probability $P(Y=k)$.

b) *The characteristic function*

The generating function is defined only for non-negative, integer random variables. In the general case the so-called characteristic function plays a similar role than the generating function has played with integer variables. The characteristic function of a random variable X is defined by the mean of the complex random variable e^{itX}. Let denote the characteristic function of random variable X by $\varphi_X(t)$, then:

$$\varphi_X(t) = E(e^{itX}) = E(\cos tX) + iE(\sin tX) \quad (i^2 = -1).$$

The characteristic function will be used in this book only to continuous random variables. In this case:

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, d(x) \tag{2.62}$$

where $f(x)$ is the density function of random variable X. A characteristic function is called Fourier-transform in mathematical analysis. Some important features of the characteristic function are:

$$|\varphi_X(t)| = \left| \int_{-\infty}^{\infty} e^{itx} f(x)\, dx \right| \leqq \int_{-\infty}^{\infty} |e^{itx}| f(x)\, dx = 1 \tag{2.63}$$

because

$$|e^{itx}| = +\sqrt{\cos^2 tx + \sin^2 tx} = +\sqrt{1} = 1.$$

It should be noted that

$$\varphi_X(-t) = E(e^{-itX}) = E[\cos(-tX) + i\sin(-tX)] =$$
$$= E(\cos tX) - iE(\sin tX) = \bar{\varphi}_X(t).$$

According to the theory of the Fourier-transform, if

$$\int_{-\infty}^{\infty} |\varphi(t)|\, dt < +\infty$$

then

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx}\varphi(t)\, dt. \tag{2.64}$$

In other words, if the characteristic function is known the density function may be determined. A distribution is unambiguously defined by a characteristic function. During our discussions even the inverse formula of (2.64) will not be needed due to the fact that only of a few continuous distribution will their characteristic functions be determined and from the well-known characteristic function the identification of the distribution will be not difficult.

In the following an important theorem related to characteristic functions will be presented for the determination of the distribution of the sums of independent random variables.

Theorem: Let $X_1, X_2, \ldots, X_n$ be independent random variables and suppose that $Y = X_1 + X_2 + \ldots + X_n$. In this case the characteristic function of random variable Y is equal to the product of the characteristic function of the addends. So,

$$\varphi_Y(t) = E(e^{itY}) = E[e^{it(X_1+\ldots+X_n)}] = E[e^{itX_1} \cdot e^{itX_2} \ldots e^{itX_n}] = \tag{2.65}$$
$$= E(e^{itX_1}) \cdot E(e^{itX_2}) \ldots E(e^{itX_n}) = \varphi_{X_1}(t) \ldots \varphi_{X_n}(t) \cdot \varphi_{X_n}(t).$$

Namely, if $X_1, X_2, \ldots X_n$ are independent then $e^{itX_1}, e^{itX_2}, \ldots e^{itX_n}$ are also independent and the mean of their product is equal to the product of their mean values.

If a random variable Y is a linear function of another random variable X, say $Y = aX + b$, then

$$\varphi_Y(t) = E(e^{itY}) = E[e^{it(aX+b)}] = e^{itb}E[e^{i(at)X}] = e^{itb}\varphi_X(at). \tag{2.66}$$

It will be shown now, how the moments of a continuous random variable are determined by aid of the characteristic function (if the moments exist at all).

Expression

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx$$

differentiated according to t, gives:

$$\varphi'_X(t) = \int_{-\infty}^{\infty} ixe^{itx} f(x)\, dx.$$

The mean of the complex random variable $Y=e^{itX}=\cos t\,X+i\sin tX=X_1+iX_2$ is defined by

$$E(Y)=E(X_1)+iE(X_2)$$

and its variance by

$$D^2(Y)=E(|Y-E(Y)|^2).$$

So,

$$\varphi'_X(0)=i\int_{-\infty}^{\infty}xf(x)\,dx=i\alpha_{1'}=iE(X). \tag{2.67}$$

Further

$$\varphi''_X(t)=\int_{-\infty}^{\infty}(ix)^2e^{itx}f(x)\,dx$$

and

$$\varphi''_X(0)=i^2\int_{-\infty}^{\infty}x^2f(x)\,dx=i^2E(X^2).$$

From this:

$$E(X)=\frac{1}{i}\varphi'_X(0)=-i\varphi'_X(0)$$

$$D^2(X)=E(X^2)-[E(X)]^2=\frac{1}{i^2}\varphi''_X(0)-\frac{1}{i^2}\varphi'_X(0)^2$$

$$D^2(X)=[\varphi_X(0)]^2-\varphi''_X(0). \tag{2.68}$$

In general:

$$\varphi_X^{(k)}(t)=i^k\int_{-\infty}^{\infty}x^ke^{itx}f(x)\,dx$$

$$\varphi_X^{(k)}(0)=i^k\int_{-\infty}^{\infty}x^kf(x)\,dx=i^k\alpha_k. \tag{2.69}$$

where α_k is for the kth moment of random variable x. Based on this, the Taylor-series of function $\varphi_X(t)$ may be obtained as follows:

$$\varphi_X(t)=\varphi_X(0)+\frac{\varphi'_X(0)}{1!}t+\frac{\varphi''_X(0)}{2!}t^2+\ldots=1+\alpha_1it+\alpha_2\frac{(it)^2}{2!}+\ldots. \tag{2.70}$$

2.2. REVIEW OF PROBABILITY DISTRIBUTIONS OCCURRING OFTEN IN HYDROLOGY

The simple alternative

2.2.1. THE INDICATOR VARIABLE OF AN EVENT

The simplest discrete distribution is the simple alternative. This distribution is valid if we are interested during an experiment in the fact whether event A has occurred or not. Let suppose that the distributions of $P(A)=p$ and $P(\bar{A})=1-p=q$ are known.

Let now assign to our experiment random variable X. This random variable will take a value 1 if A has occurred and 0 if $\bar{A}$. In mathematical terms $P(X=1)=p$, $P(X=0)=q=1-p$.

The distribution of variable X is usually depicted by the following scheme:

$$X: \begin{array}{c|c} 1 & 0 \\ \hline p & q \end{array}.$$

Let calculate the mean and variance of such an indicator variable. According to the definition of the mean, and based on the distribution of X:

$$E(X) = 1 \cdot p + 0 \cdot q = p. \tag{2.71}$$

And

$$E(X^2) = 1^2 \cdot p + 0^2 \cdot p = p$$

$$D^2(X) = E(X^2) - E^2(X) = p - p^2 = p(1-p) = pq. \tag{2.72}$$

The generating function of an indicator variable is:

$$G(x) = \sum_{k=0}^{1} p_k x^k = qx^0 + px = px + q. \tag{2.73}$$

2.2.2. THE BINOMIAL DISTRIBUTION

In hydrological practice the most important discrete probability distribution from the point of view of theory and everyday application is the binomial or Bernoulli-distribution. It is often experienced that the following problem is of interest: will event A occur or not (will event $\bar{A}$ occur), and our experiment is repeated several times in consequence.

If an experiment consists of a series of simple alternatives and the outcomes are independent of each other then this experiment is called Bernoulli serial experiment. Such experiment is, for example, the so-called head-tail game by tossing a coin. Let now present an example for a series of alternatives in hydrological practice. During the past 100 years (1876—1976) the following annual water stage maxima had been registered for the Tisza river at Szeged:

Table T.2

Tisza river
Annual maximum stages

Year	Tokaj Annual max., cm	Szolnok Annual max., cm	Szeged Annual max., cm
1876	784	753	786
1877	710	688	795
1878	694	638	720
1879	755	763	806
1880	660	608	627
1881	780	764	845
1882	685	675	691
1883	649	634	738
1884	738	639	613
1885	642	538	565
1886	576	658	534
1887	604	558	660
1888	872	818	847
1889	735	728	805
1890	654	576	566
1891	665	640	668
1892	640	621	630
1893	670	591	726
1894	588	545	568
1895	815	827	884
1896	640	600	525
1897	688	684	730
1898	608	580	604
1899	590	472	460
1900	644	556	525
1901	686	685	680
1902	666	619	668
1903	596	604	508
1904	396	428	450
1905	581	518	518
1906	550	544	550
1907	759	738	758
1908	658	629	595
1909	676	626	642
1910	528	534	496
1911	522	528	563
1912	726	713	753
1913	723	722	802
1914	700	715	778
1915	825	808	791
1916	688	778	791
1917	562	582	614

Table T.2

Year	Tokaj Annual max., cm	Szolnok Annual max., cm	Szeged Annual max., cm
1918	516	462	349
1919	854	882	916
1920	736	716	708
1921	374	378	325
1922	770	784	774
1923	672	634	637
1924	802	846	870
1925	857	574	681
1926	773	778	759
1927	590	540	488
1928	571	572	542
1929	508	522	458
1930	639	586	496
1931	663	602	603
1932	856	894	923
1933	695	662	660
1934	540	572	526
1935	606	618	594
1936	552	564	472
1937	722	750	703
1938	570	621	638
1939	502	586	579
1940	818	880	847
1941	804	856	855
1942	636	728	780
1943	462	430	366
1944	651	662	654
1945	603	638	560
1946	558	617	525
1947	640	633	602
1948	781	784	714
1949	617	578	495
1950	632	602	517
1951	613	634	550
1952	767	734	648
1953	748	801	706
1954	535	549	454
1955	693	646	657
1956	671	678	689
1957	659	645	604
1958	756	708	730
1959	550	490	436
1960	600	599	582
1961	507	394	394
1962	794	836	820

Table T.2

Year	Tokaj Annual max., cm	Szolnok Annual max., cm	Szeged Annual max., cm
1963	740	691	587
1964	857	853	764
1965	725	793	748
1966	755	855	799
1967	831	881	790
1968	740	673	600
1969	597	659	626
1970	858	909	961
1971	630	608	521
1972	564	563	606
1973	518	427	475
1974	801	840	807
1975	686	757	692

Let now the event denote by A that the annual maximum at Szeged is higher than 700 cm. The probability of A should be: $P(A)=p$.

Let calculate the probability that A will occur K times in n years! The number of occurrences of event A should be X during those n years in question. The possible values of X are now $0, 1, 2, \ldots, n$ non-negative integers. The probability $p_k=P(X=k)$ is wanted. (Annual water stage maxima are considered independent events. See Chapter 6, Section 5.1).

Let the occurrence of event A or $\bar{A}$ be registered by 1 and zero, respectively. Then, the space of elementary events of this experiment formulated by n observations will consist of a set of the numbers 0 and 1.

This kind of space of events has been investigated already in Section 1.1 of Chapter 2. In this case, however, the probabilities of the elementary events are not the same. If some ω_i elementary event consists of k ones (and $n-k$ zeros), then

$$P(\omega_i) = p^k q^{n-k}.$$

Event $\{X=k\}$ will occur if the result of our experiment (consisting of n observations) is an elementary event ω_i in which k ones and $(n-k)$ zeros can be found. The number of such ω_i elementary events is $\binom{n}{k}$, each with a probability of $p^k q^{n-k}$. Therefore,

$$P(X = k) = \binom{n}{k} p^k q^{n-k}. \tag{2.74}$$

If we write 1 for the occurrence of A and 0 for $\bar{A}$ then to every observation the indicator variable of event A has been assigned. Obviously,

$$X = X_1 + X_2 + \ldots + X_n$$

where X_i is the indicator variable of observation i. The sum of the values of indicator variables is exactly the number of occurrences of event A since the occurrence of $\{X=k\}$ is identified by k indicator variables taking the value 1 and by $(n-k)$ taking the value of zero.

If X stands for the number of rainy days in May of a given year and the indicator variables for the individual days are $X_1, X_2, \ldots, X_n$ (being 1 if it has rained and 0 if not) then the value of X is obviously equal to the number of ones among the addends:

$$X = X_1+X_2+\ldots+X_n = 1+0+0+1+\ldots+0.$$

The expectation of random variable X with a binomial distribution is:

$$E(X) = E(X_1)+E(X_2)+\ldots+E(X_n) = np. \tag{2.75}$$

Because the mean of each indicator variable is p. The variance of X is

$$D^2(X) = D^2(X_1)+\ldots+D^2(X_n) = pq+pq+\ldots+pq = npq \tag{2.76}$$

according to Eq. (2.44′).
So, according to the standard deviation of the binomial distribution:

$$D(X) = \sqrt{npq}\,. \tag{2.77}$$

The generating function — based on Eq. (1.55) — is:

$$G(x) = \sum_{k=0}^{n} p_k x^k = \sum_{i=0}^{n} G_i(x) = (px+q)^n \tag{2.78}$$

where $G_i(x)=px+q$ stands for the generator function of the characteristic variable X_i. In the generating function $G(x)$ the coefficient of x^k will equal the probability $p_k=P(X=k)$, which in this case will take the form:

$$p_k = \binom{n}{k} p^k q^{n-k}.$$

Mean and variance are easily calculated by the help of the generating function

$$G(x) = (px+q)^n$$

because according to Eqs. (2.58) and (2.59)

$$E(X) = G'(1) = [n(px+q)^{n-1}p]_{x=1} = np \tag{2.79}$$

and

$$D^2(X) = G''(1)+G'(1)-[G'(1)]^2 = \tag{2.80}$$

$$= [n(n-1)p^2(px+q)^{n-2}]_{x=1}+np-n^2p^2 = n^2p^2-np^2+np-n^2p^2 =$$

$$= np(1-p) = npq.$$

The question arises that if p is given, which p_k probability will be the greatest? The values of the numbers p_k will increase up till

$$\frac{p_k}{p_{k-1}} \geqq 1,$$

or till

$$\frac{\binom{n}{k} p^k q^{n-k}}{\binom{n}{k-1} p^{k-1} q^{n-k+1}} = \frac{n-k+1}{k} = \frac{p}{1-p} \geqq 1.$$

In other words till

$$(n+1)p \geqq k.$$

If $(n+1)p$ is integer, the maximum value of the distribution is p_k where the suffix is $k=(n+1)p$. In this case $p_k=p_k=p_{k-1}$ and there are two identical maximums. If $(n+1)p$ is not an integer then the value of p_k will be maximum at a k which is the largest integer still included in $(n+1)p$. If n is large then $(n+1)p \approx np$ due to $p>1$.

Let take an example to illustrate the above discussed topic.

The distribution of rainy days at Budapest, Hungary was on the basis of 50 years of observations (rain is assumed to have a depth of minimum 1 mm):

month	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
rainy days	7.6	6.8	7.3	7.4	8.5	8.0	6.5	6.3	6.2	7.5	8.8	9.1

So, the probability that 7 rainy days will be observed in April of a given year is:

$$p \approx \frac{1}{4}$$

$$p_7 = \binom{30}{7}\left(\frac{1}{4}\right)^7\left(\frac{3}{4}\right)^{23} = 2\,035\,800\,\frac{3^{23}}{4^{30}} \approx 0.16.$$

The calculation of such probabilities is cumbersome due to the presence of the binomial coefficients. Usually, approximations are introduced. E.g. the value of $\binom{30}{7}$ is retrievable from the table of binomial coefficients, the value of $3^{13}\times 4^{-30}$ may be easily determined by logarithms. (See [2.25] and [2.28]).

It is visible that the maximum value of a binomial random variable has a fairly small probability. The situation is different if the question sounds: what is the probability of falling, say, between 5 and 11 of the number of rainy days in April?

$$P = \sum_{k=5}^{10} \binom{30}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{30-k}.$$

The calculation of this probability would be extremely cumbersome. Therefore, approximative methods must be introduced. See: Section 2.2.8.

2.2.3. MULTINOMIAL DISTRIBUTION

If an experiment may have different outcomes denoted by $A_1, A_2, \ldots, A_s$ ($A_1, A_2, \ldots, A_s$ compose a complete system of events), and if

$$P(A_1) = p_1, \; P(A_2) = p_2, \ldots, P(A_s) = p_s$$

and $\sum_1^s p_i = 1$, then by n continuous repetitions a so-called multinomial distribution will be obtained. The notion is basically a generalization of the binomial distribution. If $s=2$, the two distributions are the same.

The probability that event A_1 will occur k_1 times, $A_2 k_2$ times, and $A_s k_s$ times is:

$$p_{k_1, k_2, \ldots, k_s} = \frac{n!}{k_1!\, k_2! \ldots k_s!} p_1^{k_1} p_2^{k_2} \ldots p_s^{k_s}.$$

2.2.4. THE GEOMETRIC DISTRIBUTION

Let consider an experiment with two possible outcomes. This is called — from earlier chapters — simple alternative. Let denote the first alternative by A and the second by $\bar{A}$. Let be

$$P(A) = p_1, \quad P(\bar{A}) = 1 - p = q.$$

Suppose, that the experiment is repeated till event A will occur for the first time. The probability that A will occur for the first time in the kth experiment is:

$$p_k = q^{k-1} p. \tag{2.81}$$

It is easy to see that $\sum_{k=1}^{\infty} p_k = 1$, since

$$\sum_1^{\infty} p_k = p \sum_1^{\infty} q^{k-1} = \frac{p}{1-q} = \frac{p}{p} = 1.$$

The generating function of a geometric distribution is:

$$G(x) = \sum_{k=1}^{\infty} p q^{k-1} x^k = px \sum_1^{\infty} (qx)^{k-1} = \frac{px}{1-qx}. \tag{2.82}$$

Because

$$G'(x) = \frac{p(1-q)+pqx}{(1-qx)^2} = \frac{p}{(1-qx)^2},$$

and

$$G''(x) = \frac{2pq\,(1-qx)}{(1-qx)^4} = \frac{2pq-2pq^2 x}{(1-qx)^4}.$$

Expectation and standard deviation of a geometric distribution are:

$$E(X) = G'(1) = \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p} \tag{2.83}$$

$$D^2(X) = G''(1) + G'(1) - [G'(1)]^2 = \tag{2.84}$$

$$= \frac{2p^2 q}{p^4} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}.$$

2.2.5. THE POISSON-DISTRIBUTION

From among the discrete probability distributions great theoretical and practical importance is attributed to the so-called Poisson-distribution. It can be derived — as a simplest approach — from the binomial distribution — a limiting case of this latter — if the number of experiments n is large, and p (the probability of an event in which we are interested) is small. The Poisson-distribution is called sometimes the distribution of rare events or of events with a small probability. The importance of the Poisson-distribution, however, is by no means in the fact that it is a good approximation of the binomial distribution.

The Poisson-distribution is a suitable descriptor of a number of natural processes. The so-called random point "processes" (say, the number of flood waves during a $(0, t)$ time interval e.g. a season or quarter of a year) follow Poisson-distribution as this will be demonstrated in Section 2.2.6. Similarly, the Poisson-distribution is used for the description of random point scattering in certain circumstances. This will be later shown in examples.

Let denote by A the event that the annual maximum water stage exeeds a given value c. If $P(X \geqq c) = P(A) = p$, then $P(\bar{A}) = 1 - p$. The case here is the simple alternative. So, the probability that A will occur k times during n years is (supposed that the annual maximum stages are independent):

$$p_k^{(n)} = \binom{n}{k} p^k (1-p)^{n-k}.$$

The formula is cumbersome if n is large, therefore, let us write the above relation in the following form:

$$p_k^{(n)} = \frac{1}{k!} \frac{n(n-1)\dots(n-k+1)}{n^k} (np)^k \left(1 - \frac{np}{n}\right)^{n-k}$$

If $n \to \infty$ and $p \to 0$ and $np = \lambda$ is meanwhile constant, then

$$p_k^{(n)} = \frac{1}{k!} \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) \lambda^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

By letting $n \to \infty$:

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{2.85}$$

Here $e = 2.718\,281\dots$ is nothing else than the limit of the sequence $\left(1 + \frac{1}{n}\right)^n$ $(n = 1, 2, \dots)$.

If level c is high enough in connection with annual maximum stages, and the probability $P(A) = P(X \geqq c) = p$ is low, then the probability that in the coming great number of years (n is large) A will occur k times, may be well approximated by a Poisson-distribution.

If a random variable Y stands for the number of occurrences of event A, then the distribution of Y is:

k	0	1	2	...	k	...
$p_k = P(Y=k)$:	$\frac{\lambda^0}{0!}e^{-\lambda}$	$\frac{\lambda}{1!}e^{-\lambda}$	$\frac{\lambda^2}{2!}e^{-\lambda}$	...	$\frac{\lambda^k}{k!}e^{-\lambda}$	... .

If k is given, the value of p_k is a function of parameter λ. Because $\lambda = np$, and np is the mean of the binomial distribution — when n trial has been made and the probability of the event in which we are interested is p — so λ is also an expectation. This statement can be also formally proven:

$$E(Y) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \tag{2.86}$$

$$= \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

Let now calculate the variance of the Poisson-distribution:

$$D^2(Y) = E(Y^2) - [E(Y)]^2 = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 = \tag{2.87}$$

$$= \lambda e^{-\lambda} \left[\sum_{k=1}^{\infty} (k-1+1) \frac{\lambda^{k-1}}{(k-1)!} \right] - \lambda^2 =$$

$$= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} +$$

$$+ \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} - \lambda^2 =$$

$$= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda e^{-\lambda} e^{\lambda} - \lambda^2 = \lambda.$$

The standard deviation of the Poisson-distribution is:

$$D(Y) = +\sqrt{\lambda}.$$

In order to look after probabilities in the Poisson-tables it is indispensable to know the value of parameter λ. In practice it is usually estimated statistically (See Section 5.1.4).

2.2.6. EXAMPLE FOR THE POISSON-DISTRIBUTION IN FLOOD HYDROLOGY. PROBABILITY DISTRIBUTION OF THE NUMBER OF FLOODS

In flood protection it is of cardinal importance to know the expected number of floods in a given cross-section. Let now see, how the probability distribution of the number of floodwaves may be determined. It should be noted that if in the selected time interval $(0, t)$ the water regime is considered generally as homogeneous (according to experience in Hungary a quarter of a year will still fulfill this requirement), the distribution of the number of floods is easily calculated.

Let select for $(0, t)$ say the first quarter of each year, and let denote the number of exceedances by random variable v in this interval. It is obvious that v will take the values represented by 0, 1, 2, ... non-negative integers. In this context v is a discrete integer random variable. When the distribution of v is sought for the following questions must be answered: in how many first quarters of the investigated years will we have zero, one, two, three, ... etc. floodwaves? To discover something for the future let us turn for information to the past. Suppose, we are investigating again the springtime water regime of the Tisza River — at the water stage of Szolnok — from the point of view of the number of floods. Exceedances observed during the period of 1903—1970 are summed up in Table 2.2. From this table it is apparent that during 68 years there were 34 spring-seasons when there was no exceedance above $c=$ $=600$ cm, in other words, the frequency of event $\{v=0\}$ was 34. The frequencies of events $\{v=1\}$, $\{v=2\}$ and $\{v=3\}$ were 26, 6 and 2, respectively. According to observations, no more exceedances than 3 occurred in one spring period. During these 68 years altogether 44 exceedances were recorded at Szolnok, which has set the average numbe of the spring floods to $\lambda=\frac{44}{68}\approx 0.65$.

If the frequency of event $\{v=k\}$ is denoted by v_k and its relative frequency is $\frac{v_k}{n}$, the following table can be constructed:

Table 2.2

Water stage Szolnok 1st quarter $c=600$ cm 1903—1970

k	v_k	Poisson-distribution: np_k	$\lambda=0.7$
0	34	34	
1	26	23.8	
2	6	8.16	
3	2	2.04	

If the distribution of a random variable is at stage it is not appropriate to rely solely on the data of a single gauging station. Let see, therefore, the situation at another stage e.g. at Szeged (also along the Tisza River). What was the number of exceedances in the 2nd quarter (between April 1 and June 30) above level $c=650$ cm in the period of 1901—1970. The gathered information is presented in Table 2.3

Table 2.3

Szeged 2nd quarter $c=650$ cm 1901—1970

k	v_k	Poisson-distribution: np_k	$\lambda=0.4$
0	45	45.6	
1	21	18.4	
2	4	4.1	

(In the last column of these tables the corresponding values of the Poisson-distribution are presented, later being referred to).

Let now introduce the following combinatorial solution for the determination of the distribution of v!

Let consider the first (or second, etc.) quarters of consecutive years like adjecent disjoint time intervals on the real axis of the line of numbers.

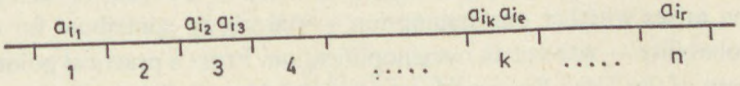


Figure 25

Suppose, we have n years of observations and r exceedances were measured above a given c level. Let denote these by $a_1, a_2, \ldots, a_r$! Think about having n numbered cells for the consecutive n years in which the exceedances $a_1, a_2, \ldots, a_r$ are placed, see Figure 25. Let now suppose that every position of the exceedances has the same probability*: $\frac{1}{n^r}$.

In practice the number of cells n is given, it is the number of years for which observations exist. The number of the allocated exceedances r is a function of the selected c level. If c is high, r is obviously small.

Expression $\frac{r}{n}=\lambda$ is for the average number (the mean) of exceedances located in one cell. In case of our model we may easily calculate the probability of having exactly k exceedances in a randomly selected cell. From among the r exceedances k are selected (this can be done in $\binom{r}{k}$ ways) and placed in a given cell. The rest ($r-k$ exceedances) may be placed in $n-1$ cells according to $(n-1)^{r-k}$ variations. (Repetitive variation.)

The number of all possible allocations is n^r. Accordingly, the probability that exactly k exceedances should be allocated in a given cell is:

$$p_k = \frac{\binom{r}{k}(n-1)^{r-k}}{n^r} = \binom{r}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{r-k} = \tag{2.87}$$

$$= \binom{r}{k} p^k (1-p)^{r-k}$$

$$\text{where } p = \frac{1}{n}.$$

* This assumption may seem unrealistic for the first glance. The calculation shown in page 90 convinces us that in a position in which many balls were placed — their common probability was neglectable.

This is the well-known binomial distribution. If n and r are equally large but with the condition that $\frac{r}{n}=\lambda$ (const) then due to the smallness of $p=\frac{1}{n}$ the binomial distribution can be approximated by a Poisson-distribution, like:

$$p_k \approx \frac{\lambda^k}{k!} e^{-\lambda}. \tag{2.88}$$

The question arises whether our assumption — namely to contribute for each cell the same probability — was not an oversimplification. From a practical point of view, e.g. in the case of the Tisza River we have observed 31 exceedances at Szeged during 70 years. The assumption that all of these could fall in one quarter of a year or in some quarters of the years is absurd because we know that not more than 3 exceedances did fall in any of the quarters.

In reality if $r=31$ and $n=70$ are put in Eq. (2.87), then

$$p_0 = \left(1-\frac{1}{70}\right)^{31} \approx 0.64$$

$$p_1 = \binom{31}{1}\frac{1}{70}\left(1-\frac{1}{70}\right)^{30} \approx 0.29$$

$$p_2 = \binom{31}{2}\frac{1}{70^2}\left(1-\frac{1}{70}\right)^{29} \approx 0.06$$

$$p_3 = \binom{31}{3}\frac{1}{70^3}\left(1-\frac{1}{70}\right)^{28} \approx 0.009$$

$$p_0+p_1+p_2+p_3 \approx 0.999.$$

2.2.7. THE NORMAL DISTRIBUTION (GAUSSIAN DISTRIBUTION)

If the outcome of an experiment is influenced by a great number of factors — independent or almost independent from each other — and the individual factors, one by one, are contributing in an extremely limited way to the overall fluctuation of the outcome, moreover the effects of the individual factors may be simply summed up like additive terms, we are faced by a so-called normal distribution. Such circumstances are often encountered in our practice. This fact leads to the central role of the normal or Gaussian distribution in probability theory. The most often covered areas by this distribution are: investigation of statistical populations, error-analysis and approximation of the binomial distribution.

By a more accurate, mathematical definition: the distribution of a random variable is normal if its density-function has the following form:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{2.89}$$

where m and σ are constants, and $\sigma>0$. The quantities m and σ are called parameters of the normal distribution. It is easy to see from Eq. (2.89) that if m is a finite number and $\sigma>0$, then the value of function $f(x)$ is positive for each x. The first factor because of $\sigma>0$, and the second due to

$$1/\exp\left(\frac{x-m}{\sigma\sqrt{2}}\right)^2$$

which is always positive, or zero. Consequently, $f(x)$ will be maximum if $x=m$. Then

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}.$$

The smaller the value of paramter σ, the larger the maximum of $f(x)$. If $\sigma=1$, the maximum is

$$f_{\max}(x)=\frac{1}{\sqrt{2\pi}}\approx 0.4.$$

If the value of σ is fixed, then $f(x)$ will depend only on $(x-m)^2$ which is the same whatsoever the sign of the difference $x-m$ would be. The value of $f(x)$ is a function of the distance of variable x from the point $x_0=m$. Also, $f(x)$ is symmetric for $x_0=m$. The shape of function $f(x)$ is the so-called bell-curve. Its changes are depicted in Fig. 26 for different σ-s.

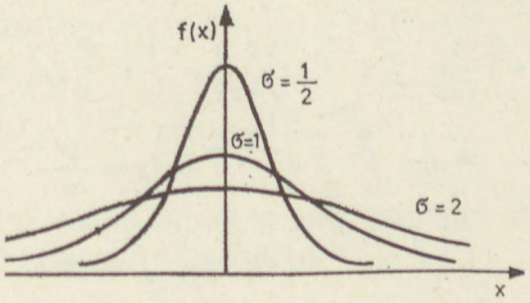


Figure 26

If Eq. (2.89) is a density-function, the area below the curve must be unity. It is not difficult to show the reality of such a statement. Let introduce in the following integral

(2.90) $$I=\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)}{2\sigma^2}}dx$$

substitution $u=\dfrac{x-m}{\sigma}$. Then $x=\sigma u+m$ and $dx=\sigma du$, so

$$I=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{u^2}{2}}du,$$

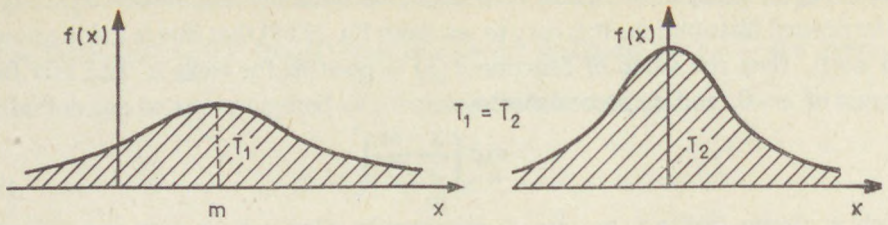


Figure 27

having the same value like Eq. (2.90). From this expression it is apparent that for a curve with parameters of any m and any $\sigma>0$ the area enclosed by the function and the x-axis is the same as for a curve with parameters $m=0$ and $\sigma=1$. In other words, the extent of the area below the density-function of a normal distribution does not depend on its paramters m and σ. So, it is sufficient to prove that

$$I=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{u^2}{2}}\,du=1. \tag{2.90'}$$

This can be done easily if the validity of $I^2=1$ can be proved.

$$I^2=\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{u^2}{2}}\,du\right)^2=$$

$$=\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx\right)\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy\right)=$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy=\frac{1}{2\pi}\cdot$$

$$\cdot\int_0^{2\pi}\left[\int_0^{\infty} re^{-\frac{r^2}{2}}\,dr\right]d\varphi=\frac{1}{2\pi}\int_0^{2\pi} d\varphi=1.$$

This has proven the unity of the area below Eq. (2.89) which has turned out to be a density-function $f(x)$ having always positive ordinates.

The distribution function of the normal distribution has the following form:

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}}\,dt. \tag{2.91}$$

If substitution $u=\dfrac{t-m}{\sigma}$ is introduced in Eq. (2.91), then

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-m}{\sigma}} e^{-\frac{u^2}{2}}\,du=\Phi\left(\frac{x-m}{\sigma}\right) \tag{2.92}$$

where $\Phi(x)$ denotes the normal distribution function with parameters $m=0$ and $\sigma=1$. The values of function $\Phi(x)$ are presented in Table I.

The essence of Eq. (2.92) is here discussed in more detail. Let be X a normal random variable with a distribution function of changing m and $\sigma>0$ parameters by short notation: $X: N(m, \sigma)$. The cumulative distribution function

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

will give then the probability of X being smaller than x.

Instead of the actually received values of random variable X (the outcomes of our experiment) let now investigate the values of $X^*=\frac{X-m}{\sigma}$. X^* is again a random variable determined unabigously by the values of X. Random variable X^* is called a *standardized* variable. If X would take a value of $X=x_0$ in the course of an experiment then X^* will appear as $\frac{x_0-m}{\sigma}$. Let denote the cumulative distribution function of X^* as $\Phi(x)$. It is trivial that the probability of event $\{X<x\}$ will be the same like of $\left\{X^*<\frac{x-m}{\sigma}\right\}$:

$$P(X<x) = P(\sigma X^*+m<x) = P\left(X^*<\frac{x-m}{\sigma}\right),$$

or

$$F(x) = \Phi\left(\frac{x-m}{\sigma}\right).$$

The probability that a random variable normally distributed and with parameters m and $\sigma>0$ will fall in an interval (a, b) is:

(2.93) $$F(b)-F(a) = P(a \leqq X < b) = P(a \leqq \sigma X^*+m < b) =$$

$$= P\left(\frac{a-m}{\sigma} \leqq X^* < \frac{b-m}{\sigma}\right) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right).$$

The tabular presentation of function $\Phi(x)$ is given, in general, for positive x-s. It can be shown that the values of $\Phi(x)$ for negative x-s are also easily obtainable.

The density-function of the standardized normal random variable X^* can be derived from Eq. (2.89) by substitutions $m=0$ and $\sigma=1$.

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Function $\varphi(x)$ is symmetric for point $x=0$, so

$$\Phi(-x) = \int_{-\infty}^{-x} \varphi(t)dt = \int_{x}^{\infty} \varphi(t)dt = \tag{2.94}$$

$$= 1 - \int_{-\infty}^{x} \varphi(t)dt = 1 - \Phi(x)$$

which is apparent also from the diagram of the distribution function $\Phi(x)$, see Figure 28.

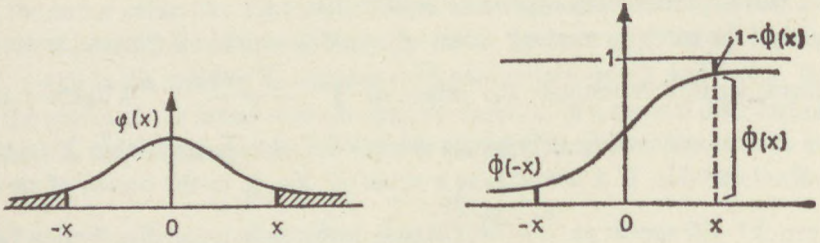


Figure 28

The probability that the value of random variable X^* is in the interval $(-x, x)$ is:

$$\Phi(x) - \Phi(-x) = \Phi(x) - [1 - \Phi(x)] = 2\Phi(x) - 1. \tag{2.95}$$

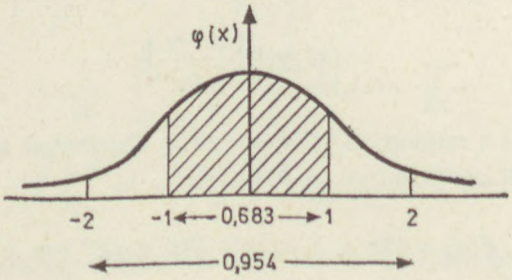


Figure 29

It is easy to understand from Eq. (2.92) that the following two tables are equivalent:

$X^*: N(0, 1)$	$X: N(m, \sigma)$
$P(-0.67 \leqq X^* < 0.67) = 0.5$	$P(m - 0.67\sigma \leqq X < m + 0.67\sigma) = 0.5$
$P(-1 \leqq X^* < 1) = 0.689$	$P(m - \sigma \leqq X < m + \sigma) = 0.683$
$P(-2 \leqq X^* < 2) = 0.954$	$P(m - 2\sigma \leqq X < m + 2\sigma) = 0.954$
$P(-3 \leqq X^* < 3) = 0.997$	$P(m - 3\sigma \leqq X < m + 3\sigma) = 0.997$

It is obvious from these tables that the probability is extremely small, less than 0.05, that the observed value of a random variable X in a normal distribution with parameters m and σ will fall beyond the distance of 2σ along the x-axis. This fact will be often used in mathematical statistics. This is the so-called 2σ-rule. Practically, it is almost to be taken as certain that the observed value of X will be in between the interval $(m-3\sigma, m+3\sigma)$. The above presented facts tend to suggest that m is the expectation and σ is the standard deviation of random variable X.

And really

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(x-m)^2}{2\sigma^2}} dx =$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m+m) e^{-\frac{(x-m)^2}{2\sigma^2}} dx =$$

$$= \frac{\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{x-m}{\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx + \frac{m}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-m)^2}{2\sigma^2}} dx.$$

Let now apply the substitution $u = \frac{x-m}{\sigma}$. Then,

$$E(X) = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-\frac{u^2}{2}} du + \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = m. \tag{2.96}$$

Moreover,

$$D^2(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \tag{2.97}$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-\frac{u^2}{2}} du = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \cdot u e^{-\frac{u^2}{2}} du =$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \left[-u e^{-\frac{u^2}{2}}\right]_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = \sigma^2.$$

In the following, the characteristic function of the normal distribution will be often used. It was first assessed for random variable X^* with a standard normal distribution $N(0, 1)$.

(2.98)

$$\varphi(t) = E(e^{itX^*}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-it)^2}{2}} e^{-\frac{t^2}{2}} dx = e^{-\frac{t^2}{2}}.$$

Here we apply substitution $x - it = z$ and get the result usually obtained in mathematical analysis:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty-it}^{+\infty-it} e^{-\frac{z^2}{2}} dz = 1.$$

If now the distribution of random variable X is $N(m, \sigma)$, in other words, if

$$X = \sigma X^* + m$$

then according to relationship (2.66)

$$\varphi_X(t) = E(e^{it(\sigma X^* + m)}) = e^{itm} E(e^{i\sigma t X^*}) = e^{itm} e^{-\frac{\sigma^2 t^2}{2}} = e^{itm - \frac{\sigma^2 t^2}{2}}. \tag{2.99}$$

Of course, with the help of the derivatives of the characteristic function $\varphi_X(t)$ taken at $t=0$ the moments of the normal distribution are easily obtained. (Let leave this derivation to the reader.)

Let now have two normally distributed independent random variables X_1 and X_2, with distributions $N(m_1; \sigma_1)$ and $N(m_2; \sigma_2)$, respectively. Let be $X = X_1 + X_2$. The task is to determine the distribution of X. Due to the fact that a characteristic function contains in itself precisely the character of the distribution it is sufficient to determine the characteristic function. According to formula (2.99):

$$\varphi_X(t) = E[e^{it(X_1+X_2)}] = E[e^{itX_1} \cdot e^{itX_2}] =$$

$$= \varphi_{X_1}(t) \cdot \varphi_{X_2}(t) = e^{itm_1 - \frac{\sigma_1^2 t^2}{2}} \cdot e^{itm_2 - \frac{\sigma_2^2 t^2}{2}} = e^{it(m_1+m_2) - \frac{(\sigma_1^2+\sigma_2^2)t^2}{2}}.$$

From this it is apparent that

$$E(X) = m_1 + m_2$$

$$D^2(X) = \sigma_1^2 + \sigma_2^2. \tag{2.100}$$

The result is a proof of being the sum of independent normally distributed random variables again a normally distributed random variable. This is true for any desired finite number of independent normally distributed random variables.

In the following, attention will be called to certain important properties of the normal distribution. These are extremely useful from the point of practical applications.

If X is normally distributed with expectation of $E(X)=m$ and standard deviation of $D(X)=\sigma$, moreover $Y=aX+b$ is a linear function of X then Y will also be normally distributed with expectation $E(Y)=am+b$ and standard deviation $D(Y)=$ $=a\sigma$.

This statement may be easily proven by the use of relationship (2.16). The density-function of random variable X is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

and according to formula (2.16) the density-function of variable Y is:

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right) = \frac{1}{|a|\sigma\sqrt{2\pi}} e^{-\frac{\left(\frac{y-b}{a}-m\right)^2}{2\sigma^2}},$$

in other words

$$g(y) = \frac{1}{|a|\sigma\sqrt{2\pi}} e^{-\frac{(y-am-b)^2}{2a^2\sigma^2}}. \tag{2.101}$$

This latter formula is our proof of the above formulated statement since (2.101) is a normal density-function in which the expectation is $am+b$ and the standard deviation: $a\sigma$.

Normal distributions are uniquely determined by their parameters m and $\sigma>0$. From this, it follows that if we are encountered with two normally distributed random variables — with $X: N(m_1, \sigma_1)$ and $Y: N(m_2, \sigma_2)$ — then we will always find numbers $a>0$ and $b>0$, by which the distribution of $Y=aX+b$ is $N(m_2, \sigma_2)$.

In other terms, a normally distributed random variable can be converted in another normal distribution by linear transformation. The numbers $a>0$ and $b>0$ should be selected in such a way that $am_1+b=m_2$ and $a^2\sigma_1^2=\sigma_2^2$. Because $a>0$, it is enough that one should have

$$a\sigma_1 = \sigma_2$$

$$a = \frac{\sigma_2}{\sigma_1} \quad \text{and} \quad b = m_2 - \frac{\sigma_2}{\sigma_1} m_1.$$

2.2.8. APPROXIMATION OF THE BINOMIAL DISTRIBUTION BY NORMAL DISTRIBUTION. THE MOIVRE—LAPLACE THEOREM

It was mentioned at the discussion of the binomial distribution

$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \tag{2.102}$$

that the calculation of the probabilities p_k — if n is large — becomes extremely cumbersome. Therefore, it has been aimed at to replace formula (2.102) by some other, simple, easily used approximation. Depending on the value of parameter p two procedures are applied if binomial distributions must be approximated. If the value of p is between the limits:

$$\frac{0.637}{\sqrt[3]{n}} < p < 1 - \frac{0.637}{\sqrt[3]{n}},$$

(where n is the number of trials), the approximation will be:

$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}}. \tag{2.103}$$

This approximation may be defined — in other words — like this: if the value of p is in the range of the two aforementioned limits, the binomial distribution is substituted by a normal distribution with expectation np and standard deviation $\sqrt{npq}$. If the value of p is outside the limits (close to zero or 1) the Poisson-distribution should be used to approximate the binomial. Eq. (2.103) works best if n is large and p is close to $\frac{1}{2}$, or if $p \approx q = 1-p$.

If substitution $x=\dfrac{k-np}{\sqrt{npq}}$ is introduced in formula (2.103), then

$$p_k \approx \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{npq}} = \varphi(x) \frac{1}{\sqrt{npq}}, \tag{2.104}$$

where $\varphi(x)$ is a standard normal density-function.

In practice it is not often that the values of individual terms of the binomial distribution are of interest but rather the sum of certain terms.

The mathematical derivation of Eq. (2.103) by means of mathematical analysis is simple. Nothing else is needed than the Stirling-formula and the Taylor-series of function $y=\ln(1+x)$.

By use of the Stirling-formula:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \frac{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}}{\left(\frac{k}{e}\right)^k \left(\frac{n-k}{e}\right)^{n-k} \sqrt{2\pi k}\sqrt{2\pi(n-k)}} =$$

$$= \frac{1}{\sqrt{2\pi n \frac{k}{n}\left(1-\frac{k}{n}\right)}} \cdot \frac{n^n}{k^k (n-k)^{n-k}}.$$

After this, the probability p_k may be written in the following form:

$$p_k = \binom{n}{k} p^k q^{n-k} \frac{1}{\sqrt{2\pi n \frac{k}{n}\left(1-\frac{k}{n}\right)}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}.$$

Let introduce substitution $x=k-np$, then

$$k = np+x$$

$$n-k = nq-x$$

and

$$p_k \approx \frac{1}{\sqrt{2\pi n\left(p+\frac{x}{n}\right)\left(q-\frac{x}{n}\right)}} \cdot \tag{2.105}$$

$$\cdot \frac{1}{\left(1+\frac{x}{np}\right)^{np+x} \left(1-\frac{x}{nq}\right)^{np-x}}.$$

It is easy to see that if n increases then $\dfrac{x}{n} \to 0$ and the first factor will tend to $\dfrac{1}{\sqrt{2\pi npq}}$, the same as in formula (2.103).

The logarithm of the second factor is:

$$(np+x)\ln\left(1+\frac{x}{np}\right)+(nq-x)\ln\left(1-\frac{x}{nq}\right)=$$

$$=(np+x)\left[\frac{x}{np}-\frac{x^2}{2n^2\,p^2}+\frac{x^3}{3n^3\,p^3}\right]-\ldots-$$

$$-(nq-x)\left[\frac{x}{nq}+\frac{x^2}{2n^2\,q^2}+\frac{x^3}{3n^3\,q^3}+\ldots\right]=$$

$$=x-\frac{x^2}{2np}+\frac{x^3}{3n^2\,p^2}-\ldots+\frac{x^2}{np}-\frac{x^3}{2n^2\,p^2}+\ldots-$$

$$=-x-\frac{x^2}{2nq}-\frac{x^3}{3n^2\,q^2}-\ldots+\frac{x^2}{nq}+\frac{x^3}{2n^2\,q^2}+\ldots=$$

$$=\frac{x^2}{2n}\left(\frac{1}{p}+\frac{1}{q}\right)-\frac{x^3}{6n^2}\left(\frac{1}{p^2}-\frac{1}{q^2}\right)-\ldots\approx\frac{x^2}{2npq}.$$

Beginning with the second term of this expression subsequent members may be neglected because $\frac{x^3}{n^2}\to 0$ if $n\to\infty$. To agree with this one should see that $x=k-np$, and the deviation of random variable k from its expectation has an order of magnitude of $\sqrt{n}$, which is coming quite straight from the Chebyshev-inequality. Followingly,

$$\frac{x^3}{n^2}\approx\frac{x^{3/2}}{n^2}\approx\frac{1}{\sqrt{n}}\to 0.$$

The denominator of the second factor of Eq. (2.105) is $\approx e^{-\frac{x^2}{2npq}}$ and the probability denoted by p_k is:

$$p_k\approx\frac{1}{\sqrt{2\pi npq}}e^{-\frac{x^2}{2npq}}.$$

2.2.9. TWO-DIMENSIONAL NORMAL DISTRIBUTION

From among a number of multi-dimensional distributions only the two-dimensional normal will be discussed here in detail due to its great practical importance.

Random vector $\vec{Z}=(X, Y)$ is called two-dimensional normal distribution if its density-function is:

(2.106)
$$h(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}}\,exp\left\{-\frac{1}{2(1-\varrho^2)}\cdot\right.$$

$$\left.\cdot\left[\frac{(x-m_1)^2}{\sigma_1^2}-2\varrho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2}+\frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}$$

The geometry of this density-function is a continuous surface with the approximative shape of a bell.

There are five parameters (constants) in the formula. Depending on the value of these paramters the shape of the surface is like that of a real bell or of a compressed bell from two sides. This vague similarity will be specified based on geometrical investigations described later on. Let, however, see first what the meaning of the parameters in Eq. (2.106) would be.

It will be shown that m_1 and σ_1 are the mean and standard deviation of random variable X, m_2 and σ_2 of random vaiable Y, and ϱ is the correlation coefficient calculated between X and Y.

This should be derived as it follows. On the basis of the theory of two-dimensional distributions see (Eq. (1.15)) the density-function of X was:

$$f(x) = \int_{-\infty}^{\infty} h(x, y)dy.$$

Let introduce in Eq. (2.106) the following substitutions:

$$u = \frac{x - m_1}{\sigma_1}, v = \frac{y - m_2}{\sigma_2}$$

then

$$f(x) = \int_{-\infty}^{\infty} h(x, y)dy = \frac{1}{2\pi\sigma_1\sqrt{1-\varrho^2}}$$

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2(-\varrho^2)}[u^2 - 2\varrho uv + v^2]} dv =$$

$$= \frac{e^{-\frac{u^2}{2}}}{2\pi\sigma_1\sqrt{1-\varrho^2}} \int_{-\infty}^{\infty} e^{-\frac{(v-\varrho u)^2}{2(1-\varrho^2)}} dv.$$

By introducing variable $z = \frac{v - \varrho u}{\sqrt{1-\varrho^2}}$, $dv = \sqrt{1-\varrho^2}dz$ and so:

$$f(x) = \frac{e^{-\frac{u^2}{2}}}{2\pi\sigma_1} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{u^2}{2}} \cdot$$

$$\cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{u^2}{2}},$$

and

$$f(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}}$$

which is a proof that X is normally distributed with mean and standard deviation of m_1 and σ_1, respectively. Formula (2.106) is symmetric for variables X and Y, therefore, the density-function of Y may be obtained in an analog way:

$$g(y) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}}$$

which is again a proof that Y is normally distributed with mean and standard deviation of m_2 and σ_2.

After this short investigation the roles played by parameters m_1, m_2 and σ_1, σ_2 could be better defined and it became clear that the common distribution of random variables X and Y is a two-dimensional normal distribution in which the components are also normally distributed. In other words, the marginals of two-dimensional normal distributions are one-dimensional normal distributions.

It will be shown that parameter ϱ found in Eq. (2.106) is the correlation coefficient between X and Y. The covariance between variables X and Y is, on the basis of Eq. (1.46):

$$C = E(X-m_1)(Y-m_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-m_1)(y-m_2)\,h(x,y)dxdy =$$

$$= \frac{1}{2\sigma_1\sigma_2\sqrt{1-\varrho^2}} \int_{-\infty}^{\infty} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}} dy \cdot$$

$$\cdot \int_{-\infty}^{\infty} (x-m_1)(y-m_2) e^{\frac{1}{2(1-\varrho^2)}\left(\frac{x-m_1}{\sigma_1}-\varrho\frac{y-m_2}{\sigma_2}\right)^2} dx.$$

If now the following substitutions are applied:

$$u = \frac{1}{1-\varrho^2}\left(\frac{x-m_1}{\sigma_1} - \frac{y-m_2}{\sigma_2}\right), \quad v = \frac{y-m_2}{\sigma_2}$$

then

$$C = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\sigma_1\sigma_2\sqrt{1-\varrho^2}\,uv + \varrho\sigma_1\sigma_2 v^2)$$

$$e^{-\frac{u^2}{2}} \cdot e^{-\frac{v^2}{2}} du\,dv =$$

$$= \frac{\varrho\sigma_1\sigma_2}{2\pi} \int_{-\infty}^{\infty} v^2 e^{-\frac{v^2}{2}} dv \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du +$$

$$+ \frac{\sigma_1\sigma_2\sqrt{1-\varrho^2}}{2\pi} \int_{-\infty}^{\infty} ve^{-\frac{v^2}{2}} dv \int_{-\infty}^{\infty} ue^{-\frac{u^2}{2}} du = \varrho\sigma_1\sigma_2$$

in other words

$$\varrho = \frac{c}{\sigma_1 \sigma_2} = \frac{E[(X-m_1)(Y-m_2)]}{\sigma_1 \sigma_2}$$

which is exactly the definition of the correlation coefficient.

By analysing Eq. (2.106) it is visible that if X and Y are uncorrelated, i.e. $\varrho=0$, then

$$h(x, y) = \frac{1}{2\pi \sigma_1 \sigma_2} e^{-\frac{1}{2}\left[\frac{(x-m_1)^2}{\sigma_1^2}+\frac{(y-m_2)^2}{\sigma_2^2}\right]} = \tag{2.107}$$

$$= \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(y-m_2)^2}{2\sigma_1^2}} .$$

In this situation the joint density-function of variables X and Y is the product of the density-function of each single variable. This is, on the other hand, a proof of independence between X and Y. Our result was — summing up the above presented ideas — that if the joint distribution of two random variables was a two-dimensional normal distribution and there is no correlation between these variables, then they are independent of each other.

2.2.10. THE LOG-NORMAL DISTRIBUTION

Random variable Y is log-normally distributed if $\ln Y$ has a normal distribution. Let be $X=\ln Y$ and let X be normally distributed with parameters m and σ. In this case, the density function of X is:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} .$$

Based on Eq. (2.17), the density-function of Y is:

$$g(x) = \frac{1}{|x|} f(\ln x) = \frac{1}{\sigma \sqrt{2\pi}} \cdot \frac{1}{x} e^{-\frac{(\ln x-m)^2}{2\sigma^2}} \quad (x > 0). \tag{2.108}$$

Let calculate the mean and standard deviation of random variable Y:

$$E(Y) = \frac{1}{\sigma \sqrt{2\pi}} \int_0^\infty e^{-\frac{(\ln x-m)^2}{2\sigma^2}} dx.$$

By using substitution

$$\frac{\ln x - m}{\sigma} = u$$

then

$$x = e^{\sigma u+m} \quad \text{and} \quad dx = \sigma e^{\sigma u+m} du.$$

Finally,

$$E(Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\sigma u + m} e^{-\frac{u^2}{2}} du = \tag{2.109}$$

$$= e^{m+\frac{\sigma^2}{2}} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{(u-\sigma)^2}{2}} du = e^{m+\frac{\sigma^2}{2}}$$

$$E(Y^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(\ln x - m)^2}{2\sigma^2}} dx = e^{2m+2\sigma^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(u-2\sigma)^2}{2}} du = e^{2m+2\sigma^2}$$

$$D^2(Y) = E(Y^2) - [E(Y)]^2 = e^{2m+2\sigma^2} - e^{2m+\sigma^2} = e^{2m+\sigma^2}(e^{\sigma^2} - 1). \tag{2.110}$$

Log-normal distribution is generally applied in the processes of comminution, cell-division, or disintegration as the distribution of weight, volume or some other dimension of the end-product.

2.2.11. UNIFORM AND RECTANGULAR DISTRIBUTIONS +

In the case of tossing of coins or rolling dice, random variables attached to such experiments will take their possible values with the same probability. This kind of random variables are called variables with uniform distribution. We may distinguish between discrete and continuous uniform distributions. Let X be a discrete random variable with a finite number of possible values $x_1, x_2, \ldots, x_n$ values and let be

$$P(X = x_i) = \frac{1}{n} \qquad (i = 1, 2, \ldots, n). \tag{2.111}$$

Then the random variable X will have uniform distribution on the numbers of $x_1, x_2, \ldots, x_n$. In classical probability theory only this kind of random variables were examined. Discrete uniform distributions are defined only on a finite set of different outcomes. Problems concerning random variables with discrete uniform distribution may be answered by combinatorial methods. (See Section 1.1.4.)

Mean and variance of a random variable with a distribution described in (2.111) are formulated by the following:

$$E(X) = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

which is the arithmetic mean of the possible values.

$$D^2(X) = \frac{1}{n} \sum_{1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{1}^{n} x_i^2 - \bar{x}^2.$$

A continuous random variable X has a rectangular distribution in an interval $[a, b]$ if its density function is the following:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \leqq x < b \\ 0, & \text{otherwise.} \end{cases} \tag{2.112}$$

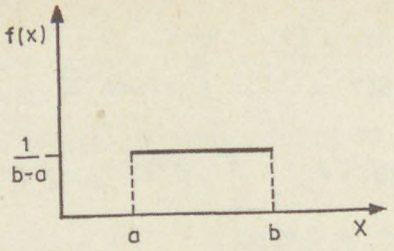


Figure 30

The distribution function of (2.112) is then:

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \dfrac{x-a}{b-a}, & \text{if } a \leqq x < b \\ 1, & \text{if } x \geqq b. \end{cases} \tag{2.113}$$

Mean and variance can also be easily calculated:

$$E(X) = \int_a^b \frac{a}{b-a}\,dx = \frac{1}{b-a}\,\frac{b^2-a^2}{2} = \frac{a+b}{2} \tag{2.114}$$

$$D^2(X) = \int_a^b \frac{x^2}{b-a}\,dx - \left(\frac{a+b^2}{2}\right) = \frac{(b-a)^2}{12}. \tag{2.115}$$

2.2.12. THE GAMMA DISTRIBUTION-FAMILY: GAMMA-, EXPONENTIAL- AND χ^2-DISTRIBUTIONS

Let the density function of a continuous random variable X be:

$$f(x;\,\alpha, p) = \begin{cases} \dfrac{\alpha^p}{\Gamma(p)}\,x^{p-1}e^{-\alpha x}, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.116}$$

Here $\Gamma(p)$ is the so-called Euler's gamma function with parameter p:

$$\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}\,dx. \tag{2.117}$$

It is easy to see that $f(x;\,\alpha : p)$ is really a density function, because

$$\int_0^\infty x^{p-1}e^{-\alpha x}\,dx = \frac{1}{\alpha^p}\int_0^\infty (\alpha x)^{p-1}e^{-\alpha x}\,d(\alpha x) = \frac{\Gamma(p)}{\alpha^p}$$

and, so

$$\int_0^\infty f(x;\,\alpha, p)dx = 1.$$

Random variable X with a density function like (2.116) is eventually called gamma-distributed variable with paramter p.

Let calculate the characteristic function of a gamma distribution with parameters α, p!

$$\varphi(t) = \int_0^\infty e^{itx} f(x;\, \alpha, p)dx = \frac{\alpha^p}{\Gamma(p)} \int_0^\infty x^{p-1} e^{-(\alpha - it)x} dx = \tag{2.118}$$

$$= \frac{1}{\left(1 - \frac{it}{\alpha}\right)^p}.$$

It can be seen that these relationships are valid if α is a complex number but the real part of it must be positive.

By changing the values of parameters α and p in density function $f(x; \alpha, p)$, one may obtain different distributions considered as very important in practical applications. E.g. if $p=1$

$$f(x;\, \alpha, 1) = \alpha e^{-\alpha x} \tag{2.119}$$

which is the well-known exponential distribution. This distribution plays an important role in the hydrology of floods.

Its cumulative distribution function is:

$$F(x) = 1 - e^{-\alpha x}. \tag{2.120}$$

The characteristic function of the exponential distribution may be formulated as:

$$\varphi(t) = \left(1 - \frac{it}{\alpha}\right)^{-1}. \tag{2.121}$$

By aid of the characteristic function one may obtain the mean and variance of this distribution according to (2.67) and (2.68)

$$E(X) = \frac{1}{i}\varphi'(0) = \frac{1}{\alpha} \tag{2.122}$$

$$D^2(X) = [\varphi'(0)]^2 - \varphi''(0) = \frac{1}{\alpha^2}, \quad \text{so} \quad D(X) = \frac{1}{\alpha}. \tag{2.123}$$

It is visible that the mean and standard deviation of an exponentially distributed variable is numerically the same.

If $\alpha = \frac{1}{2}$ and $p = \frac{n}{2}$ are selected as parameters of the density function of (2.116) then the density function of the so-called χ^2-distribution is obtained:

$$k_n(x) = f\left(x;\, \frac{1}{2}, \frac{n}{2}\right) = \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \quad (x > 0). \tag{2.124}$$

This distribution is very important in statistical applications. Squared sums of independent normally distributed random variables are χ^2-distributed. More precisely, if $X_1, X_2, \ldots, X_n$ independent random variables have a standard normal distribution $N(0; 1)$ and if

$$X = X_1^2 + X_2^2 + \ldots + X_n^2$$

then the density function of variable X will take the form of (2.124).

In order to verify this statement let see first the distribution of the addendants. If X is standard normally distributed $N(0; 1)$ and $Y = X^2$ then the distribution function of Y is:

$$G(x) = P(Y < x) = P(X^2 < x) = P(-\sqrt{x} \leqq X < \sqrt{x}) = \tag{2.125}$$

$$= 2\Phi(\sqrt{x}) - 1.$$

The density function is then:

$$g(x) = 2\Phi'(\sqrt{x}) \cdot \frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}} = \tag{2.126}$$

$$= \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} x^{-\frac{1}{2}} e^{-\frac{x}{2}}.$$

So, random variable $Y = X^2$ has a gamma-distribution with parameters $\alpha = \frac{1}{2}$ and $p = \frac{1}{2}$. The characteristic function of variable Y — based on formula (2.118) — may be set then, as:

$$\varphi_Y(t) = \frac{1}{(1-2it)^{\frac{1}{2}}}. \tag{2.127}$$

Due to the fact that the distribution of every random variable $X_i^2 (i = 1, 2, \ldots, n)$ playing a role in the summation of $X = \sum_1^n X_i^2$ is identical with the distribution of Y, the characteristic function of variable X will be:

$$\varphi_X(t) = \prod_{i=1}^{n} \varphi_{X_i^2}(t) = \left[\frac{1}{(1-2it)^{\frac{1}{2}}}\right]^n = \tag{2.128}$$

$$= \frac{1}{(1-2it)^{\frac{n}{2}}}.$$

If this formula is compared with function (2.118) it becomes evident that the latter is the characteristic function of a gamma distribution with parameters

$$\alpha = \frac{1}{2} \quad \text{and} \quad p = \frac{n}{2}.$$

The density function of X is the one given in formula (2.124). The shape of this function is plotted in Fig. 31 for $n=1$, $n=2$ and $n=6$.

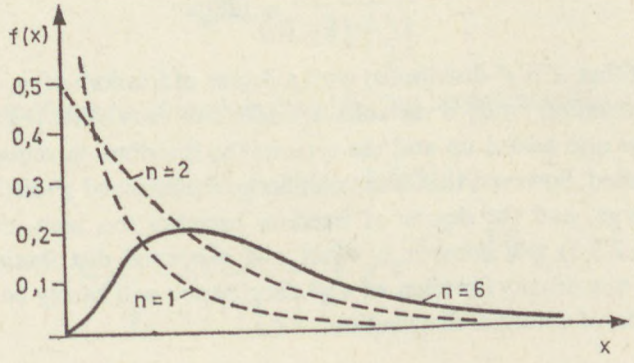


Figure 31

The density function denoted by $k_n(x)$ and formulated by (2.124) which has been gained as the distribution of the squared-sum of n independent $N(0:1)$ normally distributed random variables is sometimes defined as a distribution with n degrees of freedom. The sense of the degree of freedom is explained in Section (6.3.2). By aid of the characteristic function of (2.128) one may easily calculate the moments of χ^2-distributions with n degrees of freedom. When the mean and variance are calculated one has to consider, that

$$\varphi'_x(0) = -\frac{n}{2}\left[(1-2it)^{-\frac{n}{2}-1}(-2i)\right] = in$$

$$\varphi''_x(0) = -\frac{n}{2}\left[-\left(\frac{n}{2}+1\right)\right]\left[(1-2it)^{-\frac{n}{2}-2}\right]_{t=0} =$$

$$= i^2(n^2+2n).$$

So, after considering formulae (2.67) and (2.68)

(2.129) $$E(X) = n,\ D^2(X) = -n^2+n^2+2n = 2n;$$

$$D(X) = \sqrt{2n}.$$

By the use of the characteristic function it is provable that χ^2-distributions are additive. This attributum has the consequence that the sums of two independent χ^2-

distributed variables are again χ^2-distributed. If X_1 has n_1 degrees of freedom and is χ^2-distributed, and X_2 has n_2 degrees of freedom and is also χ^2-distributed and they are independent of each other then the characteristic function of the sum $X=X_1+X_2$ is the product of the two original characteristic functions:

(2.130)
$$\varphi_X(t) = \varphi_{X_1}(t) \cdot \varphi_{X_2}(t) =$$
$$= (1-2it)^{\frac{n_1}{2}} \cdot (1-2it)^{\frac{-n_1}{2}} =$$
$$= (1-2it)^{-\frac{n_1+n_2}{2}}.$$

This is a proof that X is χ^2-distributed with a degree of freedom of n_1+n_2.

So, if χ^2-distributed random variables are added to each other the degrees of freedom should be also added up and the type of the distribution remains unchanged. It should be noted, however, that if the number of the summed χ^2-distributed random variables is large, and the degree of freedom becomes too high, the central limit theorem (see: 2.3.9) will come into effect and a normal distribution is obtained. E.g. if $n=30$, the density function of a χ^2-distribution will hardly be different of the density function of a normal distribution.

2.2.13. THE STUDENT-t DISTRIBUTION

Some important problems in mathematical statistics require the analysis of distributions like:

(2.131)
$$t = \frac{\sqrt{n}X}{\sqrt{X_1^2+X_2^2+\ldots+X_2^2}}$$

where $X_1, X_2, \ldots, X_n$ and X are independent random variables with standard normal distribution, $N(0; 1)$.

To calculate the density function of the random variable t we have to assess first the density function of the numerator, then that of the denominator and finally, the formula defined for the calculation of the density function of the quotient of random variables must be applied.

Let assume that random variable Y is a continuous function of another random variable X:

$$Y = \varphi(X).$$

If the density function of X is denoted by $f(x)$, then the density function of Y is:

$$h(y) = f[\varphi^{-1}(y)] \cdot \left| \frac{d\varphi^{-1}(y)}{dy} \right|.$$

Let now calculate the density function of the numerator of (2.131):

$$Y = \sqrt{n}\,X; \quad \varphi^{-1}(y) = \frac{1}{\sqrt{n}}\,y;$$

so

$$h(x) = \frac{1}{\sqrt{2\pi n}} e^{-\frac{x^2}{2n}}.$$

The density function of the denominator of (2.131) is obtained easily if the density function of a χ^2-distribution with n degrees of freedom is known. The denominator is namely X and X is χ^2-distributed with n degrees of freedom. Thus

$$g(y) = \frac{2y^{n-1}e^{-\frac{y^2}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}.$$

If now the density function of random variable t has been denoted by $s_n(x)$, then:

$$s_n(x) = \int_0^\infty yh(x, y)g(y)dy =$$

$$= \frac{2}{\sqrt{2\pi n}\, 2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^\infty y^n e^{-\left(\frac{x^2}{n}+1\right)\frac{y^2}{2}} dy.$$

By introducing substitution $u=\left(\frac{x^2}{n}+1\right)\frac{y^2}{2}$; we obtain

(2.132)
$$s_n(x) = \frac{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \int_0^\infty u^{\frac{n-1}{2}} e^{-n} du =$$

$$= \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

which is the density function of variable t called Student-t variable.

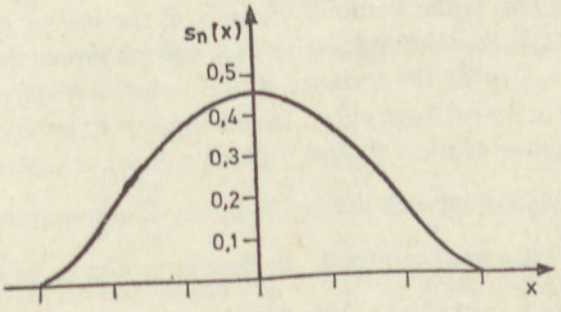


Figure 32

2.3.1. THE BERNOULLI-FORMULA OF THE LAW OF LARGE NUMBERS

In practice the problem often arises that an unknown probability $P(A)=p$ of event A in which we are extremely interested is approximated by aid of the well-known relative frequency $\frac{k}{n}$ in n trials, and we want to know what the difference between this relative frequency and our unknown probability might be. We may be convinced about the smallness of this difference if n is enough large if the law of large numbers formulated by Bernoulli is considered.

Let apply Chebyshev's inequality for the binomial distribution.

In formula (2.45), that is in

$$P(|X-E(X)| > \lambda\sigma) \leqq \frac{1}{\lambda^2}$$

let introduce the following substitutions:

$$X = k,\ E(X) = np \quad \text{and} \quad \sigma = \sqrt{npq},$$

then

$$P(|k-np| > \lambda\sqrt{npq}) \leqq \frac{1}{\lambda^2}.$$

Let divide the inequality in brackets by n:

$$P\left(\left|\frac{k}{n}-p\right| > \lambda\sqrt{\frac{pq}{n}}\right) \leqq \frac{1}{\lambda^2}.$$

If the notation $\lambda\sqrt{\frac{pq}{n}}=\varepsilon$ is introduced, then

$$\frac{1}{\lambda^2} = \frac{qp}{n\varepsilon^2} \leqq \frac{1}{4n\varepsilon^2}$$

and

$$P\left(\left|\frac{k}{n}-p\right| > \varepsilon\right) \leqq \frac{1}{4n\varepsilon^2}. \tag{2.134}$$

Relationship (2.134) is the Bernoulli-formula of the law of large numbers. The right side of (2.134) is approaching zero if $n \to \infty$ which proves that if the number of trials is increased unlimited, the deviation between relative frequency and probability will be small with a desired large safety. In practice, we do not have the opportunity to increase the number of trials without limitation so we are extremely interested to know what the deviation between relative frequency $\frac{k}{n}$ and probability p will be with a large probability if n has been fixed; or how large n must be that this deviation $\left(\frac{k}{n}-p\right)$ be less than a prescribed ε. This question may be answered by using (2.134)

but our n would be larger than needed. The reason for this is explained by a remark at the end of Section 2.3.1.

It was shown earlier that a binomial distribution may be approximated by a normal under conditions presented in Section 2.2.8. From a table below Fig. 34 it is apparent that a normally distributed X random variable will deviate by less than its triple standard deviation from its mean with a fairly large probability, so

$$P(|X-m| < 3\sigma) \approx 0.997.$$

So again, the probability is approximately 0.997 that the deviation of the frequency k of event A from its mean np is smaller than the triple of its standard deviation:

$$P(|k-np| < 3\sqrt{npq}) \approx 0.997. \tag{2.135}$$

In other words, the probability of being $|k-np|$ larger than $3\sqrt{npq}$ is so small that practically it should not be reckoned with. This is the so-called 3σ-rule.

In (2.135) it is allowed to divide both sides by n, so

$$P\left(\left|\frac{k}{n}-p\right| < 3\sqrt{\frac{pq}{n}}\right) \approx 0.997. \tag{2.136}$$

This is an evidence of the fact that the deviation of the relative frequency from our unknown probability is less than $3\sqrt{\frac{pq}{n}}$, with a large probability.

With regard to the fact that the value of $pq=p(1-p)$ cannot exceed $1/4$ it is practically sure, that

$$\left|\frac{k}{n}-p\right| < \frac{3}{2\sqrt{n}} = \frac{1.5}{\sqrt{n}}, \tag{2.137}$$

so it is almost certain that p will fall between

$$\frac{k}{n}-\frac{1.5}{\sqrt{n}} \quad \text{and} \quad \frac{k}{n}+\frac{1.5}{\sqrt{n}}.$$

This error-bound is considered as acceptable if our unknown probability is near to $1/2$. If p is substantially different from $1/2$ then pq is smaller than $1/4$* and the standard deviation of the relative frequency is $\sqrt{\frac{pq}{n}} < \frac{1}{2\sqrt{n}}$.

In such cases we may write the value of relative frequency $\frac{k}{n}$ instead of probability p and $1-\frac{k}{n}=\frac{n-k}{n}$ instead of $q=(1-p)$, so

$$\sqrt{\frac{pq}{n}} \approx \frac{1}{n}\sqrt{\frac{k(n-k)}{n}}.$$

* Function $f(p)=p(1-p)$ has its maximum where $f'(p)=1-2p=0$, and this point is $p=\frac{1}{2}$. Because $f''(p)=-2$, our extreme is really a maximum.

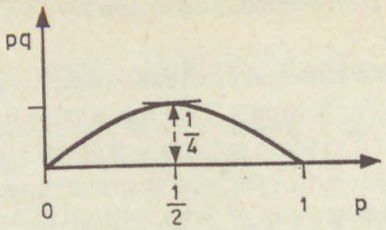


Figure 33

A further improvement of the estimation of the standard deviation may be attained if instead of n, $n-1$ is written in the denominator.

It is practically true, with a very large probability, that

$$\left|\frac{k}{n}-p\right| < \frac{3}{n}\sqrt{\frac{k(n-k)}{n-1}}.$$

If the question must be answered what the value of n, the number of trials should be in order to let relative frequency $\frac{k}{n}$ deviate from our unknown probability by less than e.g. $\varepsilon=0.1$ with a large safety, then based on (2.137) only the following equation must be solved:

$$\frac{1.5}{\sqrt{n}} = 0.1.$$

For n we will obtain 225.

$\varepsilon=0.1$ is not always satisfactory in practice. If accuracy is to be improved, the number of trials will increase. If $\varepsilon=0.05$ is selected, (2.137) will give

$$n = 900$$

and if $\varepsilon=0.01$ (this accuracy is rarely wanted in practice), the number of necessary trials is

$$n = 22\,500.$$

2.3.2. THE CENTRAL LIMIT THEOREM

The central limit theorem is for a theoretical explanation of the fact, why so often the normal distribution is valid while natural processes are investigated. The essence of this theorem is in the finding that if the random fluctuation of a variable X is the resultant of the sum of independent random components exerting only a small influence individually on the fluctuation of these sums, then X is normally distributed.

We saw at the investigation of the binomial distribution that a binomially distributed random variable X is the sum of n independent indicator variables:

$$X = X_1+X_2+\ldots+X_n$$

where the members X_i may take the value of zero or one, independent of each other. It has been shown that a binomial distribution is close to the normal if the value of parameter p is not too small, or not too large, in other words, if the variables do not take in sequence — or predominantly — one of their possible two values. (We may say, if the distribution of X_i is not too degraded).

Approximation of the binomial by normal distribution is a special case, although a very important one, of a far more general regularity which may be defined as it follows: if a very large number of independent random variables are summed up and the individual components do have finite variances then their sum will be normally distributed without any regard on the distribution of the individual components.

The central limit theorem may be defined mathematically in several ways. Due to the fact that in mathematical statistics we are faced usually with independent, identically distributed components, we will discuss first the case in which the independent components have the same distribution.

Theorem: If $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with finite standard deviation, e.g. $E(X_k)=m$, $D^2(X_k)=\sigma^2 (k=1, 2, \ldots, n)$ then

$$\lim_{n\to\infty} P\left(\frac{X_1+X_2+\ldots+X_n-nm}{\sigma\sqrt{n}} < x\right) = \tag{2.138}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du.$$

Proof: Let denote the characteristic function of random variable X_k-m by $\varphi(t)$. Then the characteristic function of variable $\dfrac{X_k-m}{\sigma\sqrt{n}}$ is:

$$\varphi\left(\frac{t}{\sigma\sqrt{n}}\right).$$

Let be:

$$Y = \frac{X_1+X_2+\ldots+X_n}{\sigma\sqrt{n}} = \sum_{k=1}^{n} \frac{X_k-m}{\sigma\sqrt{n}}.$$

Because of independence of the components

$$\varphi_Y(t) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

Let introduce the notation $\dfrac{t}{\sigma\sqrt{n}} = u$ and let expand function $\varphi(u)$ in Taylor-series:

$$\varphi(u) = \varphi(0) + \frac{\varphi'(0)}{1!}u + \frac{\varphi''(0)}{2!}u^2 + \ldots.$$

The derivatives of the characteristic function at $u=0$ may be expressed by the moments according to (2.70):

$$\varphi(u) = 1+\alpha_1 iu+\alpha_2 \frac{(iu)^2}{2!}+\ldots+$$

$$+\alpha_k \frac{(iu)^k}{k!}+o(u^k),$$

where α_j is the jth moment of variable $\dfrac{X_k-m}{\sigma\sqrt{n}}$.

$$\alpha_1 = E\left(\frac{X_k-m}{\sigma\sqrt{n}}\right) = 0$$

$$\alpha_2 = E\left(\frac{X_k-m}{\sigma\sqrt{n}}\right)^2 = \frac{1}{n\sigma^2}E(X_k-m)^2 = \frac{1}{n}.$$

Based on this:

$$\varphi\left(\frac{t}{\sigma\sqrt{n}}\right) = 1-\frac{t^2}{2n}+0\left(\frac{t^2}{n\sigma^2}\right) = 1-\frac{t^2}{2n}+0\left(\frac{1}{n}\right).$$

And finally:

(2.139)
$$\lim_{n\to\infty}\left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \lim_{n\to\infty}\left[1-\frac{t^2}{2n}+\right.$$

$$\left.+0\left(\frac{1}{n}\right)\right]^n = e^{-\frac{t^2}{2}}.$$

As it was discussed with the normal distribution this function is exactly the characteristic function of the standard normal distribution.

A next formulation of the theorem does not stipulate that the components should have the same distribution but would postulate the existence of their third absolute moments.

The theorem of Laplace and Liapunov: if $X_1, X_2, \ldots, X_n$ are independent random variables having third absolute moment, if $E(X_k)=0$ $(k=1, 2, 1\ldots, n)$ and further if:

$$\lim_{n\to\infty}\frac{\sum_1^n \beta_k}{S_n^3} = 0$$

where

$$\beta_k = E(|X_k|^3),\ S_n = \sqrt{\sum_1^n \sigma_k^2};\ \sigma_k^2 = D^2(X_k)$$

then

(2.140)
$$\lim_{n\to\infty} P\left(\frac{X_1+X_2+\ldots+X_n}{S_n} x\right) =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{u^2}{2}}\,du.$$

The central limit theorem is of great practical importance. On the one hand, if the conditions of the theorem are fulfilled — at least with acceptable approximation — one may assume that the standardized variable is normally distributed which can be checked easily by the methods of mathematical statistics. On the other hand, it is well-known that statistical functions used usually in statistical analysis consist in a very large number of the sums of independent random variables being usually normally distributed individually due to the large number of observations upon which they are based. As a consequence, the characteristics of the normal distribution and the function-tables of the normal distribution are important aids in statistical decision-making.

CHAPTER 3

Markov-chains. Markov-processes.

3.1. MARKOV-CHAINS

3.1.1. THE NOTION OF THE MARKOV-CHAIN, EXAMPLES FOR MARKOV-CHAINS

In the previous chapters sequences of experiments were discussed where the outcome of experiments are independent. In hydrology another type is most often used where the outcome of the experiments is more or less dependent on the results of previous ones. Hydrological observations are usually recorded in the form of time series measured in equal time intervals. It is trivial that e.g. daily water stages are not independent from each other. Even the weekly or the monthly mean discharges display some dependence. In this chapter sequences of experiments will be investigated with certain stochastic relations in their consecutive outputs. Let us first investigate the simplest type of such models when the $(N+1)$th trial depends on the Nth trial but without direct dependence on trials $(N-1)$, $(N-2)$, ... (Indirect dependence means dependence through the outcome of the Nth trial.) More precisely: let $A_1, A_2, \ldots, A_n$ be a complete system of events. The Nth trial is characterized by the random variable X_N, if the result of the Nth experiment is A_j, then $X_N=j$. Such an example is: the water stages of a river are measured at a gauging station at time points $t=0, 1, 2, \ldots$. If the state is denoted by X, then event A_j means $A_j: \{a_j \leqq X < a_{j+1}\} (j=0, 1, 2, \ldots, n)$, where a_0 and a_{n+1} are the possible smallest and highest observable stages. After a while a time series $X_0, X_1, X_2, \ldots$ of random variable X will be at our disposal. Event $\{X_i=j\}$ will occur if in the time point i the measured stage falls into the interval (a_j, a_{j+1}).

Another example is the changes of the stored amount of water in a reservoir which may be characterized in a similar way if our observations are carried out in discrete time points.

In general the values of the random variable $X_0, X_1, X_2, \ldots$ should be considered a sequence of states of a physical system.

Let us now consider a physical system with a finite number of possible states. At the beginning let our system be in state X_0 and after N steps in state X_N (the possible values of X_N are $0, 1, 2, \ldots, n$).

Transitions $X_0 \to X_1 \to X_2 \to \ldots$ are of random nature and let us suppose that they are satisfying the following rule: if the system is in state $X_N=i$ at the Nth step then it should reach state $X_{N+1}=j$ with a probability p_{ij} independently of its earlier states.

$$p_{ij} = P(X_{N+1} = j \,|\, X_N = i) \quad (i, j = 0, 1, 2, \ldots, n), \tag{3.1}$$

of course, as one of the A_j's certainly will occur. The sequence $\{X_N\}$ is called here a homogeneous Markov-chain, and the conditional probabilities p_{ij} are called transition-probabilities.

Beyond the transition-probabilities the so-called initial distribution i.e. the distribution of X_0 is of importance if a Markov-chain is to be determined:

$$p_i^{(0)} = P(X_0 = i) \quad (i = 0, 1, 2, \ldots, n). \tag{3.2}$$

Let us denote by $p_k^{(N-1)} = P(X_{N-1} = k)$ the probability that our system is in state k after $N-1$ steps. The probability that the system is in state j after N steps — supposed that it has reached state k after $N-1$ steps — is: p_{kj}. By virtue of the total probability rule:

$$P(X_N = j) = \sum_k P(X_N = j \,|\, X_{N-1} = k) P(X_{N-1} = k). \tag{3.3}$$

Equation (3.3) leads to the following recursive relation for the probabilities $p_j^{(N)}$:

$$p_j^{(0)} = p_j^{(0)}; \; p_j^{(m)} = \sum_k p_k^{(m-1)} p_{kj} \quad (m = 1, 2, \ldots). \tag{3.4}$$

If at the beginning the system was in state i with probability 1 then the initial distribution is : $p_i^{(0)} = 1, p_k^{(0)} = 0$ if $k \neq i$. In this case probability $p_j^{(N)}$ is identical with the transition-probability $p_{ij}^{(N)}$ which stands for the probability of finding the system in state j after N steps if the initial state was i:

$$p_{ij}^{(N)} = P(X_N = j \,|\, X_0 = i) \quad (j = 0, 1, 2, \ldots, n). \tag{3.5}$$

In case of this special initial distribution, equation (3.4) will lead to the following recursion for the calculation of the transition-probabilities:

$$p_{ij}^{(0)} \begin{cases} 1 & \text{if} \quad j = i \\ 0 & \text{if} \quad j \neq i; \end{cases} \tag{3.6}$$

$$p_{ij}^{(N)} = \sum_{kj} p_{ik}^{(N-1)} p_{kj} \quad (N = 1, 2, \ldots).$$

The probabilities $p_{ij}^{(m)}$ may be arranged in a matrix. (Assume that $i, j = 0, 1, 2, \ldots, n$)

Matrix $\mathbf{P}^{(0)} = [p_{ij}^{(0)}]$ is obviously a unitmatrix:

$$\mathbf{P}^{(0)} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} = \mathbf{E} = \langle 1, \ldots, 1, 1 \rangle.$$

$$\mathbf{P}^{(1)} = [p_{ij}] = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0n} \\ p_{10} & p_{11} & \cdots & p_{1n} \\ p_{n0} & p_{n1} & \cdots & p_{nn} \end{bmatrix} = \mathbf{P} \tag{3.7}$$

which is the matrix of one-step transition-probabilities. Based on relationship (3.6) $\mathbf{P}^{(2)}=[p_{ij}^{(2)}=\sum_k p_{ik}p_{kj}=\mathbf{P}^2$. The elements of matrix $\mathbf{P}^{(2)}$ are calculated in the following way: elements of the ith row of matrix $\mathbf{P}$ are composed by elements of the jth column of the same matrix which corresponds to the multiplication of matrices. Similarly,

$$\begin{aligned}\mathbf{P}^{(3)} &= \sum_k p_{ik}^{(2)} p_{kj} = \mathbf{P}^2 \cdot \mathbf{P} = \mathbf{P}^3 \\ &\vdots \\ \mathbf{P}^{(N)} &= \sum_k p_{ik}^{(N-1)} p_{kj} = \mathbf{P}^{N-1} \cdot \mathbf{P} = \mathbf{P}^N.\end{aligned} \tag{3.8}$$

Obviously, the matrix of the m-step transition-probabilities equals the m-th power of the matrix of the one-step transition-probabilities. So, the determination of m-step transition-probability matrices of Markov-chains is a problem of calculating the powers of a given matrix which often needs computer-aid.

It should be noted that the one-step transition-probability matrix $\mathbf{P}$ given in (3.6) is a so-called stochastic matrix having only non-negative elements with a sum of 1 in every row. The product of two stochastic matrices is also a stochastic matrix. It follows that $\mathbf{P}^N$ is a stochastic matrix for every N as well. On the basis of (3.4) the N-step absolute probabilities may be expressed with the aid of the initial distribution vector $p^{(0)}=[p_0^{(0)}, p_1^{(0)}, \ldots, p_n^{(0)}]$ and of the one-step transition-probability matrix $\mathbf{P}=[p_{ij}]$ in the following way:

$$p_N^* = [p_0^{(N)}, p_1^{(N)}, \ldots, p_n^{(N)}] =$$

$$= [p_0^{(0)}, p_1^{(0)}, \ldots, p_n^{(0)}] \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0n} \\ p_{10} & p_{11} & \cdots & p_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n0} & p_{n1} & \cdots & p_{nn} \end{bmatrix} = {}_0{}^* \cdot \mathbf{P}^N.$$

3.1.1. RANDOM WALK BETWEEN ABSORBING BARRIERS

Our problem should be confined to a random walk over the integers of the real line. A particle should begin its motion from a point i for which $0<i<n$. Let us suppose that in points $x=0$ and $x=n$ the particle will be absorbed, in other words, absorbing barriers are present. This means that if the particle reaches one of these during its random motion, it stops and remains at that point. Let be $p_{00}=1$, $p_{nn}=1$, $p_{ii}=0$ $(i=1, 2, \ldots, n-1)$ and

$$p_{ij} = \begin{cases} p & \text{if } j = i+1 \\ q & \text{if } j = i-1 \\ 0 & \text{otherwise.} \end{cases}$$

The matrix of the one-step transition-probabilities is:

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ q & 0 & p & 0 & \dots & 0 \\ 0 & q & 0 & p & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & & 1 \end{bmatrix},$$

being an $(n+1)\times(n+1)$ matrix.

If $p=q=\frac{1}{2}$, the random walk is called symmetric. Let us define the N-step transition-probabilities!

It is easy to see that by certain changes in the rows and the columns our matrix $\mathbf{P}_1$ may be brought into the following form:

$$\mathbf{P}_1 = \begin{bmatrix} \mathbf{E} & \mathbf{O} \\ \mathbf{B}_1 & \mathbf{P} \end{bmatrix}$$

where $\mathbf{P}$ is the following matrix of order $n-1$:

$$\mathbf{P} = \begin{bmatrix} 0 & p & 0 & 0 & \dots & 0 & 0 \\ q & 0 & p & 0 & \dots & 0 & 0 \\ 0 & q & 0 & p & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & q & 0 \end{bmatrix}, \quad E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

It is easy to see that

$$\mathbf{P}_1^N = \begin{bmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{B}_N & \mathbf{P}^N \end{bmatrix}. \tag{3.9}$$

In connection with this random walk the following problems are of practical importance

a) What is the probability that the particle will be absorbed in N steps by wall $x=0$?

b) What is the probability that the particle will be absorbed in N steps by wall $x=n$?

c) What is the probability that the particle will return to its starting point in $2N$ steps?

d) What is the probability that the particle will not be absorbed during N steps?

e) What is the mean number of steps till absorption?

We turn now to the answer of these questions:

Question a) will be answered if we determine the probability that the particle, starting from point $x=i$, reaches point $x=1$ after $N-1$ steps (without earlier absorption; this probability is given by the elements of matrix $\mathbf{P}^{N-1}$) and this probability is then multiplied by q:

$$p_{i0}^{(N)} = q p_{i1}^{(N-1)}.$$

Answer to question b) is similar:

$$p_{in}^{(N)} = p \cdot p_{i,n-1}^{(N-1)}.$$

So, we need the $(N-1)$th power of matrix $\mathbf{P}$; it can be obtained without the effective performance of this operation. Namely the following factorization of matrix $\mathbf{P}$ holds:

$$\text{(3.10)} \qquad \mathbf{P} = \begin{bmatrix} 0 & p & 0 & 0 & \dots & 0 & 0 \\ q & 0 & p & 0 & \dots & 0 & 0 \\ 0 & q & 0 & p & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 0 & p \\ 0 & \dots & \dots & \dots & \dots & q & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & \\ 0 & \sqrt{\frac{q}{p}} & 0 & \dots & \\ 0 & 0 & \left(\sqrt{\frac{q}{p}}\right)^2 & \dots & \\ & & & & \left(\sqrt{\frac{q}{p}}\right)^{n-2} \end{bmatrix} \cdot$$

$$\cdot \begin{bmatrix} 0 & \sqrt{pq} & 0 & \dots & \dots \\ \sqrt{pq} & 0 & \sqrt{pq} & \dots & \dots \\ 0 & \sqrt{pq} & \cdot & \sqrt{pq} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \sqrt{pq} & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & \sqrt{\frac{p}{q}} & 0 & \dots & \dots \\ 0 & 0 & \left(\sqrt{\frac{p}{q}}\right)^2 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \left(\sqrt{\frac{p}{q}}\right)^{n-2} \end{bmatrix} = \mathbf{T}\mathbf{\Pi}_1\mathbf{T}^{-1}.$$

$$\text{(3.11)} \qquad \text{where } \mathbf{\Pi}_1 = \sqrt{pq} \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = \sqrt{pq}\,\mathbf{\Pi}.$$

The canonic decomposition of matrix $\mathbf{\Pi}$ is well-known (see e.g. A. 18). Its eigenvalues and eigenvectors are given by the following formulae:

$$\text{Eigenvalues: } \lambda_k = 2\cos\frac{k\pi}{n} \quad (k = 1, 2, \dots, n-1).$$

$$\text{Eigenvectors: } u_k^* = \sqrt{\frac{2}{n}}\left[\sin\frac{k\pi}{n}, \dots, \sin\frac{(n-1)k\pi}{n}\right].$$

$$\text{Consequently: } \mathbf{\Pi} = \sum_{k=1}^{n-1} \lambda_k\, u_k\, u_k^*.$$

It is known that in this case: $\Pi^N = \sum_k \lambda_k^N u_k u_k^*$. Based on (3.9) and (3.10):

(3.12)
$$\mathbf{P}^N = T \cdot \sqrt{pq}\, \Pi^N T^{-1} =$$
$$= \sqrt{pq}\, \mathbf{T}\left(\sum_k \lambda_k^N u_k u_k^*\right) T^{-1}.$$

To answer questions a) and b) the elements of matrix $\mathbf{P}^{N-1}$ are needed:

(3.13)
$$p_{ij}^{(N-1)} = \frac{2}{n} \sum_{k=1}^{n-1} \left(2\sqrt{pq} \cos\frac{k\pi}{n}\right)^{N-1} \left(\sqrt{\frac{q}{p}}\right)^{i-1}$$
$$\sin\frac{ik\pi}{n} \left(\sqrt{\frac{p}{q}}\right)^{j-1} \sin\frac{jk\pi}{n} =$$
$$= \frac{2}{n} p^{\frac{N-1-i+j}{2}} q^{\frac{N-1+i-j}{2}} \sum_{k=1}^{n-1} \cos^{N-1}\frac{k\pi}{n}$$
$$\sin\frac{ik\pi}{n} \sin\frac{jk\pi}{n}.$$

Finally, the answer to question a) is:

(3.14)
$$p_{i0}^{(N)} = q p_{i,1}^{(N-1)} = \frac{2^N}{n} p^{\frac{N-i}{2}} q^{\frac{N+i}{2}} \sum_{k=1}^{n-1}$$
$$\cos^{N-1}\frac{k\pi}{n} \sin\frac{ik\pi}{n} \sin\frac{jk\pi}{n}.$$

Based on this, the reader may answer easily question b).
(It should be mentioned that (3.13) was first obtained by Feller using the method of generating-functions, see: A. 6.)

If question c) is to be answered then the probability that the particle will return to its starting point $x=i$ after $2N$ steps (without being absorbed) is to be determined. This is the following:

(3.15)
$$p_{ii}^{2N} = \frac{2^{2N}}{n} p^N q^N \sum_{k=1}^{n-1} \cos^{2N}\frac{k\pi}{n} \sin^2\frac{ik\pi}{n}.$$

In connection with question d), the probability that after N steps our particle is still not absorbed is obviously equal to the probability that the particle started its motion in point $x=1$ then reached some point $x=1$, $k=2$, or ... $x=n-1$ after N steps. Due to the fact that these events are mutually exclusive, the probability is exactly the sum of the ith row of matrix $\mathbf{P}^N$.

(3.16)
$$P(X > N) = \frac{2^{N+1}}{n} p^{\frac{N-i}{2}} q^{\frac{N+i}{2}} \sum_{k=1}^{n-1} \cos^N\frac{k\pi}{n} \sin\frac{ik\pi}{n} \left[\sum_{j=1}^{n-1} \left(\sqrt{\frac{p}{q}}\right)^j \sin\frac{jk\pi}{n}\right],$$

where random variable X stands for the number of steps before absorption. If in the above formula $p=q=\frac{1}{2}$ is substituted the relationship presented by Loéve (see: A. 18) is obtained:

$$P(X > N) = \frac{2}{N} \sum_{k'=1}^{n-1} \cos^N \frac{k\pi}{n} \sin \frac{ik\pi}{n} \operatorname{ctg} \frac{k\pi}{2n}, \tag{3.16'}$$

where k' is for the rule that summation should be carried out only for odd k-s. Finally, if question c) is answered, the mean of random variable X denoting the number of steps before absorption, will be:

$$E(X) = \sum_{N=0}^{\infty} NP(X = N)$$

where

$$P(X = N) = P(X > N-1) - P(X > N).$$

By aid of (3.16), $P(X=N)$ can be easily determined. In case of symmetric random walk this probability is obtained from (3.16'). The mean of the number of steps before absorption is in this case:

$$E(X) = \frac{1}{n} \sum_{k=1}^{n-1} - \frac{\sin \frac{ik\pi}{n} \cos \frac{k\pi}{n}}{\sin^3 \frac{k\pi}{2n}}. \tag{3.17}$$

The reader's attention is called to the fact that (3.14) is used in a method to test homogeneity (see Section 6.4.3).

Random walk between absorbing walls may be discussed in a similar way. Also, by aid of relatively simple matrix-theoretical methods the problem of random walk among the lattice-points of an n-dimensional space may be investigated between absorbing or reflective barriers (see e.g. B. 29).

Although the above presented, bounded random walk model may be successfully used in a series of practical problems, it is often the case that a more general model would be needed allowing for the particle not only to step onto the neighbouring lattice-point but also to jump onto others. Such a more general model is needed for the description e.g. of the sequence of changes of the states of a reservoir.

3.1.3. THE PROBABILITY OF EMPTYING AND THE OVERSPILL OF A RESERVOIR

Let be the possible states of a reservoir (by using appropriate coding), 0, 1, 2, ..., k. Assure, that the initial state was $X=z$. If now, during a properly selected time interval an amount of l m^3 will enter the reservoir its state will be $z+l$. If a m^3 water is withdrawn from the same reservoir its state will be $z+l-a$. If $l-a>0$ or $l-a<0$ the reservoir will step in a larger or smaller state than z.

Let now assume that a particle is pursuing random walk in a $[0, k]$ interval of the line where absorbing walls had been placed in points $x=0$ and $x=k$. Our particle will take steps to the left or to the right starting from point $x=z$. The length of its steps is the difference between inflow and release, $l-a$, where l is a random variable (inflow in m^3/time unit) and a is a fixed number. (The possible values of $l-a$ are: $-a, -a+1, \ldots, -1, 0, 1, 2, \ldots$.) The state of the reservoir is characteristic to the amount of stored water (in integer m^3-s), the location of the particle is — on the other hand — the state of the reservoir. The absorbing walls in points $x=0$ and $x=k$ indicate the following extremities: if the particle reaches or surpasses point $x=0$ the reservoir emptied, if it reaches or surpasses point $x=k$, overspill has occurred. Naturally if an absorbing point has been reached or jumped over, motion of the particle ceases.

Assume that the lengths of the individual steps are represented by random variables $X_1, X_2, \ldots$. Let assume that the individual variables X_i are independent and uniformly distributed:

$$P(X_i = h) = p_h \quad \text{for every} \quad i, h = \ldots, -3, -2, -1, 0, 1, 2 \quad (h = l-a).$$

(Assume that this distribution — which is basically the distribution of inflow — is known). Motion of the particle will be terminated if

$$X_1+X_2+\ldots+X_N < z \quad \text{or} \quad X_1+X_2+\ldots+X_N \geqq k-z$$

in any N time point.

The above investigated generalized random walk is basically a Markov-chain with a one-step transition-probability matrix represented by the following stochastic matrix:

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 \\ r_1 & p_0 & p_1 & p_2 & \cdots & p_{h-2} & q_1 \\ r_2 & p_{-1} & p_0 & p_1 & \cdots & p_{h-3} & q_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{h-1} & p_{h+2} & \cdots & \cdots & \cdots & p_0 & q_{h-1} \\ 0 & 0 & \ldots & \ldots & \ldots & 0 & 1 \end{bmatrix}$$

where $r_h = p_h + p_{h+1} + \ldots q_h = p_{k-h} + p_{k-h+1} + \ldots$ N-step transition-probabilities can be computed by matrix-involution but this operation needs usually computer-help in case of larger matrices. It will be shown that the probability of emptying (absorption at $x=0$) or overspilling (absorption at $x=k$) of a reservoir may be approximated with simple considerations by Feller's method (see: A. 6).

The probability that a particle will move from a point z in point x in one step, is: p_{x-z}. The probability that a particle being in point x will be absorbed in point zero is denoted by: u_x.

The probability that a particle starting from point z is absorbed in one step at $x=0$, is: r_z. Obviously,

$$r_z = p_{-z} + p_{-z-1} + p_{-z-2} + \ldots$$

(r_z may take the value zero, too).

After these, the probability of emptying of our reservoir (initial state: z) will be:

$$u_z = \sum_{x=1}^{k-1} u_x p_{x-z} + r_z \quad (z = 1, 2, \ldots, k-1). \tag{3.18}$$

In this way we will have $(k-1)$ linear equations and $(k-1)$ unknowns, and $(z=1, 2, \ldots, k-1)$. This system of equations is not homogeneous, because if motion into negative direction is possible at all (less inflow than intake), the following unequality must hold: $r_1>0$.

It is a requirement that adequately homogeneous systems of equations, like

$$u_z = \sum_{x=1}^{k-1} u_x p_{x-z} \tag{3.19}$$

should have only trivial solutions. If a nontrivial solution of (3.19) was available then one of the values $u_z(z=1, 2, \ldots, k-1)$ would be the largest absolute value.

Let be $u_z=M>0$. Assume that $p_{-1}\neq 0$. Because the sum of coefficients p_{x-z} in (3.18) is at best 1, equality in (3.18)

$$M = u_1 p_{1-z} + u_2 p_{2-z} + \cdots$$

is possible only if the coefficients of the probabilities p_{x-z} other than zero are M and their sum is 1. Assumption $p_{-1}\neq 0$ immediately leads to $u_{z-1}=M$. With a similar logic $u_{z-2}=\ldots u_1=M$. If $z=1$, then

$$M = u_1 - \sum_{x=1}^{k-1} u_x p_{x-1} = u_1 p_0 + u_0 p_1 + \ldots +$$

$$+ u_{k-1} p_{k-z} = M \sum_{i=0}^{k-2} p_i.$$

Because $p_1\neq 0$ and $\sum_{i-0}^{k-2} p_i<1$, consequently $M=0$. Similar logic is used if $p_{-1}=0$ but another probability $p_h(h<0)$ is different from zero. This leads to unique solution of (3.18). Let introduce the following boundary conditions:

$$\begin{aligned} u_x &= 1 \quad \text{if} \quad x \leqq 0 \\ u_x &= 0 \quad \text{if} \quad x \geqq k. \end{aligned} \tag{3.20}$$

(3.18) can be written, in this case, in the following from:

$$u_z = \sum_{x=-\infty}^{\infty} u_x p_{x-z}. \tag{3.21}$$

If k is large, the direct solution of $k-1$ equations would be tedious. Another approach is here used which is extremely advantageous if the distribution $\{p_k\}$ has relatively few positive components.

Assume that positive probabilities p_h may be found only if $-\nu \leqq h \leqq \mu$. This involves that the possible largest positive jump is ν and the possible largest negative jump is μ. One may assume without restricting generality that the mean of distribution $\{p_k\}$ is zero. So,

$$\sum h p_h = 0.$$

In this case $A+Bz$ is a formal solution to equation (3.21), namely:

$$u_z = \sum_x (A+Bx)p_{x-z} = A\sum p_h + B(z+h)p_h =$$

$$= A\sum p_h + Bz\sum p_h + B\sum h p_h = A+Bz.$$

If A and B are selected so that

$$A+Bz = 0 \quad \text{should hold if} \quad z = k+\mu-1$$

and $A+Bz=1$ if $z=0$
then:

$$A+Bx \geqq 1 \quad \text{if} \quad x \leqq 0, \quad A+Bz \geqq 0 \quad \text{if} \quad k \leqq x \leqq k+\mu.$$

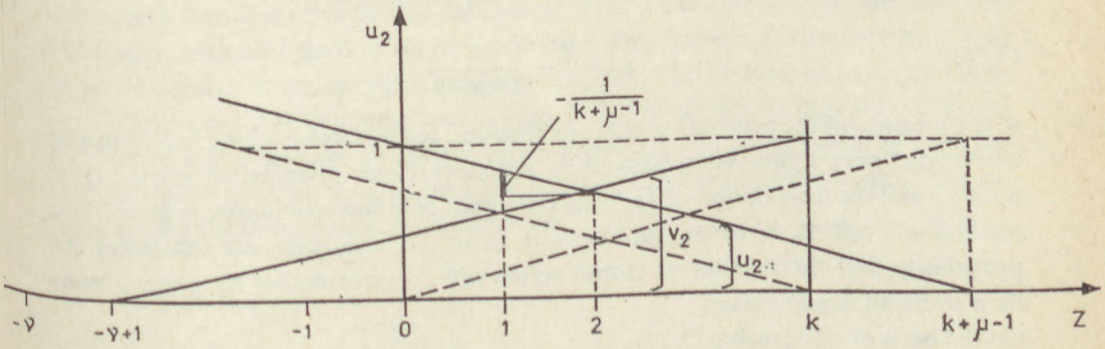


Figure 34

Consequently, $A+Bz$ will satisfy the following boundary conditions:

(3.22)
$$u_x \geqq 1 \quad \text{if} \quad x \leqq 0$$

$$u_x \geqq 0 \quad \text{if} \quad x \leqq k.$$

Based on the conditions of (3.22):

$$A+B(k+\mu-1) = 0$$

$$A = 1 \quad \text{and} \quad B = -\frac{1}{k+\mu-1}, \text{ finally:}$$

$$u_z = 1 - \frac{z}{k+\mu-1}.$$

If, however, A and B are selected as to have

$$A+Bz=0 \quad \text{if} \quad z=k$$

$$A+Bz=1 \quad \text{if} \quad z=-v+1$$

then

$$A+Bk=0; \quad Bk=-A$$

$$A+B(-v+1)=1; \quad -Bk-Bv+B=1; \quad B=\frac{-1}{k+v-1};$$

$$A=\frac{k}{k+v-1}$$

$$u_z \equiv A+Bz=\frac{k-z}{k+v-1}.$$

Based on these, the following limits can be assessed for the probability of u_z:

$$\frac{k-z}{k+v-1} \leqq u_z \leqq 1-\frac{z}{k+\mu-1}. \tag{3.23}$$

For the sake of safety the upper limit is used if u_z is estimated:

$$u_z \approx 1-\frac{z}{k+\mu-1} \tag{3.24}$$

u_z is a measure of the probability that our reservoir reaches its zero state (which can be determined only somewhat arbitrarily in practice) if our investigation started from a state z and the distribution of $\{p_h\}$, the difference of inflow and intake, is given. This method will tell us, however, substantially more. We may calculate (estimate) the probability that the amount of stored water will decrease earlier by given λ units than overspill would occur.

On the basis of relationship (3.24):

$$u_{z-\lambda} \approx 1-\frac{z-\lambda}{k+\mu}=u_z+\frac{\lambda}{k+\mu}. \tag{3.25}$$

By similar considerations one may have the (approximate) probability that the reservoir will sooner overspill than being emptied if $\{p_h\}$ is the given probability distribution with an assumed expectation of zero. Let denote this probability by v_z. Then, its estimation will be:

$$v_z \approx \frac{1}{k+v} z=\frac{v-1}{k+v}. \tag{3.26}$$

(The appropriate geometrical construction is shown in Fig. 34.)

Also, it can be calculated what the initial state z should be from which overspill may occur by a larger probability than emptying if the probability distribution was $\{p_h\}$. This will be the case, if

$$z>\frac{(k+1)(k+\mu)}{2k+\mu+v}.$$

As this was mentioned at the definition of a Markov-chain, the different states of a Markov-chain are not independent of each other. Future states are functions of the present state, which in turn, is a function of past states. The question was still not investigated how the different states are stochastically dependent on each other. It was earlier mentioned that in hydrology Markov-chains are usually used to model time series. The task of a hydrologist is generally the following: there is a $X_1, X_2, \ldots, \ldots, X_n$ realization of a time series based on observations in regular time intervals and the problem is how to approximate the elements (components) of this time series by a Markov-chain. Statistical analyses are usually started by a categorization of the values of random variable $X_1, X_2, \ldots$ in certain classes $1, 2, \ldots, n$. Then the statistics of the frequency, or relative frequency of the transitions $i \to j$ $(i, j = 1, 2, \ldots, n)$ are prepared and placed in an $n \times n$ size matrix. If the series of information is enough long, this matrix will be the approximation of the one-step transition-probability matrix $\mathbf{P} = [p_{ij}]$. If the rows of matrix $\mathbf{P}$ are near to identical, in other words, if in a column the same numbers are discovered (the columns, of course, may differ from each other), then it is obvious that the time series reaches a given state with the same probability, with no regard on its previous state. This implies, at the same time, that the series consists of independent random variables. The time series is, in this case, not a Markov-chain but an independent sequence of observations. The more different the rows of a $\mathbf{P}$ transition-matrix the more dependent the components of our time series, in case of Markov-chains the subsequent states of the Markov-chain.* Dependence among the individual states is strongest, if

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ . & . & . & \ldots & . \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} = \mathbf{E}.$$

Here, if the Markov-chain is in state i, then it will remain there which is symbolized by $X_{k+1} = X_k$ $(k = 1, 2, \ldots)$. The matrix of the one-step transition-probabilities is:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & 0 & \ldots & 0 \\ . & . & . & . & . & \ldots & . \\ 0 & 0 & 0 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 0 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

which is a so-called cyclic-matrix, and the series of the states of the chain is:

$$0 \to 1 \to 2 \to \ldots \to n \to 1.$$

* (In this case the numbers in a column are also different compared to each other.)

It should be noted that, in this case, $\mathbf{P}$ is double-stochastic. ($\mathbf{E}$ is also double-stochastic). Markov-chains applied in practice usually allow transition from a given state into more possible states. Let now consider a finite Markov-chain with possible states of $1, 2, \ldots, m$. Let assume that its one-step transition-probability matrix $\mathbf{P}=[p_{ij}]$ may have an exponent — say N — for which

$$\mathbf{P}^N = [p_{ij}^{(N)}]$$

and for which its elements are positive, say

$$\min_{i,j} p_{ij}^{(N)} = \delta > 0.$$

This assures that the chain may move from any state i into any state j with a positive probability (with another terminology: state j can be attained from state i in N steps, $i, j = 1, 2, \ldots, m$). If now the transition-probability matrix is future involuted, matrix $\mathbf{P}^{N+1}$ will be obtained:

$$\mathbf{P}^{N+1} = \begin{bmatrix} p_{11} & p_{12} \cdots p_{1m} \\ p_{21} & p_{22} \cdots p_{2m} \\ \cdot & \cdot \quad \cdots \cdot \\ p_{m1} & p_{m2} \cdots p_{mm} \end{bmatrix} \begin{bmatrix} p_{11}^{(N)} & p_{12}^{(N)} \cdots p_{1m}^{(N)} \\ p_{21}^{(N)} & p_{22}^{(N)} \cdots p_{2m}^{(N)} \\ \cdot & \cdot \quad \cdots \cdot \\ p_{m1}^{(N)} & p_{m2}^{(N)} \cdots p_{mm}^{(N)} \end{bmatrix} = \begin{bmatrix} p_{11}^{(N+1)} & p_{12}^{(N+1)} \cdots p_{1m}^{(N+1)} \\ p_{21}^{(N+1)} & p_{22}^{(N+1)} \cdots p_{2m}^{(N+1)} \\ \cdot & \cdot \quad \cdots \cdot \\ p_{m1}^{(N+1)} & p_{m2}^{(N+1)} \cdots p_{mm}^{(N+1)} \end{bmatrix}$$

where $p_{ij}^{(N+1)} = \sum_k p_{ik} p_{kj}^{(N+1)}$ is the weighted arithmetic mean of the elements of column j of matrix $\mathbf{P}^N$ and the weights are the elements of row i of matrix $\mathbf{P}$. The convex arithmetic mean of m positive members must be in between the largest and smallest values (even if more positive weights were applied), so

$$\min_i p_{ij}^{(N)} \leqq \min_i p_{ij}^{(N+1)} \leqq \max_i p_{ij}^{(N+1)} \leqq \max_i p_{ij}^{(N)} \quad (j = 1, 2, \ldots, m).$$

This, however, will lead to an ever decreasing difference between the largest and smallest values with an increase of n (and assuming that more positive elements exist in every row of matrix $\mathbf{P}$) in every column of matrix $\mathbf{P}^n (n = N+1, N+2, \ldots)$ if the involution of matrix $\mathbf{P}$ is continued beyond any limit. $\lim_{n\to\infty} \mathbf{P}^n = \mathbf{P}^*$ is a matrix in which the columns are identically built up, or in other words, where any row is identical:

$$\lim_{n\to\infty} \mathbf{P}^n = \lim_{n\to\infty} \begin{bmatrix} p_{11}^{(m)} & p_{12}^{(m)} \cdots p_{1m}^{(m)} \\ \cdot & \cdot \quad \cdots \cdot \\ p_{m1}^{(m)} & p_{m2}^{(m)} \cdots p_{mm}^{(m)} \end{bmatrix} = \begin{bmatrix} p_1^* & p_2^* \cdots p_m^* \\ p_1^* & p_2^* \cdots p_m^* \\ \cdot & \cdot \quad \cdots \cdot \\ p_1^* & p_2^* \cdots p_m^* \end{bmatrix} = \mathbf{P}^*.$$

In case of Markov-chains with finite number of states $\mathbf{P}^*$ is a stochastic matrix: $p_j^* \geqq 0$, $\sum_{k=1}^{m} p_k^* = 1$. The rows of matrix $\mathbf{P}^*$ represent the same distribution, called limit-distribution. This limit-distribution is independent of the corresponding initial distribution. The chain reaches state j — in a limit-case — with the same probability p_j^* no matter what the initial state was. Markov-chains for which the limits

$\lim_{n\to\infty} p_{ij}^{(n)} = p_j^*(j=1, 2, \ldots, m)$ exist, which are independent of i and for which $\sum_1^m p_j^* = 1$, are called ergodic. The necessary and sufficient condition for a Markov-chain to be ergodic is defined by Markov's theorem, in the following form.

Markov's theorem: A homogeneous Markov-chain with a finite number of states is ergodic if its one-step transition-probability matrix

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ . & . & \cdots & . \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}$$

has an Nth power-matrix in which the elements of one column at least are all positive. The velocity of convergence toward the limits $p_j^*(j=1, 2, \ldots, m)$ is exponential:

$$|p_{ij}^{(n)} - p_j^*| \leqq (1-\delta N_1)^{\frac{n}{N}-1} \tag{3.43}$$

where N_1 is the number of columns containing only positive elements in matrix $\mathbf{P}^N$, and δ is the smallest value in these columns.

Proof to this theorem is found in literature (See: A. 23). It should be noted that because in a given column of the limit-matrix $\mathbf{P}^*$ the probabilities are the same, after a longer time (large n) the Markov-chain will attain state j with the same probability whatever its initial state was. We may say that the chain does not remember its past, the dependence on its initial state ceases, for a matrix with identical rows is the representation of a series of independent states.

It is easy to see that the elements of any of the rows of a limit-matrix $\mathbf{P}^*$ are the limits of the n-step absolute probability-elements $p_1^{(n)}, p_2^{(n)}, \ldots, p_n^{(n)}$, at the same time, since

$$p_k^{(n)} = \sum_{i=1}^m p_i^{(0)} p_{ik}^{(n)},$$

and

$$\lim_{n\to\infty} p_k^{(n)} = \lim_{n\to\infty} \sum_{i=1}^m p_i^{(0)} p_{ik}^{(n)} = \sum_{i=1}^m p_i^{(0)} p_k^* = p_k^* \sum_{i=1}^m p_i^{(0)} = p_k^*.$$

The question arises how the limit-matrix $\mathbf{P}^*$ or, which is equivalent, the limit-probabilities $\lim_{n\to\infty} p_{ij}^{(n)} = p_j^*$ can be calculated? It will be shown that the problem is rather simple, it requires only the solution of a linear system of equation with m unknowns.

Let start with the Markovian relationship

$$p_{ik}^{(n+1)} = \sum_{j=1}^m p_j^* p_{jn}$$

and by executing the $n\to\infty$ limit-transition:

$$p_k^* = \sum_{j=1}^m p_j^* p_{jk}. \tag{3.44}$$

In a matrix-form:

$$(3.45) \qquad (p_1^*, p_2^*, \ldots, p_m^*) = (p_1^*, p_2^*, \ldots, p_m^*) \begin{bmatrix} p_{11} & p_{12} \cdots p_{1m} \\ p_{21} & p_{22} \cdots p_{2m} \\ \vdots & \\ p_{m1} & p_{m2} \cdots p_{mm} \end{bmatrix}.$$

It will be shown that this system of equation has a solution consisting of m numbers which are non-negative and their sum is 1. Namely, if $q_1, q_2, \ldots, q_n$ are m real numbers for which

$$q_j \geqq 0, \quad \sum_1^m q_j = 1$$

and

$$(3.46) \qquad q_k = \sum_{j=1}^m q_j p_{jk},$$

then by multiplying both sides of (3.46) by p_k and then summed up, we obtain:

$$(3.47) \qquad q = \sum_{k=1}^m q_k p_k = \sum_{j=1}^m q_j \sum_{k=1}^m p_{jk} p_k = \sum_{j=1}^m q_j p_j^{(2)}.$$

If the procedure is continued then for every positive n it is true that

$$(3.48) \qquad q_k = \sum_{j=1}^m q_j p_{jk}^{(n)}.$$

If $n \to \infty$, then

$$(3.49) \qquad q_k = \lim_n \sum_{j=1}^n q_j p_{jk}^{(n)} = \sum_{j=1}^m q_j p_k = p_k \quad (k = 1, 2, \ldots, m).$$

So, in case of ergodic Markov-chains the calculation of limit-probabilities is a simple problem of linear algebra.

It should be noted that the $\mathbf{P} = [p_{ij}]$ transition-probability matrix is double-stochastic and if the Markov-chain is ergodic, then the limit-matrix takes the form:

$$(3.50) \qquad \mathbf{P}^* = \begin{bmatrix} \frac{1}{m} & \frac{1}{m} \cdots \frac{1}{m} \\ \frac{1}{m} & \frac{1}{m} \cdots \frac{1}{m} \\ \cdot & \cdot \cdots \cdot \\ \frac{1}{m} & \frac{1}{m} \cdots \frac{1}{m} \end{bmatrix}$$

due to the rule that any power of a double-stochastic matrix is double-stochastic and in the columns of a limit-matrix uniform elements can be found.

The Markov-chain is stationary if the absolute-probabilities $p_j^{(n)} (j=1, 2, \ldots, m)$ are independent of n, so

$$(p_1^{(1)}, p_2^{(1)}, \ldots, p_m^{(1)}) = (p_1^{(0)}, p_2^{(0)}, \ldots, p_m^{(0)}) \begin{bmatrix} p_{11} & p_{12} \cdots p_{1m} \\ p_{21} & p_{22} \cdots p_{2m} \\ . & . \quad \ldots . \\ p_{m1} & p_{m2} \cdots p_{mm} \end{bmatrix} = (p_1^{(0)}, p_2^{(0)}, \ldots, p_m^{(0)}).$$

Further:

$$(p_1^{(2)}, p_2^{(2)}, \ldots, p_m^{(2)}) = (p_1^{(1)}, p_2^{(1)}, \ldots, p_m^{(1)}) [\mathbf{P}] = (p_1^{(0)}, p_2^{(0)}, \ldots, p_m^{(0)}) [\mathbf{P}^2].$$

etc.

It is visible that in case of ergodic Markov-chains (3.45) has only a single solution:

$$(p_1^{(0)}, p_2^{(0)}, \ldots, p_m^{(0)}) = (p_1^*, p_2^*, \ldots, p_m^*).$$

If the Markov-chain is ergodic the limiting distribution $p_j^* (j=1, 2, \ldots, m)$ is stationary, at the same time.

3.2. MARKOV PROCESSES WITH FINITE OR COUNTABLE INFINITE STATES

The termonology introduced for the description of Markov-chains is applied once again in case of a physical system with continuously changing states in time. (Such a physical system is e.g. the set of discharges of a river at a gauging-station). According to our present assumption the state-transitions may occur in any time t, regularity between the consecutive steps is not a prerequisite. Another assumption should be, on the other hand, that the system would have finite or countable infinite possible states denoted by 0, 1, 2, ... integers. Such a physical system is called a system with discrete state-space. Let denote the state of our system in time point t by X_t. Let assume that the system is stochastic, and X_t is a random variable for every t. A further assumption is that if the system is in state i at a time point s then it will reach state j at time $s+t$ with a probability $p_{ij}(t)$ independent of its former behaviour:

$$p_{ij}(t) = P(X_{s+t} = j \mid X_s = i) \quad (i, j = 0, 1, 2, \ldots). \tag{3.67}$$

This model is called a homogeneous Markov-process, an analog of the homogeneous Markov-chain. The probabilities $p_{ij}(t)$ are called transition-probabilities for this case again. Similarly to the Markov-chains the initial distribution is used for the determination of the process:

$$p_i(0) = P(X_0 = i) \quad (i = 0, 1, 2, \ldots). \tag{3.68}$$

Let denote by $p_j(t)$ the probability that the system is in state j in time point t:

$$p_j(t) = P(X_t = j). \tag{3.69}$$

Let be further:

$$p_{ij}(0) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (i, j = 1, 2, \ldots).$$

In a similar way than in the case of Markov-chains the following relationships hold:

$$p_j(t) = \sum_i p_i(0)\, p_{ij}(t)$$

$$p_{ij}(s+t) = \sum_k p_{ik}(s)\, p_{jk}(t) \quad \text{for desired} \quad s \geqq 0, \quad t \geqq 0. \tag{3.70}$$

The transitions of a Markov-process with continuous time parameter and discrete state-space may occur in any desired t time point, as this was earlier mentioned. The run, or life-time of such a process can be presented by a function. Let denote the time points in which the transitions occur by $t_1 < t_2 < \ldots < t_n \ldots$. In a coordinate-system (t, X_t) a step-function X_t stands for the behaviour of the process. This function has jumps at time points $t_1, t_2, \ldots$ which in turn is a function of the transition from one state into another. In interval (t_i, t_{i+1}) the function is constant, see Figure 35.

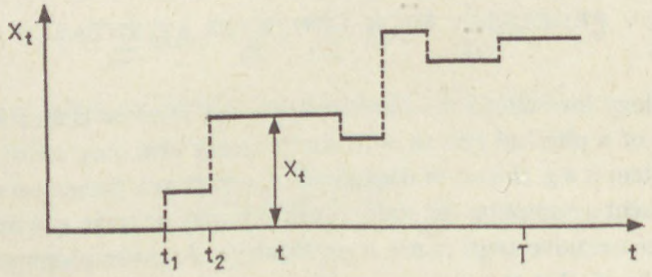


Figure 35

Every function X_t is a possible realization of the process. The set of possible realizations should be denoted by $\{X_t\}$ which in turn may be considered a Markov-process $\{X_t\}$ with a discrete state-space. X_t is a random variable for desired $t-s$. The time interval between two jumps of the state is also a random variable. Let denote this by $\tau (\tau \geqq 0)$. It will be shown that due to the Markovian property (3.70) τ is exponentially distributed.

Let assume that in time point $t = t_0$ the system was in state i. Then, another state has been reached in a time point $t_0 + \tau$ being also of random nature. What is the probability that $P(\tau \geqq t)$?

Let be

$$P(\tau \geqq t) = \varphi(t). \tag{3.71}$$

If now $s < t$ and $\tau > s$ then the system is still in state i in time point $t_0 + s$ on the basis of relationship (3.71):

$$P(\tau > s+t | \tau > s) = P(\tau \geqq t) = \varphi(t).$$

Because

$$P(\tau > s+t) = P(\tau > s+t | \tau > s) P(\tau > s) = \varphi(t)\varphi(s)$$

the following function-equation may be obtained:

$$\varphi(s+t) = \varphi(s)\varphi(t). \tag{3.72}$$

It can be shown that a function of the form $\varphi(t)=e^{-\lambda t}$ is a unique solution to equation (3.72). As a consequence,

$$P(\tau \geqq t) = e^{-\lambda t}. \tag{3.73}$$

The distribution function of τ is then:

$$F(t) = P(\tau < t) = 1-e^{-\lambda t} \quad (t \geqq 0), \tag{3.74}$$

which is the well-known exponential distribution itself. (See Chapter 2. Section 12). Parameter λ is called intensity of movement and is nothing else than the reciprocal of the mean of λ, because $E(\tau)=\frac{1}{\lambda}$ (See Chapter 2., Section 2.102). If $\lambda=0$, the system is remaining in the same state. If $\lambda>0$, then the probability that the system will leave a given state during a small Δt time interval is:

$$1-\varphi(\Delta t) = \lambda \Delta t + 0(\Delta t) \tag{3.75}$$

$\left(\text{where } (0)\cdot \text{ is a quantity for which } \lim_{\Delta t \to 0} \frac{0(\Delta t)}{\Delta t} = 0\right)$.

To conceive (3.75) the following should be known:

$$\lim_{\Delta t \to 0} \frac{1-e^{-\lambda \Delta t}}{\Delta t} = 1.$$

Let now denote by $\tau_1, \tau_2, \ldots$ the time intervals lasting till the first, second, ... etc. transition. The probability $P_n(t)$ should be now calculated that in a time interval $[0, t]$ the number of transitions was exactly n.

Let denote by Y_t the number of transitions in interval $[0, t]$! Event $\{Y_t \geqq n\}$ is obviously equivalent to event $\{\tau_1+\tau_2+\ldots+\tau_n<t\}$. It is easy to see that:

$$P_n(t) = P(Y_t = n) = P(Y_t \geqq n) - P(Y_t \geqq n+1). \tag{3.76}$$

By introducing $F_n(t)=P(\tau_1+\ldots+\tau_n<t)$ one may obtain:

$$P_n(t) = F_n(t) - F_{n+1}(t). \tag{3.77}$$

$F_n(t)$ is the distribution function of the sum of n independent, exponentially distributed random variables. Because it was shown in Chapter 2 and Section 2.12 that the sum of independent, exponentially distributed random variables is gamma-distributed,

therefore:

$$P_n(t) = F_n(t) - F_{n+1}(t) = \int_0^t \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x} dx - \tag{3.78}$$

$$- \int_0^t \frac{\lambda^{n+1}}{n!} x^n e^{-\lambda x} dx.$$

If substitution $\lambda x = u$ is introduced, then by partial integration we obtain

$$\int_0^t \frac{(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x} d(\lambda x) = \frac{1}{(n-1)!} \int_0^{\lambda t} u^{n-1} e^{-u} du = \frac{u^n}{n!} e^{-u} \Big|_0^{\lambda t} + \frac{1}{n!} \int_0^{\lambda t} u^n e^{-n} du =$$

$$= \frac{(\lambda t)^n}{n!} e^{-\lambda t} + \frac{1}{n!} \int_0^t (\lambda x)^n e^{-\lambda x} dx.$$

It follows that

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \tag{3.79}$$

Relationship (3.79) shows that the number of transitions of random variable Y_t in interval $[0, t]$ is Poisson-distributed with a paramter λt. The number of states in interval $(0, T)$ may be described again by help of the function Y_t. If the transitions occurred at time points $t_1, t_2, \ldots, t_n, \ldots$ where Y_t has had unit jumps, then its value will be constant in between, see Figure 36.

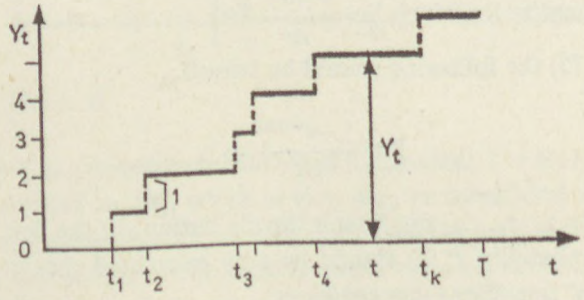


Figure 36

The step-function gained so far is a realization of the random process $\{Y_t\}$. The set of all these step-functions is called a Poisson-process $\{Y_t\}$. The Poisson-process is also a Markov-process with discrete state-space. The set of all possible realizations of the Poisson-process is considered now as the state-space of an experiment consisting of the counting of the transitions in interval $[0, T]$ of a given Markov-process. The state-space is a set of functions in this case, denoted by Ω. Any function-realization is now an elementary ω event. Event $A_k = \{Y_t = k\}$ consists of a set of elementary events, in other words of a set of step-functions which will run to a height k at time point t. This function-set is a subset of Ω.

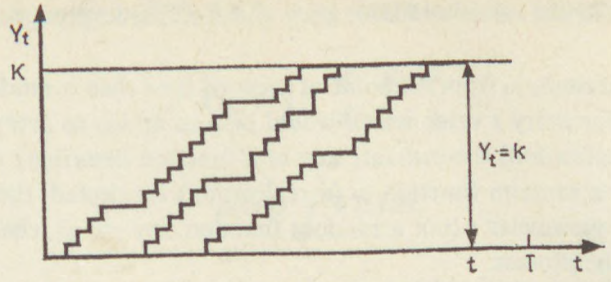


Figure 37

According to formula (3.79)

$$P(A_k) = P(Y_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (t \geqq 0).$$

So by fixing t a probability has been determined on the subsets of a state-space Ω consisting of a set of functions. Random variable Y_t will attach to each elementary event ω, in other words, to each step-function an ordinate belonging to t.

It must be noted that a Markov-process with discrete state-space is a special case of the random processes fluctuating according to time, or by different terminology of the stochastic processes. Our experiment may consist of continuously measured stages in an interval $[0, T]$ at a given site which is called the realization of a X_t stage-process. The set of all such realizations may be considered as a state-space Ω, as a function-set defined in interval $[0, T]$. Again, every random realization is considered an ω elementary event. Subsets of event-space Ω may be determined as follows: let select time points $0 \leqq t_1 < t_2 < \ldots < t_n = T$, let define intervals $[a_1, b_1]$, $[a_2, b_2]$, ..., $[a_n, b_n]$ and let see the set of X_t functions which will run in the given time points between the given intervals, see Figure 38.

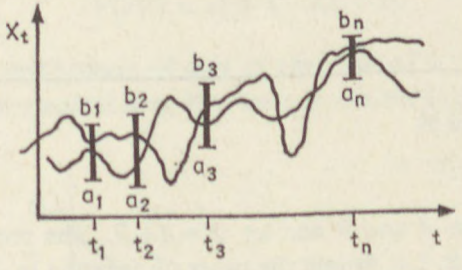


Figure 38

If A is an event that a realization is moving in between the denoted intervals at time points $t_i (i = 1, 2, \ldots, n)$, then

$$P(A) = P(a_1 \leqq X_{t_1} < b,\ a_2 \leqq X_{t_2} < b_2, \ldots, = a_n \leqq X_{t_n} < b_n).$$

In this way, a set of n-dimensional probabilities may be assessed over an event-space $\{X_t\} = \Omega$. If for every n and for any desired $t_1 < t_2 < \ldots < t_n$ interval the appropriate

n-dimensional distribution is available then the stochasitic process $\{X_t\}$ has been specified.

If the process is random from the point of view of time then a random variable X_t may be defined for every t value which would help to attach to every ω elementary event (to each realization) the ordinate at t of a function describing this elementary event ω. If now a random function ω (a realization) is selected, then $X_t(\omega)$ is the function of time parameter t (not a random function any more), characterizing one concrete run of the process.

3.3. DURATION OF A FLOOD-WAVE (OF THE TIME OF FLOODING): A STOCHASTIC PROCESS

In this chapter a mathematical concept is presented to determine the seasonal average duration of a flood-wave over a given level. It should be stressed that separate analyses are made for the first, second, etc. season in order to overcome the effects of seasonality.

The values X of the stages, a stochastic process denoted by $\{X_t, 0 \leqq t < +\infty\}$ is split in two disjunct intervals by a given c level based on our decision from the point of flood-protection, see Figure 39.

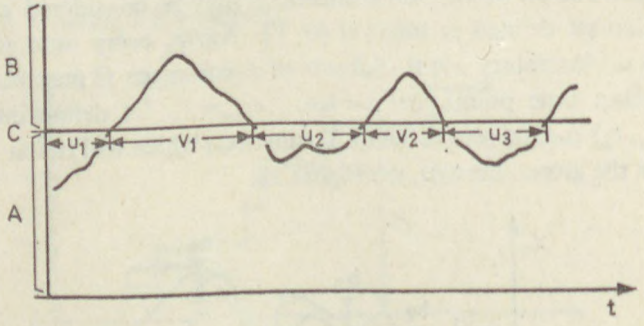


Figure 39

Let these intervals be A and B and let $X=A+B$. The process will interchange values between A and B. Let denote the times of duration in A by $U_1, U_2, \ldots, U_n$ and in B by $V_1, V_2, \ldots, V_n$.

Assume that the random variables $U_1, U_2, \ldots, V_1, V_2, \ldots$ are non-negative, independent, and

$$P(U_i < x) = G(x), \quad P(V_j \leqq x) = H(x).$$

$G(x)$ and $H(x)$ are distribution functions, continuous from the left and from the right, respectively.

Let now define process $\{X(t), 0 \leqq t < +\infty\}$ as the indicator process of state B:

$$X(t) = \begin{cases} 1 & \text{if } X_t \in B \\ 0 & \text{if } X_t \in A. \end{cases} \tag{3.80}$$

Let be

$$\beta(t) = \int_0^t X(\tau)\, d\tau. \tag{3.81}$$

Variable $\beta(t)$ is a measure of how long the process has been found in state B, while $\alpha(t) = t - \beta(t)$ is the same for state A.

Let now introduce the following notations:

$$\begin{aligned} S_n &= U_1 + U_2 + \ldots + U_n \\ T_n &= V_1 + V_2 + \ldots + V_n \end{aligned} \quad (n = 1, 2, \ldots) \tag{3.82}$$

$$P(S_n < x) = G_n(x) \tag{3.83}$$

$$P(T_n < x) = H_n(x). \tag{3.84}$$

It is obvious that if $x \geqq 0$ then $H_0(x) = 1$, if $x < 0$ then $H_0(x) = 0$ and by agreement

$$G_0(x) = 1.$$

The distribution function of variable $\beta(t)$ is:

$$P(\beta(t) < x) = Z_t(x). \tag{3.85}$$

Consequently,

$$P(\alpha(t) < x) = 1 - Z_t(t - x). \tag{3.86}$$

Let consider the time-process of stay as the path of a random-walking particle where vector U_i is a horizontal step, V_j a vertical one, see Figure 40.

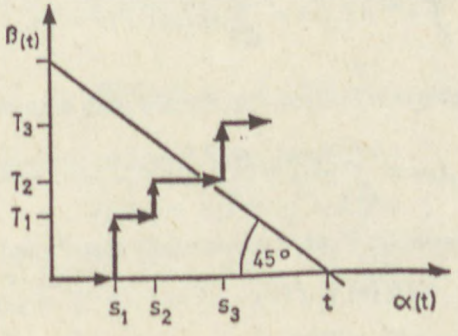


Figure 40

Finally, let be

$$\alpha = \int_0^\infty x\,dG(x), \quad \beta = \int_0^\infty x\,dH(x) \tag{3.87}$$

$$\sigma_\alpha^2 = \int_0^\infty (x-\alpha)^2\,dG(x), \tag{3.88}$$

$$\sigma_\beta^2 = \int_0^\infty (x-\beta)^2\,dH(x).$$

Let now calculate the distribution function of random variable $\beta(t)$.
It is easy to see that

$$\beta(t) = V_1+V_2+\ldots+V_\nu \tag{3.89}$$

where ν is again a random variable. (It is the number of exceedances. The last exceedance is often truncated. It should be taken as complete.) So, it is the problem of determining the distribution function of the sum of random number of random variables. The problem was discussed by several authors (e.g. Takács [B. 37], Todorovic [B. 38]). Solution may be obtained in a rather simple way by use of the theorem of complete probability:

$$P(\beta(t) < x|\nu = k) = P(V_1+V_2+\ldots+V_k < x) = H_k(x) \tag{3.90}$$

$$Z_t(x) = P(\beta(t) < x) = \sum_{k=0}^\infty H_k(x)P(\nu = k).$$

An interesting result is obtained if the characteristic functions of both sides are formulated:

$$\int_0^\infty e^{iux}\,dZ_t(x) = \sum_{k=0}^\infty \left[\int_0^\infty e^{iux}\,dH_u(x)\right] P(\nu = k). \tag{3.91}$$

Let denote the characteristic functions of the variables V_i by $\varphi(u)$.
The initial assumption was that $V_1, V_2, \ldots, V_n$ are independent and are identically distributed:

$$\int_0^\infty e^{iux}\,dZ_t(x) = \sum_{k=0}^\infty [\varphi(u)]^k P(\nu = k). \tag{3.92}$$

If now it is assumed that ν is Poisson-distributed with a parameter λt, then:

$$\varphi_{\beta(t)}(u) = \sum_{k=0}^\infty [\varphi(u)]^k \frac{(\lambda t)^k}{k!} e^{-\lambda t} = e^{-\lambda t[1-\varphi(u)]}. \tag{3.92'}$$

Let assume that the variables V_i are exponentially distributed, so

$$H(x) = P(V_i < x) = 1-e^{-\gamma x} \tag{3.93}$$

and the sum $T_k = V_1+\ldots+V_k$ will be gamma-distributed.

Namely, the characteristic function of a gamma distribution with parameters α and γ is:

$$\varphi(u) = \frac{\gamma^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\gamma x} e^{iux}\, dx = \frac{\gamma^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\gamma-iu)x}\, dx = \tag{3.94}$$

$$= \frac{\gamma^\alpha}{\Gamma(\alpha)(\gamma-iu)^\alpha} \int_0^\infty [(\gamma-iu)x]^{\alpha-1} e^{-(\gamma-iu)x}\, d[(\gamma-iu)x] = \frac{1}{\left(1-\frac{iu}{\gamma}\right)^\alpha}.$$

The characteristic function of the exponential distribution is a special case of the above formula with substitution: $\alpha=1$.

$$\varphi(u) = \frac{1}{1-\frac{iu}{\gamma}}. \tag{3.95}$$

Because the characteristic function of the sum of independent random variables is the product of the characteristic functions, it follows that

$$\varphi_{T_k}(u) = \frac{1}{\left(1-\frac{iu}{\gamma}\right)^k}$$

and

$$P(V_1+\ldots+V_k < x) = \frac{\gamma^k}{\Gamma(k)} \int_0^x z^{k-1} e^{-\gamma z}\, dz.$$

Based on this:

$$Z_t(x) = P(\beta(t) < x) = \sum_{k=0}^\infty \frac{\gamma^k}{\Gamma(k)} \left(\int_0^\infty Z^{k-1} e^{-\gamma z}\, dz \right) \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

And, on the basis of formula (3.92′):

$$\varphi_{\beta(t)}(u) = e^{\frac{i\lambda tu}{\gamma-iu}}. \tag{3.96}$$

The calculation of the exact distribution of variable $\beta(t)$ is rather complicated. It is easy, however, to determine the mean of durations $\beta(t)$.

The conditional expectation of variable $\beta(t)$ is the following by aid of the distribution function (3.93) and by assuming that $\nu=k$ and that the distribution of V_i is exponential:

$$E(V_1+V_2+\ldots+V_k) = \frac{k}{\gamma}. \tag{3.97}$$

On the basis of the theorem of complete probabilities:

$$E[\beta(t)] = E\{(\beta(t)|\nu = k)\} = \tag{3.98}$$

$$= \sum_{k=0}^\infty \frac{k}{\gamma} \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \frac{\lambda t e^{-\lambda t}}{\gamma} \sum_{k=1}^\infty \frac{(\lambda t)^{k-1}}{(k-1)!} = \frac{\lambda t}{\gamma}.$$

This result is very important from a practical point of view.

The same result may be obtained by aid of the $Z_t(x)$ distribution function:

$$\frac{dZ_t(x)}{dz} = Z_t(x) = \sum_{k=0}^{\infty} \frac{\gamma^k}{\Gamma(k)} x^{k-1} e^{-\gamma x} \frac{(\lambda t)^k}{k!} e^{-\lambda t} \tag{3.99}$$

$$E[\beta(t)] = \sum_{k=0}^{\infty} \frac{\gamma^k}{\Gamma(k)} \left(\int_0^{\infty} x^k e^{-\gamma x} \, dx \right) \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \sum_{k=0}^{\infty} \frac{\gamma^k}{\Gamma(k)} \frac{1}{\gamma^{k+1}} \cdot \tag{3.100}$$

$$\cdot \left(\int_0^{\infty} (\gamma x)^k e^{-\gamma x} \, d(\gamma x) \right) \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \frac{\lambda t}{\gamma}.$$

The result is trivial. If V_i is exponentially distributed according to distribution function (3.93) then its mean is $\frac{1}{\gamma}$ which means that during one exceedance the flood will stay in the mean for a time interval of $\frac{1}{\gamma}$ above level c. Due to the fact that the expected number of exceedances is λt, the expected value of being above level c is: $\lambda t \cdot \frac{1}{\gamma}$.

PART II
STATISTICAL INFERENCE

CHAPTER 4

4.1. MATHEMATICAL SATISTICS AS A SECTION OF PROBABILITY THEORY

Mathematical statistics is a very important section of probability theory, particularly for practice. Its scope is similarly the analysis of mass phenomena, however, its problems and, consequently, in most cases its methods are of peculiar character.

Up to now the probabilities of some events were taken as known value and the determination of probabilities belonging to more sophisticated events were problems to be solved. The distribution of a certain random variable (its cumulative distribution function or density function and their parameters) was also taken known and the questions associated with probabilities were answered in the possession thereof. On this basis certain anticipation could be got on the future course of phenomena.

For instance, it was said that if a random variable X having a normal (Gaussian) distribution with expected value m and standard deviation $\sigma>0$ the probability that observed value of X would fall into a given interval (a, b) was

$$P(a \leqq X \leqq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-m)^2}{2\sigma^2}} dx.$$

However, in practice the probabilities of events in question are usually unknown; it cannot be known whether the distribution of a given random variable is normal — although sometimes, through certain theoretical considerations such as the validity of the central limiting theorem or through previous experience, the anticipation of normal distribution may be justified — and the expected value and standard deviation of X that is parameters m and σ defining the distribution are also unknown. How to determine the unknown probability $P(A)=p$ for an event A? How to calculate the expected value of a given random variable X? How to calculate the standard deviation of a random variable X? How to decide whether the distribution of a given random variable is or isn't normal (Gaussian)?

These problems are not discussed in other sections of probability theory. To answer such questions is the task of mathematical statistics. If a hydrologist concerned with the behaviour of floods wants to utilize as an aid the methods of probability theory, too, he will face primarily problems of this kind and if these remain unanswered the application of probability theory will be impossible. To demonstrate what was said an example is given below.

Suppose that one's task is to study the statistical rules prevailing in the development of flood peaks observed at Szeged in the Tisza river. Let the random variable X be chosen as the peaks of flood waves at Szeged. The occurrence of an event $A=\{X>8\text{ m}\}$, involving inundations, represents a rather serious danger so that the knowledge of probability $P(A)=p$ would be important. Without experimental data (observations, measurements), solely through speculation, this unknown probability cannot be determined. The records of floods peaks observed in the Tisza river at Szeged should be at hand (e.g., taken from the Hydrographic Yearbook) and then, based on observations $X_1, X_2, \ldots, X_n$ (in our case $n=100$), the determination of the relative frequency k/n of event A will become possible. By virtue of the law of large numbers — if n is high enough — the relative frequency k/n will not differ greatly from the unknown probability p. In our case, as it is usual with hydrological records, n cannot be considered high enough so that the exact determination of the unknown probability by means of relative frequency is impossible, a certain approximation to — a so-called estimate of — probability p should be accepted as satisfactory. The unknown probability is a certain point within the interval [0, 1]. Now, even if the exact value of p cannot be calculated, at least a relatively short interval containing the probability p ought to be found. To achieve this the opportunity is given since the number k of outcomes when (in the course of n experiments) event A will occur is a random variable with binomial distribution whose expected value is $E(k)=np$ and standard deviation is $D(k)=\sqrt{npq}$. Consequently, the expected value and standard deviation of relative frequency k/n is $E\left(\frac{k}{n}\right)=p$ and $D\left(\frac{k}{n}\right)=\frac{pq}{n}$, respectively. Now, by virtue of the Moivre—Laplace theorem on limiting distribution (see Chapter 2, Section 2.8) and as a consequence of the known property of normal distributions

$$P\left(\left|\frac{k}{n}-p\right|<2\sqrt{\frac{pq}{n}}\right)\approx 0.95. \tag{4.1}$$

This relationship means that the interval

$$\left(\frac{k}{n}-2\sqrt{\frac{pq}{n}},\frac{k}{n}+2\sqrt{\frac{pq}{n}}\right) \tag{4.2}$$

which, since the value of its centre, $\frac{k}{n}$, is randomly dependent, will contain the unknown probability with high certainty. Formula (4.2), however, is not of great value since the length is unknown and it contains an unknown p (and a $q=1-p$). Introducing now the estimates $p\approx k/n$ and $q\approx\frac{n-k}{n}$ such a result will be obtained that the interval

$$\left(\frac{k}{n}-\frac{2}{n}\sqrt{\frac{k(n-k)}{n}},\frac{k}{n}+\frac{2}{n}\sqrt{\frac{k(n-k)}{n}}\right) \tag{4.3}$$

whose position and length is of random nature will cover the unknown probability p by a probability of about 95 per cent; this means that if the procedure of constructing the above interval on the basis of n observations is repeated many times the probability p will be contained in this interval in 95 per cent of the cases while the reverse will be true in 5 per cent. This procedure of interval estimation is the so-called method of confidence intervals which will be dealt with in Section 5.1.4 more exactly and in more detail; it can be used to estimate not only an unknown probability but also the numerical characteristics of probability distributions such as expected values, standard deviations, quantities, etc., that is to derive approximate statistical estimates for generally unknown parameters. (Note that when estimating an unknown probability p it is the parameter p of binomial distribution that is estimated.)

The statistical determinaton of an unknown (constant) parameter is dealt with in a section of mathematical statistics called *theory of estimation.*

In connection with the random variable X denoting the peaks of flood waves it is, of course, not only the probability of event $A=\{X>8\text{ m}\}$ that a hydrologist is interested in. It would be much more meaningful if, for an arbitrary x, the probability of event $\{X>x\text{ cm}\}$ or of just the reverse event $\{X<x\text{ cm}\}$ could be determined; this would mean the knowledge of the distribution function $F(x)$ belonging to the random variable X.

In mathematical statistics methods to estimate the unknown distribution function $F(x)$ (or the density function $F'(x)=f(x)$) of a given random variable X can also be found (see Chapter 4, Section 1.3 and Chapter 4, Section 1.5).

Frequently what is necessery is not to find parameters or a distribution but to decide whether one or more parameters or the very distributions of two statistical populations are or aren't identical. As to the flood waves of River Tisza the records date back to some 100 years. For instance, hydrologists may be interested in the problem whether the flood peak levels are or aren't increasing. To decide this a procedure may be, e.g., that by regarding the peak values measured in the period 1876/1936 as observed values for a random variable X and those measured in the period 1936/1976 as observations for a random variable Y and by using the methods of mathematical statistics an examination is made on whether the hypothesis $E(X)=E(Y)$ or another one, e.g., $F(x)\equiv G(x)$ can or cannot hold; here $F(x)$ denotes the distribution function of the random variable X and $G(x)$ denotes that of Y. Thus here the identity of distributions or of parameters is presumed and a special method of mathematical statistics, the so-called hypothesis test, is used to make decision on the acceptance or rejection of the hypothesis. Hypothesis testing constitutes another large domain of mathematical statistics which is based on observations (experience) as well as is the theory of estimation. A set of observations if it possesses the properties to be discussed later (see Section 4.1.1) is called statistical sample.

The theory of estimation and the statistical tests of hypotheses are two — and already classical — sections of mathematical statistics; from the viewpoint of hydrological applications both fields are considered to hold principal importance. These topics

have been integrated in the theory of statistical decision functions by Abraham Wald (see Section 6.7.1).

Based on the foregoing, in general terms the basic task of mathematical statistics can be formulated as follows: on the basis of experience (observations, measurements) inference has to be made on unknown probabilities of events or on unknown distribution functions and parameters of random variables. Mathematical statistics tries to give solutions to this fundamental problem by elaborating methods which, utilizing the observations, provide the most possible information on the required theoretical values.

The specification of the basic task of mathematical statistics as given above indicates the distribution nor the parameters of the random variable X was known in advance so that they should be determined from observations related to X. In the hydrology of floods this means that in order to have anticipations on future the past observed values of random variables associated with flood waves should be relied on.

In statistical theory a sequence of independent and identically distributed random variables

$$X_1, X_2, \ldots, X_n \tag{I}$$

is called statistical sample. This means that we have the result of n independent observations made under identical circumstances on a certain random variable X. Results of the individual observations as random variables are the elements of the sample. The number of sample elements is called the size of the sample. The sample elements are listed in the sequence of observations.

Statistical inference is a special procedure or decision of probabilistic nature; in the following our intention is to present it in connection with various problems.

This chapter will deal with the basic notions and methods of mathematical statistics. In practice there is a need to combine the different methods and to apply them in a sequence depending on the problem at hand. To demonstrate this, in the forthcoming chapters the statistical analysis of some hydrological problems associated with floods will be presented assuming that the methods outlined in the present chapter are known.

4.1.1. THE SAMPLE. PROCESSING OF HYDROLOGICAL RECORDS

As it was mentioned in the previous section in most cases in the hydrological practice neither statistical sample is a set of independent random variables whose number is finite, having identical distributions. To demonstrate this an example is given below. Let the random variable X denote the annual maximum stages at a given gauge. Making observations on the annual maximum stages at this gauge during n years sample (I) will be obtained. Through these observations, of course, concrete values, namely,

$$X_1 = x_1, \; X_2 = x_2, \ldots, X_n = x_n$$

that is n numerical values will be obtained whose set will also be called statistical sample in the following.

In flood hydrology sometimes the requirement of stochastic independence is met in approximation only. The annual maximum stages are, in general, independent because usually there is a rather long interval (several months or even more than one year) between two subsequent observations of flood stages. However, as to the diurnal stages, they are evidently not independent of one another. So by making daily records on stages during n days, by virtue of the foregoing, not a statistical sample but a so-called time series will be obtained which will be discussed in section 4.1.1. Random variables associated with floods such as the magnitude of maximum exceedances above a given level c in the individual flood waves or the maximum flood flows observed in each flood wave, etc., are generally of such kind that the elements of the associated sample are independent of each other (especially for subsequent flood waves with a rather prolonged interval in between). Statistical methods utilizable for checking the independence of results given by a sequence of hydrological observations will be described in detail later on. The most elaborate theory of mathematical statistics relates to samples consisting of independent elements; good approximations to distributions, etc., can be obtained from such samples. This is the reason why independent observations are strived for.

In hydrology the requirement that the elements in a sample should be random variables with identical distribution, i.e., that similar observations under identical circumstances should be repeated n times, can be met approximately only (since the impact of environment, man made interventions, runoff conditions, etc., that is important factors are constantly changing). Methods to check whether the distribution of observations can or cannot be considered unchanged will also be recommended (see Section 6.5).

To meet the requirement of having comparable distributions the procedure to be followed is to analyse the random variables associated with flood waves such as, e.g., the magnitude of exceedances, the duration of flood waves, etc., separately for each season.

4.1.2 THE STATISTICAL FUNCTION (STATISTICS)

As it was outlined in Section 4.1.1 the distribution of a random variable studied and the characteristics (parameters) of this distribution were to be concluded from a mathematical sample.

Let

$$P(X < x) = F(x)$$

be the distribution function of a random variable X. Information on the distribution and on its parameters are certainly included in the sample elements but in a scattered form. In order to be retrieved from the sample elements this information should be made more compact, with a view to answer the statistical questions. Another inten-

tion is to perform this compaction possibly without any loss of information. What are the items requiring retrieval of information from the sample?

They are first the distribution itself that is the distribution function $F(x)$, then its position that is its expected value, median and quantiles and, finally, the scattering of its values that is the standard deviation, etc. The way of making the information more compact by using a statistical sample is to compose one or a few data from the sample elements. These extracts of the sample elements are called statistical functions or in shorter form *statistics*. As the sample elements, depending their values on chance, are random variables the statistics calculated from them are random variables as well so that each of them also has a distribution, expected value, standard deviation, etc. Frequently the determination of the exact distribution of a statistic is not a simple task but in case of large samples there are many instances where an approximate distribution (or limiting distribution) of a statistic can still be determined. In general, what is to be estimated in connection with a statistic is its expected value and standard deviation.

4.1.3. THE EMPIRICAL DISTRIBUTION FUNCTION. THEOREM OF GLIVENKO

As a basis of a statistical analysis on the distribution of a random variable X whose distribution $F(x)$ is continuous the statistical sample

$$X_1, X_2, \ldots, X_n \tag{I}$$

will serve. The situation in the practice of hydrology is that the distribution function $F(x)$ of the random variable X is mostly unknown. As it will be seen, when the number of sample elements is high enough a rather good approach to the distribution function $F(x)$ can be found by using the following simple method.

Arrange the elements of sample (I) into a sequence of increasing magnitudes:

$$X_1^* < X_2^* < \ldots < X_n^*. \tag{II}$$

Let now a step function $F_n(x)$ be defined in the following way:

$$F_n(x) = \begin{cases} 0 & \text{if } x \leqq X_1^* \\ k/n & \text{if } X_k^* < x \leqq X_{k+1}^* \\ 1 & \text{if } x > X_n^*. \end{cases} \tag{4.4}$$

The function $F_n(x)$ is called empirical distribution function belonging to the sample concerned. It can easily be seen that the $F_n(x)$ has the properties of a distribution function its values fall in the range between 0 and 1, it is monotonically non-decreasing and it is continuous from the left. At all abscissa points X_i^* $(i=1, 2, \ldots, n)$ $F_n(x)$ has a jump upwards by $1/n$. So at a certain point x the value of $F_n(x)$ is so many times k/n as many sample elements lower than x can be found. In other words the value of $F_n(x)$ at a point x is equal to the relative frequency of the event $\{X<x\}$.

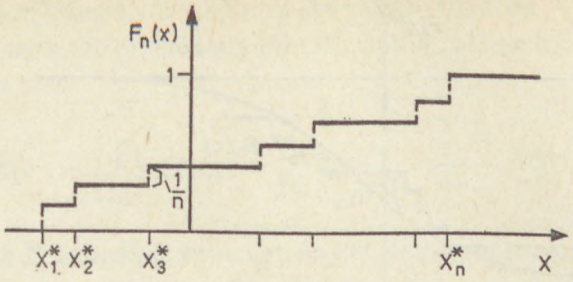


Figure 41

The probability of the same event is $P(X<x)=F(x)$. So the relation between the empirical distribution function $F_n(x)$ and the theoretical cumulative distribution function $F(x)$ is in a way the same as the relation between the relative frequency of a given event and the probability of the same event. If n is large enough there is a high probability that the relative frequency will differ only slightly from the unknown probability.

As known the frequency has binomial distribution whose parameters are the number of observations, n, and the probability of the event which is $F(x)$. Therefore

$$P\left(F_n(x)=\frac{k}{n}\right)=\binom{n}{k}[F(x)]^k[1-F(x)]^{n-k}. \tag{4.5}$$

Furthermore,

$$E[F_n(x)]=F(x) \tag{4.6}$$

and

$$D[F_n(x)]=\sqrt{\frac{F(x)[1-F(x)]}{n}}. \tag{4.7}$$

Now such a question may arise whether, with a fixed n, what an approximate difference between the empirical distribution function $F_n(x)$ and the theoretical distribution function $F(x)$ can be expected?

According to the Chebysev inequality

$$P\left(|F_n(x)-F(x)|<\lambda\sqrt{\frac{F(x)[1-F(x)]}{n}}\right)>\left(1-\frac{1}{\lambda^2}\right). \tag{4.8}$$

Since $F(x)[1-F(x)]\leqq\frac{1}{4}$ it follows that, by a high probability, the absolute value of difference for any x will be in the order of magnitude $1/\sqrt{n}$ which also involves that with increasing n the difference will converge to zero. If $n\to\infty$ then, on the one hand, the jumps of $F_n(x)$ will become gradually smaller and, on the other hand, the difference between $F_n(x)$ and $F(x)$ will be small, see Figure 42. This fact is expressed by the theorem of Glivenko stating that if n is increasing the empirical distribution

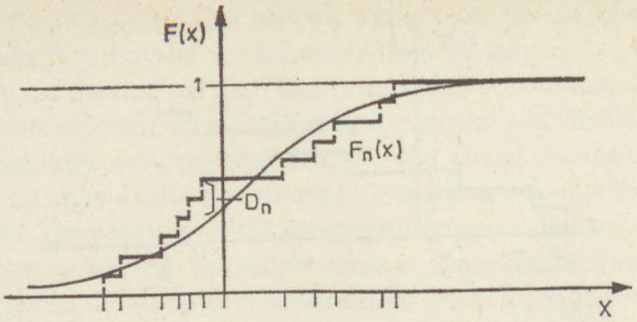


Figure 42

function *converges uniformly on the whole real line to the theoretical distribution function.* In more exact terms, if

$$D_n = \sup_{-\infty<x<\infty} |F_n(x)-F(x)| \tag{4.9}$$

then

$$P\left(\lim_{n\to\infty} D_n = 0\right) = 1. \tag{4.10}$$

Due to its great importance Glivenko's theorem is referred commonly to as the main theorem of mathematical statistics. The convergence of the maximum difference between $F_n(x)$ and $F(x)$ to zero means for the practice that if n is large enough the probability $F(x)$ can be determined approximately by $F_n(x)$. The rate of convergence of the maximum difference between $F_n(x)$ and $F(x)$ to zero is expressed by the theorem of Kolmogorov and Smirnov to be discussed in Section 4.2.5.

If a discrete random variable may take the values $x_1, x_2, \ldots$ with (unknown) probabilities $p_1, \ldots p_2, \ldots$, respectively, and v_i is the number of occurrences of x_i in a sample of size n, $\frac{v_i}{n}$ can be considered an approximation of p.

4.1.4. IMPORTANT EMPIRICAL CHARACTERISTICS. SAMPLE MEAN

One of the most important statistics often used in practice is the arithmetic mean of sample elements, i.e. the sample mean

$$\overline{X} = \frac{X_1+X_2+\ldots+X_n}{n}. \tag{4.11}$$

Being a random variable $\overline{X}$ fluctuates around the theoretical expected value $E(X)=m$. The reader is reminded of the rule that if the distribution function of X is $F(x)$ then $E(X)$ is calculated by using the formula

$$E(X) = \int_{-\infty}^{\infty} x\,dF(x).$$

$F(x)$ is unknown but, instead, the empirical distribution function $F_n(x)$ can certainly be determined by using the ordered sample (II), so if n is large enough, Glivenko's theorem yields that

$$E(X) \approx \int_{-\infty}^{\infty} x\,dF_n(x) = \sum_{i=1}^{n} X_i^* \cdot \frac{1}{n} = \frac{\sum_1^n X_i}{n} = \bar{X}. \tag{4.12}$$

So the sample mean $\bar{X}$ provides information on the position of the location parameter $E(X)$ that is on the centre of gravity of the distribution.

Empirical median

The location of a distribution is illustrated similarly by the so-called empirical median which is the middle one among the elements of the ordered sample (II) if n is an odd number. If $n=2m$ (even number) the median is calculated as the arithmetic mean of two elements in the middle. So

$$\tilde{X}_{1/2} = \frac{X_m^* + X_{m+1}^*}{2} \quad \text{if} \quad n = 2m$$
$$\tilde{X}_{1/2} = X_{m+1}^* \quad \text{if} \quad n = 2m+1. \tag{4.13}$$

Empirical quantiles

If $0<\alpha<1$ then the sample element $X^*_{[n\alpha]+1}$ is called the empirical α-quantile of the distribution. The empirical α-quantile of a distribution is a number compared to which 100α percent of the sample elements are smaller. The median is the quantile belonging to $\alpha=1/2$. In addition, the knowledge of the so-called lower and upper quantiles, $X^*_{\left[\frac{n}{4}\right]+1}$ and $X^*_{\left[\frac{3n}{4}\right]+1}$, respectively, are also highly informative. If the number of sample elements is large enough the empirical quantiles fluctuate around the theoretical quantiles of the distribution.

Estimate of the variance, the sample range

The estimate of the theoretical is the "mean square deviation" defined by the formula

$$S_n^2 = \frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n}; \tag{4.14}$$

it is used to characterize the dispersion of the sample elements around the sample mean $\bar{X}$. The square root of S_n^2 (with positive sign) is called "empirical standard deviation".

Remember that the variance of a random variable X was defined by the formula

$$D^2(X) = \int_{-\infty}^{\infty} (x-m)^2 \, dF(x).$$

Considering this,

$$D^2(X) \approx \int_{-\infty}^{\infty} (x-\bar{x})^2 \, dF_n(x) = \frac{\sum_1^n (X_i^* - \overline{X})^2}{n} = S_n^2$$

that is the estimate of variance, S_n^2, fluctuates around the theoretical variance $D^2(X)$, it is a statistical approximation thereof.

In practice — especially when n is not too large — the so-called corrected estimate of variance is used:

$$S_n^{*2} = \frac{\sum_1^n (X_i - \overline{X})^2}{n-1} = \frac{n}{n-1} S_n^2. \tag{4.15}$$

(The expediency of using this corrected estimate of variance is explained in Section 5.1.3).

As a measure of dispersion the sample range

$$R = X_n^* - X_1^* \tag{4.16}$$

is also used; this is easier to calculate than S_n^2 but, as a statistic, its reliability is lower.

To obtain a coefficient representing the relative dispersion in the case of positive variables the coefficient of variation is calculated:

$$V = \frac{S_n}{\overline{X}}. \tag{4.17}$$

The empirical moments

$$m_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r \quad (r = 1, 2, \ldots)$$

will be needed as well. Note that $m_1 = \overline{X}$ and $m_2 = S_n^2 + \overline{X}^2$.

4.1.5. DENSITY HISTOGRAM OR EMPIRICAL DENSITY FUNCTION

As it was seen, by applying Glivenko's theorem the theoretical distribution function $F(x)$ of a certain random variable X could be approximated by the empirical distribution function $F_n(x)$ and to calculate this approximation in practice was a very simple task. The density function $F'(x) = f(x)$ can give a more illustrative picture on the distribution. In order to obtain statistical approximation for a density function usually a graphic procedure is used in the following manner:

Divide the interval (a, b) whose boundaries are $a = X_1^*$ and $b = X_n^*$ into m partial intervals by means of dividing points $a = d_0 < d_1 < \ldots < d_{m-1} < d_m = b$ where the value

taken by m depends on the number of sample elements, n (the choice of m will be discussed later). If, among the sample elements $X_1, X_2, \ldots, X_n$, the number of those falling into the interval $[d_{i-1}, d_i]$ is denoted by v_i let an oblong be drawn above each interval with a height of

$$\frac{v_i}{n(d_i - d_{i-1})} \quad (i = 1, 2, \ldots, m).$$

In this way a step function will be obtained under which the whole oblong area amounts to 1. This step function is called density histogram or empirical density function. If n is large enough the area under the density histogram belonging to a certain interval (c, d) provides in approximation the probability that the observed value of X falls into (c, d)

Constructing a density histogram usually two problems will arise: how to choose the number m of intervals and how to locate their dividing points. Depending on the sample size n the practical procedure is the partition of the sample range into 8 to 14 parts which reflect the shape of the theoretical density function but depending on spatial circumstances (small or large size, the shape of the density function) this number may be less than 8 or more than 14. The dividing points may be located in a manner to obtain uniform partition (intervals with equal lengths) but the location of points may also be made dependent on the location of the sample elements (e.g., the partial intervals next to both ends can be chosen longer because here the density of sample elements is lower).

To illustrate the foregoing consider the following example:

Denote the random variable X the exceedances measured at Tokaj in the first half of each year between 1903 and 1971. With a choice of $m = 600$ cm the frequencies by which the values of X fell into the intervals

$\Delta_1 =$ 1 to 50 cm $\quad \Delta_4 =$ 151 to 200 cm

$\Delta_2 =$ 51 to 100 cm $\quad \Delta_5 =$ 201 to 250 cm

$\Delta_3 =$ 101 to 150 cm $\quad \Delta_6 =$ 251 to 300 cm

are:

Δ_i	Frequency	Relative frequency, v_i/n	$\frac{v_i}{\Delta n}$
Δ_1	34	0.36	0.72
Δ_2	24	0.26	0.52
Δ_3	15	0.16	0.32
Δ_5	14	0.15	0.28
Δ_4	4	0.04	0.08
Δ_6	3	0.04	0.06
Total	94	1.00	

If, in an orthogonal co-ordinate system, oblongs are drawn over the intervals Δ_i in a manner assuring that the area of each oblong is proportionate to v_i/n, the relative frequency of observations belonging to the respective interval, and that the sum of these areas is equal to 1, a *density histogram* shown in the figure below will be obtained, see Figure 43.

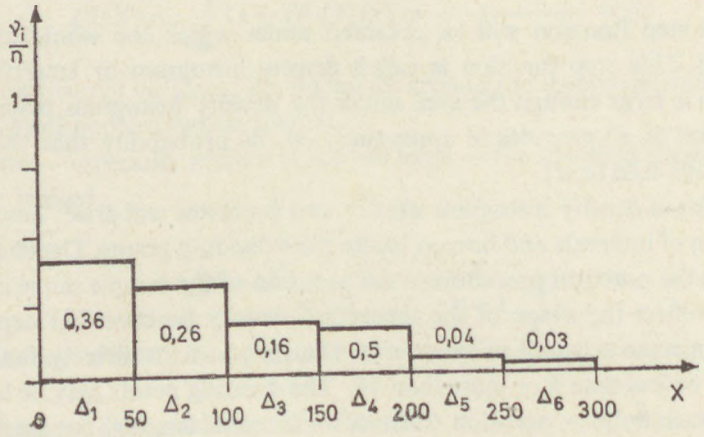


Figure 43

Note that the empirical density function can be defined in several ways. Most frequently the definition

$$f_n(x) = \begin{cases} \dfrac{v_i}{n(d_i - d_{i-1})} & \text{if } d_{i-1} < x < d_i \\ 0 & \text{otherwise} \quad (i = 1, 2, \ldots, m) \end{cases} \tag{4.18}$$

is used. It is easy to see that the emprical density function $f_n(x)$ defined by formula (4.18) is the difference ratio of the empirical distribution function $F_n(x)$ within the interval (d_{i-1}, d_i) that is

$$f_n(x) = \frac{\Delta F_n(x)}{\Delta x} \quad \text{where} \quad \Delta x = d_i - d_{i-1} \quad (i = 1, 2, \ldots, m) \tag{4.19}$$

$$\left(\text{as an analogy of the relationship } f(x) = \frac{dF(x)}{dx}\right).$$

Equation (4.19) provides a graphical technique to get the empirical density function $f_n(x)$ from the empirical distribution function $F_n(x)$ by using the so-called graphical differentiation, a procedure where a unit length is measured along the X axis backwards from each dividing point d_i towards d_{i-1} and a line is drawn parallelly to the chord connecting the points

$$[d_{i-1}, F_n(d_{i-1})], \quad [d_i, F_n(d_i)].$$

It can be shown that, applying equal length of intervals, if the number of dividing points, $m=m_n$, meets the condition

$$\frac{\log n}{\sqrt[3]{n}} < m_n < n^{1-\varepsilon} \quad (\varepsilon > 0)$$

and n is increasing the density function $f_n(x)$ as defined in formula (4.18) will, under rather general conditions, converge to the theoretical density function that is

$$n = \max_x |f_n(x) - f(x)| \to 0 \quad \text{if} \quad n \to \infty.$$

Another definition given for the empirical density function also utilizes the difference ratio of the empirical distribution function $F_n(x)$ with the only difference that the location of dividing points varies from case to case.

E. Parzen's suggestion for the definition of the empirical density function $f_n(x)$ is the formula

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} \tag{4.20}$$

where h is a suitably chosen positive number. Again with this definition the problem faced is how large the number h is to be chosen. Obviously, the choice of h will depend on the number of sample elements, n: $h=h(n)$. The larger the number of sample elements the smaller $h(n)$ may be chosen so that meeting the condition

$$\lim_{n \to \infty} h(n) = 0$$

is an apparent requirement. The numerator of the empirical density function $f_n(x)$ as defined by formula (4.20) is the relative frequency of sample elements contained by the interval whose centre is at x and length is $2h$. In his paper [B. 26] Parzen conducts an in-depth analysis on the following estimator of the density function $f(x)$:

$$f_n(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{y-x}{h}\right) dF_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right). \tag{4.21}$$

This has been derived by means of the function

$$K(y) = \begin{cases} 1/2 & \text{if} \quad |y| \leqq 1 \\ 0 & \text{if} \quad |y| > 1 \end{cases}$$

and represents essentially another form of relationship (4.20): it is an integral mean formed by utilizing the empirical distribution function $F_n(x)$. In addition he has proved [B. 26] that if the relation

$$\lim_{n \to \infty} nh(n) = 0$$

also holds then the empirical density function $f_n(x)$ calculated by formula (4.21) will meet that

$$\lim_{n\to\infty} E[f_n(x)] = f(x) \tag{4.22}$$

and

$$\lim_{n\to\infty} nhD^2[f_n(x)] = f(x) \int_{-\infty}^{\infty} K^2(y)\,dy = \frac{f(x)}{2}$$

that is

$$D[f_n(x)] \approx \sqrt{\frac{f(x)}{2nh}}. \tag{4.23}$$

Formula (4.18) provides a step function for the empirical density function while formula (4.21) is the so-called moving uniform division because once an h value has been chosen x can be caused to run along the domain of $f(x)$.

For instance consider the series of exceedances measured in the first three months in the Tisza river at the Szolnok gauging station (Table T. 1). Elements of the ordered sample are given in Table 4.1. The empirical cumulative distribution function is shown in Fig. 61. In this figure the $\hat{f}_n(x)$ estimates of the theoretical density func-

Table 4.1

Tisza river at Szolnok, first quarters (1903/1970)

1	1	0.01	21	59	0.51
2	2	0.02	22	63	0.54
3	3	0.04	23	65	0.55
4	4	0.05	24	66	0.55
5	5	0.06	25	85	0.65
6	13	0.15	26	88	0.67
7	16	0.18	27	91	0.68
8	17	0.19	28	100	0.72
9	18	0.20	29	104	0.73
10	19	0.21	30	108	0.75
11	20	0.22	31	116	0.77
12	24	0.25	32	128	0.81
13	29	0.30	33	134	0.83
14	33	0.35	34	150	0.86
15	35	0.35	35	178	0.92
16	38	0.37	36	179	0.93
17	39	0.38	37	184	0.93
18	45	0.42	38	201	0.95
19	46	0.43	39	222	0.97
20	56	0.49	40	255	0.99
			41	281	1.00

$\bar{X} = 80.98 \approx 81$ cm

tion $f(x)$ are also plotted at the abscissas $x=0.25$; $x=0.5$ m; $x=1.0$ m and $x=1.5$ m. With respect to the fact that the processing of data contained in the table requires a relatively small amount of calculation, estimation to the density function was by using the empirical density function $\hat{f}_n(x)$, utilizing formula (4.18) where the interval h was chosen as 0.25.

In the recent years the course of estimating a density function through the empirical density function $\hat{f}_n(x)$ contained in formula (4.18) and the properties of estimates have been discussed by several authors, among them Révész [B. 32], Rosenblatt [B. 33], Parzen [B. 26], Nadaraja [B. 25], Tusnády [B. 40], etc.

In case of the sequence in question*

$$\hat{f}_n(0.25) = 0.92; \quad \hat{f}_n(0.5) = 0.6; \quad \hat{f}_n(1) = 0.3; \quad \hat{f}_n(1.5) = 0.15.$$

At the same time

$$1-F_n(0.25) = 0.73; \quad 1-F_n(0.50) = 0.53; \quad 1-F_n(1) = 0.32; \quad 1-F_n(1.5) = 0.15.$$

Considering the results, above a certain c it seems to hold (in the example at least for exceedances larger than 1 m) that

$$\frac{\hat{f}_n(x)}{1-F_n(x)} \approx 1 \quad \text{(constant)}.$$

It is known (see, e.g., [A. 20]) that, when dealing with a continuous distribution of a random variable X, the relation

$$\frac{f(x)}{1-F(x)} = \lambda \quad \text{(constant)} \tag{4.24}$$

is met above a certain c, i.e., when $X>c$ then $P(X<x|X\geqq c)=1-e^{-\lambda x}$. This means, with the aforementioned stipulation, that the conditional distribution of the random variable X is exponential distribution.

Namely, if relation (4.24) fits the distribution of the random variable in question then

$$-\frac{d\ln[1-F(x)]}{dx} = \lambda$$

$$\ln[1-F(x)] = -\lambda x$$

$$1-F(x) = e^{-\lambda x}$$

$$F(x) = 1-e^{-\lambda x}.$$

If the relation (4.24) holds for all x values larger than zero the distribution is exponential.

* The shape of the empirical cumulative distribution function $F_n(x)$ and the course of values in the empirical density function refer equally to an exponential distribution.

Note that E. Zelenhasic [B. 49] has found that in the case of some North American rivers the distribution of the exceedances is exponential or gamma. Now we justify that if the distribution of a random variable X is gamma then, for a certain level of c, the conditional distribution of X, under the condition that $X>c$, will be distributed exponentially, with a good approximation. Namely, as it was seen in Section 1.4.11, the density function of gamma distribution was

$$f(x)=\begin{cases}\dfrac{\lambda^p}{\Gamma(p)}\cdot x^{p-1}e^{-\lambda x} & \text{if } x>0\\ 0 & \text{if } x\leqq 0.\end{cases} \tag{4.25}$$

Applying now L'Hospital's rule it yields that

$$\left.\frac{f(x)}{1-F(x)}\right|_{x=\infty}=-\left.\frac{f'(x)}{f(x)}\right|_{x=\infty}=$$

$$=\left.\frac{\dfrac{\lambda^p}{\Gamma(p)}(p-1)x^{p-2}e^{-\lambda x}-\lambda x^{p-1}e^{-\lambda x}}{\dfrac{\lambda^p}{\Gamma(p)}x^{p-1}e^{-\lambda x}}\right|_{x=\infty}=\left.\frac{1-p}{x}+\lambda\right|_{x=\infty}=\lambda \quad \text{(constant).}$$

So with x values large enough it holds that $\dfrac{f'(x)}{1-F(x)}=\lambda$.

Note that the (constant) value of λ contained in (4.24) is at the same time the parameter of exponential distribution. As out of our record there are a few points only where values for the empirical density function $\hat{f}_n(x)$ have been calculated a result $\lambda=1$ will be denied.

In Section 5.1.2 it will be seen that the maximum likelihood estimator of parameter λ in an exponential distribution is the statistic $\hat{\lambda}=1(\overline{X})$ (reciprocal value of the arithmetic mean of observations).

In our example $\overline{X}=0.81$ so that $\hat{\lambda}=1/0.81=1.25$. A comparison made, e.g., at the points $x=0.5$; $x=1$; $x=1.5$; $x=2$ between the values of the empirical distribution function $F_n(x)$ representing the series and the values of the exponential distribution function $F(x)=1-e^{-1.25x}$ yields the following table:

Table 4.2

x	$F(x)=1-e^{-1.25x}$	$F_n(x)$
0.5	0.47	0.47
1	0.71	0.67
1.5	0.85	0.85
2	0.92	0.90

Based on the good agreement between the theoretical and empirical values the statement is that, for the given gauge and season, the distribution of exceedances is exponential. As to the acceptance of a hypothesis presuming exponential distribution decision can be made by conducting a so-called test of exponentiality. (In literature it is often referred to as Störmer test.)

4.2. ELEMENTS OF THE THEORY OF ORDER STATISTICS

4.2.1. THE ORDERED SAMPLE

The appearance of the theory of ordered samples has opened a new chapter in the development of mathematical statistics. This theory is fundamentally important in the analysis of hydrological observations. In the sequel that part of this theory will be outlined in brief which will be applied in the further chapters.

Let X be a random variable having continuous distribution function represented by n independent observations that is let be considered a statistical sample.

Let

(I) $$X_1, X_2, \ldots, X_n$$

be a sample of size n, i.e., independent observations on the random variable X having continuous distribution. Commonly the elements are listed in random order, e.g., in temporal sequence of observations (measurements). If now the temporal sequence is disregarded and, with consideration to their numerical values, the observations are ranked in increasing order of magnitude a so-called ordered sample will be obtained:

(II) $$X_1^* < X_2^* < \ldots < X_n^*.$$

(If the elements of statistical sample (I) are plotted on the line they will line up automatically according to their magnitudes.)

Anyway, among the elements of sample (I) there will be a smallest one, a subsequent second smallest one, etc. Since the sample elements are independent and identically distributed the smallest observed value X_1^* may be any element of the series X_1, $X_2, X_3, \ldots, X_n$ by a probability of just $1/n$. Similarly, elements $X_2^*, X_3^*, \ldots$ of the ordered sample may be originated from any element of sample (I) by the same probability. So the elements of the ordered sample can be considered random variables as well as the elements of statistical sample (I). However, the elements in the ordered sample (II) are no longer independent of each other as there is a rank relation among them so that their distribution will also differ: the distribution of X_1^* is not the same as that of X_2^* and the respective distribution of $X_3^*, X_4^*, \ldots, X_n^*$ will be different again. It might seem that by ranking the sample elements the advantageous properties of sample (I), i.e., the independence of elements and the identity of their distributions have been lost. In fact, however, the operation of odering — this simple trick — leads to far-reaching mathematical consequences through which new results in the probability theory and efficient techniques in mathematical statistics could be developed.

4.2.2. DISTRIBUTION OF THE ORDERED SAMPLE ELEMENTS

Let the distribution function of a continuous random variable X be $F(x)$. Now the distribution of the ordered sample (II) will be derived from $F(x)$.

It is the distribution of the largest element in the ordered sample that is the easiest to find. Let the cumulative distribution function of the largest sample element X_n^* be denoted by $F_{nn}(x)$:

$$F_{nn}(x) = P(X_n^* < x). \tag{4.26}$$

Obviously, the largest observed value X_n^* can be less than x only if each observed value is less as well. Since the elements of sample (I) are independent and have the same distribution that is the distribution function of each sample element is $F(x)$ it follows that

$$P(X_1 < x,\ X_2 < x,\ \ldots,\ X_n < x) =$$

$$= P(X_1 < x)P(X_2 < x)\ldots P(X_n < x) = [F(x)]^n$$

so that

$$F_{nn}(x) = P(X_n^* < x) = [F(x)]^n. \tag{4.27}$$

The derivation of the distribution of the smallest sample element X_1^* can also be performed easily. If the distribution function of the smallest sample element X_1^* is denoted by $F_{n,1}(x)$ then

$$1 - F_{n,1}(x) = P(X_1^* \geqq x) = [1 - F(x)]^n.$$

This is because an event $\{X_1^* \geqq x\}$ can occur only if each observed value is greater than or equal to x.

$$P(X_1 \geqq x,\ X_2 \geqq x,\ \ldots,\ X_n \geqq x) = P(X_1 \geqq x)\ldots P(X_n \geqq x) = [1 - F(x)]^n$$

so that

$$F_{n,1}(x) = 1 - [1 - F(x)]^n. \tag{4.28}$$

Denote by $F_{n,k}(x)$ the distribution function of the k-th element X_k^* in the ordered sample. Obviously, an event $\{X_k^* < x\}$ can occur either if there are k observed values less than x and the other $n-k$ observed values are greater or if there are $k+1$ observations less than x and another $(n-k-1)$ observations greater or if each one is less than x, etc. These events are mutually exclusive so that the probability of event $\{X_k^* < x\}$ is given by the sum of their probabilities.

The probability that out of n independent observed values there will be exactly i ones less than x and the other $n-i$ values will be greater is, according to the binomial distribution, the following:

$$\binom{n}{i} F^i(x)[1 - F(x)]^{n-i}.$$

Therefore

$$P(X_k^* < x) = F_{n,k}(x) = \sum_{i=k}^{n} \binom{n}{i} [F(x)]^i [1 - F(x)]^{n-i}. \tag{4.29}$$

The distribution function $F_{n,k}(x)$ can be written in other form, too. Such an event that the k-th sample element X_k^* in the ranked series is less than x can occur in the following way.

If t denotes an optional value $t<x$ the probability that out of n independent sample elements one falls into the narrow interval $(t-\Delta t, t)$ is $nf(t)\cdot\Delta t$; at the same time $k-1$ values are less than t and the rest, $(n-k)$ values, are greater than t. Probability of the latter is:

$$\binom{n-1}{k-1}[F(t)]^{k-1}[1-F(t)]^{n-k}.$$

By virtue of the rule that the probability of joint occurrence of independent events is the product of their probabilities this probability is

$$n\binom{n-1}{k-1}[F(t)]^{k-1}[1-F(t)]^{n-k}f(t)\cdot\Delta t.$$

If now the interval $(-\infty, x)$ is divided into adjacent intervals with lengths of Δt then, by adding up the probabilities of the aforementioned mutually exclusive events one can obtain that

$$F_{n,k}(x) \approx \sum_k n\binom{n-1}{k-1}[F(t)]^{k-1}[1-F(t)]^{n-k}f(t)\cdot\Delta t \to \tag{4.30}$$

$$\to n\binom{n-1}{k-1}\int_{-\infty}^{x}[F(t)]^{k-1}[1-F(t)]^{n-k}f(t)\,dt.$$

From (4.30) the density function of the k-th element X_k^* in the ordered sample is obtained by derivation:

$$f_{n,k}(x) = n\binom{n-1}{k-1}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x). \tag{4.31}$$

Now an important transformation utilized frequently in practice is referred to. Let $F(x)$ be a function increasing strictly monotonically and introduce the new variable $u=F(t)$; thus $du=f(t)dt$ and the formulae above will take the following forms:

$$F_{n,1}(x) = n\int_0^{F(x)}(1-u)^{n-1}\,du \tag{4.32}$$

$$\vdots$$

$$F_{n,k}(x) = n\binom{n-1}{k-1}\int_0^{F(x)}u^{k-1}(1-u)^{n-k}\,du \tag{4.33}$$

$$\vdots$$

$$F_{n,n}(x) = n\int_0^{F(x)}u^{n-1}\,du. \tag{4.34}$$

So if X is distributed uniformly in the interval $(0, 1)$ — let now its distribution be denoted by $G(y)$ — that is if

$$G(y) = \begin{cases} 0 & \text{if } y<0 \\ y & \text{if } 0\leqq y<1 \\ 1 & \text{if } y\geqq 1 \end{cases}$$

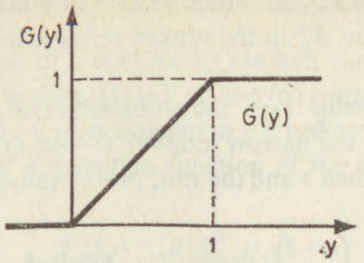


Figure 44

then the distribution functions written above will take simpler forms as follows:

$$
\begin{aligned}
&G_{n,1}(y) = n \int_0^y (1-u)^{n-1}\,du = 1-(1-y)^n \\
&\vdots \\
&G_{n,k}(y) = n \binom{n-1}{k-1} \int_0^y u^{k-1}(1-u)^{nk}\,du \\
&\vdots \\
&G_{n,n}(y) = n \int_0^y u^{n-1}\,du = y^n.
\end{aligned}
\tag{4.35}
$$

Concludingly, in this case the element Y_k^* of the ordered sample follows beta distribution with parameter $(k, n-k+1)$. The expected value and variance of the k-th element in the ordered sample are found as

$$E(Y_k^*) = \frac{k}{n+1}; \quad D^2(Y_k^*) = \frac{k(n-k+1)}{(n+1)^2(n+2)}.$$

Furthermore, it is worth mentioning that the distribution of the sample range

$$R_n = X_n^* - X_1^*$$

can be determined easily; it has the form

$$W_n(r) = n \int_{-\infty}^{\infty} [F(r+u)-F(u)]^{n-1} f(u)\,du. \tag{4.36}$$

Note that if the distribution function of X is the strictly monotonic $F(x)$ and if $Y=F(X)$ the random variable Y is distributed in interval (0, 1) uniformly since

$$P(Y < y) = P[F(X) < y] = P\big(X < F^{-1}(y)\big) = F[F^{-1}(y)] = y. \tag{4.37}$$

Thus, within interval (0, 1) the series of random variables $F(X_1), F(X_2), \ldots, F(X_n)$ can be taken as a statistical sample related to a random variable with uniform distribution.

The joint distribution for pairs, triads, etc. constituted by the elements of an ordered sample can also be derived (see, e.g., [*A*. 25]).

Due to the fact that exponential distribution plays a highly important part in flood hydrology consider now in somewhat more detail the distribution of elements belonging to an ordered sample

(I) $$X_1^* < X_2^* < \ldots < X_n^*$$

derived from a statistical sample

(II) $$X_1, X_2, \ldots, X_n$$

which relates to an exponentially distributed random variable X.

As it is known (see, e.g., [A. 22]) a characteristic feature of exponential distribution is that

$$P(X < y+x \mid X \geqq y) = P(X < x) \quad (\text{if } x > 0,\ y > 0).$$

Namely, since $F(x)=1-e^{-\lambda x}$ one may write that

(4.38) $$P(X < x+y \mid X \geqq y) = \frac{F(x+y)-F(y)}{1-F(y)} =$$

$$= \frac{1-e^{-\lambda(x+y)}-[1-e^{-\lambda y}]}{e^{-\lambda y}} = \frac{e^{-\lambda y}-e^{-\lambda x}\cdot e^{-\lambda y}}{e^{-\lambda y}} = 1-e^{-\lambda x}.$$

First the distribution of differences $X_{k+1}^*-X_k^*$ are calculated from the ordered sample (I). Then, maintaining the condition that $X_k^*=y$, consider the following probability:

(4.39) $$P(X_{k+1}^*-X_k^* \geqq x \mid X_k^* = y) = P(X_{k+1}^* \geqq x+y \mid X_k^* = y).$$

The event in the right hand side of Eq. (4.29) means that having the condition $X_k^*=y$ each of the random variables $X_{k+1}^*, X_{k+2}^*, \ldots, X_n^*$ is greater than $x+y$. This means at the same time that under the condition $X_k^*=y$ there are exactly $(n-k)$ sample elements greater than $x+y$. As in sample (II) the sample elements are independent identical exponentially distributed random variables and as

(4.40) $$[P(X \geqq x)]^{n-k} = e^{-(n-k)\lambda x}$$

the conditional distribution function of differences $X_{k+1}^*-X_k^*$ under the condition that $X_k^*=y$ is

(4.41) $$P(X_{k+1}^*-X_k^* < x \mid X_k^* = y) = 1-e^{-(n-k)\lambda x}.$$

As it can be seen the value of the conditional distribution function (4.41) doesn't depend on y so that (4.41) simultaneously provides the unconditional distribution function of $X_{k+1}^*-X_k^*$.

The validity of this statement follows from the theorem of total probability, too (see: Section 1.1.5):

Denote by $G_k(y)$ the distribution function of the random variable X_k^*. So

$$P(X_{k+1}^*-X_k^*<x)=\int_0^\infty P(X_{k+1}^*-X_k^*<x|X_k^*=y)\,dG_k(y)= \tag{4.42}$$

$$=[1-e^{-(n-k)\lambda x}]\int_0^\infty dG_k(y)=1-e^{-(n-k)\lambda x}.$$

$\left(\int_0^\infty dG_k(y)\right.$ is, of course, equal to one since $G_k(y)$ is a distribution function.$\Big)$ Thus Eq. (4.42) shows that the differences $X_{k+1}^*-X_k^*$ are also distributed exponentially with expected value of

$$\frac{1}{(n-k)\lambda} \quad (k=1,2,\ldots,n-1). \tag{4.43}$$

The distribution of the smallest element X_1^* of the ordered sample is exponential as well; this can be realized plainly by introducing an auxiliary variable $X_0^*=0$ but the same can be conceived by considering Eq. (4.42):

$$P_{n,1}(x)=1-[1-F(x)]^n=1-e^{-n\lambda x}. \tag{4.44}$$

So the expected value of X_1^* is equal to $\frac{1}{n\lambda}$. Based on the foregoing it can be seen that all the differences

$$\delta_{k+1}=(n-k)(X_{k+1}^*-X_k^*) \quad (k=0,1,\ldots,n-1) \tag{4.45}$$

are distributed exponentially with expected value of $1/\lambda$. It can also be realized easily that the random variables $\delta_1,\delta_2,\ldots,\delta_n$ as a whole are independent since the conditional probability

$$P(X_{k+1}^*-X_k^*<x|X_1^*=y_1,X_2^*-X_1^*=y_2,\ldots,X_k^*-X_{k-1}^*=y_k) \quad (y_j\geqq 0) \tag{4.46}$$

doesn't depend on the values of $y_1,y_2,\ldots,y_k$. This is because Eq. (4.46) is equivalent to the probability

$$P(X_{k+1}^*-X_k^*<x|X_k=y_1+\ldots+y_k)=1-e^{-(n-k)\lambda x} \tag{4.47}$$

which, as it was seen, is independent of the condition. Because, by virtue of Eq. (4.45),

$$X_k^*=(X_1^*-X_0^*)+(X_2^*-X_1^*)+\ldots+(X_k^*-X_{k-1}^*)=$$

$$=\frac{\delta_1}{n}+\frac{\delta_2}{n-1}+\ldots+\frac{\delta_k}{n-k+1}$$

the consequence is that

$$E(X_k^*)=E\left(\frac{\delta_1}{n}\right)+E\left(\frac{\delta_2}{n-1}\right)+\ldots+E\left(\frac{\delta_k}{n-k+1}\right)= \tag{4.48}$$

$$=\frac{1}{\lambda}\left(\frac{1}{n}+\frac{1}{n-1}+\ldots+\frac{1}{n-k+1}\right).$$

The expected value of the largest element of an ordered sample, derived specially from exponential distribution, is

$$E(X_n^*) = \frac{1}{\lambda}\left(\frac{1}{n}+\frac{1}{n-1}+\ldots+\frac{1}{2}+1\right) \approx \frac{\ln n}{\lambda}. \tag{4.49}$$

Furthermore,

$$D^2(X_k^*) = D^2\left(\frac{\delta_1}{n}\right)+\ldots+D^2\left(\frac{\delta_k}{n-k+1}\right) = \tag{4.50}$$

$$= \frac{1}{\lambda^2}\left(\frac{1}{n^2}\right)+\frac{1}{(n-1)^2}+\ldots+\frac{1}{(n-k+1)^2}.$$

The variance of the largest element belonging to exponential distribution is

$$D^2(X_n^*) = \frac{1}{\lambda^2}\left(1+\frac{1}{2^2}+\ldots+\frac{1}{n^2}\right) \approx \frac{\pi}{6\lambda^2}. \tag{4.51}$$

Based on formula (4.27), when exponential distribution is dealt with, the distribution function of the largest element of an ordered sample can be found as

$$P(X_n^* < x) = F_{nn}(x) = (1-e^{-\lambda x})^n.$$

Introducing the notation $x = E(X_n^*)+z \approx \frac{\ln n}{\lambda}+z$ the limiting relation

$$G(z) = P\left(X_n^* < \frac{\ln n}{\lambda}+z\right) = (1-e^{-\lambda z-\ln n})^n = \left(1+\frac{-e^{-\lambda z}}{n}\right)^n \to e^{-e^{-\lambda z}} \tag{4.52}$$

is obtained which is called commonly extremal distribution of the first kind.

4.2.4. THE DISTRIBUTION OF THE LARGEST EXCEEDANCES

In general, from the viewpoint of flood control in a given river, it is primarily the distribution of the annual maximum stages that is considered essential. In our opinion the distribution of total exceedances in the individual seasons may also provide useful information. The distribution of seasonal maximum exceedances above a level c chosen suitable from the viewpoint of flood control seems to be especially important. Let be examined therefore the probability distribution of maximum exceedances observed in a given time interval $[0, t)$; this will provide supplementary information on the regime of floods.

As it was seen in Section 4.1.2 if, by a suitable choice, the level c was high enough the distribution of exceedances would be in general exponential (at least as far as rivers covered here, e.g., the Tisza river, are concerned but the same is probable for other rivers with medium flow rates, too). Generally, the number of exceedances can be approximated by Poisson distribution. More than one exceedance in the given period $[0, t)$ can, of course, well occur. Suppose that the number of exceedances

observed in the given interval $[0, t)$ is ν and let these exceedances be denoted by

$$X_1, X_2, \ldots, X_\nu$$

where ν itself is also a random variable. Denote by Z_t the magnitude of maximum exceedance above the level c within an interval $[0, t)$ that is

$$Z_t = \sup_{1 \leq i \leq \nu} \{X_1, X_2, \ldots, X_\nu\}. \tag{4.53}$$

Find the probability of events $\{Z_t < x\}$. When there is no exceedance within interval $[0, t)$, i.e., when $\{\nu = 0\}$ then $Z_t = 0$ automatically. So, if ν has Poissonian distribution whose parameter is λt it follows that

$$P(Z_t = 0) = P(\nu = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}. \tag{4.54}$$

Consequently, if the probability distribution of flood events is known (see Section 2.2.6) the probability of the event $\{Z_t = 0\}$ is known as well. Therefore, it is the conditional distribution function of the random variable Z_t that is to be derived by keeping in mind the condition that $\{\nu > 0\}$. Denote by

$$F_t(x) = P(Z_t < x | \nu > 0)$$

the conditional distribution function of the maximum exceedances Z_t. According to Todorovic and Zelenhasic [B. 49] $F_t(x)$ can be derived as follows:

Let the sample elements $X_1, X_2, \ldots, X_\nu$ be arranged in increasing order of magnitude that is let be formed the ordered sample

$$X_1^* < X_2^* < \ldots < X_\nu^*$$

where the number of sample elements, ν, is also a random variable. Thus the distribution function $F_t(x)$ is the distribution function of the largest element of an ordered sample composed of a random number of elements, under the condition that exceedance(s) did occur that is at least one sample element does exist.

The possible values of the random variable ν are the non-negative integers $0, 1, 2, \ldots$ and the probabilities by which these are taken by ν are $P(\nu = 0)$, $P(\nu = 1)$, $\ldots$, . The probability of the joint occurrence of events $\{Z_t < x\}$ and $\{\nu > 0\}$ is

$$P(\{Z < x\} \cap \{\nu > 0\}) = \sum_{k=1}^{\infty} P(\{Z_t < x\} \cap \{\nu = k\}) = \tag{4.55}$$

$$= \sum_{k=1}^{\infty} P(Z_t < x | \nu = k) P(\nu = k).$$

Resting on the theory of ordered samples (Section 4.2.2) it is easy to see that if $H(x)$ is the distribution function of each of the random variables $X_i (i = 1, 2, \ldots, \nu)$ then

$$P(Z_t < x | \nu = k) = H^k(x)$$

since in this case Z_t is the largest element of an ordered sample containing k elements (see Eq. (4.27)). So it follows that

$$F_t(x) = P(Z_t < x | v > 0) = \frac{\sum_{k=1}^{\infty} H^k(x) P(v = k)}{P(v > 0)}.$$

When the number v of exceedances is distributed according to the Poisson law, having expected value λt in the given interval $[0, t)$ then

$$F_t(x) = \frac{\sum_{k=1}^{\infty} H^k(x) \frac{(\lambda t)^k}{k!} e^{-\lambda t}}{1 - e^{-\lambda t}} = \tag{4.56}$$

$$= \frac{e^{-\lambda t[1-H(x)]} - e^{-\lambda t}}{1 - e^{-\lambda t}}.$$

From this equation it can also be seen that if $H(0)=0$ then $F_t(0)=0$ and $F_t(+\infty)=1$.

As it was seen in Section 4.1.2 the exceedances observed at the gauges of the Tisza river is distributed exponentially that is $H(x)=1-e^{-\beta x}$ $(\beta>0)$. So

$$F_t(x) = \frac{e^{-\lambda t e^{-\beta x}} - e^{-\lambda t}}{1 - e^{-\lambda t}}. \tag{4.57}$$

It will be seen later that for instance for the Tisza river the distribution of maximum exceedances fits this conditional distribution function rather well.

It was shown in Section 2.2.3 that if the distribution of the random variables $X_1, X_2, \ldots, X_v$ was exponential defined by a cumulative distribution function $H(x)=1-e^{-\beta x}$ the random variables

$$\delta_1 = vX_1^*, \quad \delta_2 = (v-1)(X_2^* - X_1^*), \ldots, \tag{4.58}$$

and

$$\delta_k = (v-k+1)(X_k^* - X_{k-1}^*)$$

would be independent and distributed exponentially, with expected value $E(\delta_i)=1/\beta$ and variance $D^2(\delta_i)=1/\beta^2$ $(i=1, 2, \ldots, v)$. As

$$X_k^* = \frac{\delta_1}{v} + \frac{\delta_2}{v-1} + \ldots + \frac{\delta_k}{v-k+1} \tag{4.59}$$

it is involved that when having exponential distribution the expected value of the k-th element in the ordered sample is

$$E(X_k^*) = \frac{1}{\beta}\left(\frac{1}{v} + \frac{1}{v-1} + \ldots + \frac{1}{v-k+1}\right). \tag{4.60}$$

Utilizing this relationship the conditional expected value of Z_t, keeping in mind the stipulation that within the period $[0, t)$ the number of exceedances was k, can be given as

$$E(Z_t|\nu = k) = \frac{1}{\beta}\sum_{i=1}^{k}\frac{1}{i}. \tag{4.61}$$

Because the random variables $\delta_1, \delta_2, \ldots, \delta_k$ are independent it follows that

$$D^2(X_k^*) = \frac{1}{\beta^2}\left[\frac{1}{\nu^2}+\frac{1}{(\nu-1)^2}+\ldots+\frac{1}{(\nu-k+1)^2}\right] \tag{4.62}$$

so that

$$D^2(Z_t|\nu = k) = \frac{1}{\beta^2}\sum_{k=1}^{k}\frac{1}{i^2}.$$

The unconditional expected value of the largest random variable Z_t is equal to the expectation of its conditional expected value (see Section 1.3.4) that is

$$E(Z_t) = E[E(Z_t|\nu)] = \sum_{k=1}^{\infty}E(Z_t|\nu = k)P(\nu = k) = \tag{4.63}$$

$$= \frac{e^{-\lambda t}}{\beta(1-e^{-\lambda t})}\left[\sum_{1}^{\infty}\frac{(\lambda t)^k}{k!}\left(\sum_{i=1}^{k}\frac{1}{i}\right)\right].$$

The numerical calculation of Eq. (4.63) requires usually computer. Here an attempt is made to derive a lower and an upper limit for $E(Z_t)$.

To obtain a lower limit it is sufficient to consider that the expected value of the largest exceedance may not be less than the expected value of all (any) exceedances; now, since exceedance X follows exponential distribution having a distribution function $H(x)=1-e^{-\beta x}$ where $E(X)=1/\beta$, this involves that

$$E(Z_t) \geqq 1/\beta. \tag{4.64}$$

To derive the upper limit for $E(Z_t)$ note that $1+\frac{1}{2}+\ldots+\frac{1}{k}-\ln(k+1)\to\gamma$ (const.) where $\gamma=0.577\ldots$, the so-called Euler constant. Thus

$$\sum_{i=1}^{k}\frac{1}{i} \approx \ln(k+1)+\gamma < k \quad (\text{if } k \geqq 2)$$

so that

$$E(Z_t) = \frac{e^{-\lambda t}}{\beta(1-e^{-\lambda t})}\left[\sum_{k=1}^{\infty}\frac{(\lambda t)^k}{k!}\left(\sum_{i=1}^{k}\frac{1}{i}\right)\right] \leqq \tag{4.65}$$

$$\leqq \frac{e^{-\lambda t}}{\beta(1-e^{-\lambda t})}\left[\lambda t+\lambda t\sum_{l=1}^{\infty}\frac{(\lambda t)^l}{l!}\right] =$$

$$= \frac{e^{-\lambda t}}{\beta(1-e^{-\lambda t})}[\lambda t+\lambda t(e^{\lambda t}-1)] = \frac{\lambda t}{\beta(1-e^{-\lambda t})}.$$

Consequently,

$$\frac{1}{\beta} \leqq E(Z_t) \leqq \frac{\lambda t}{\beta(1-e^{-\lambda t})} \tag{4.66}$$

where, if $\lambda t \leqq 1$, the right hand side of this inequality is surely an upper limit. (In practice, as far as our own studies are concerned, with interval $[0, t)$ as a quarter of year the condition $\lambda t \leqq 1$ was always satisfied for the expected value of exceedances.)

The unconditional variance of the maximum exceedance Z_t can be calculated by formula

$$D^2(Z_t) = E[D^2(Z_t|\nu)] + D^2[E(Z_t|\nu)] \tag{4.67}$$

which leads to a rather sophisticated relationship when the magnitude of exceedances are distributed exponentially and the number thereof follows Poisson distribution. No complicated calculation is needed to guess but an upper limit for $D^2(Z_t)$. As

$$D^2(Z_t|\nu = k) = \frac{1}{\beta^2} \sum_{i=1}^{k} \frac{1}{i^2} \tag{4.68}$$

and

$$\sum_{i=1}^{\infty} \frac{1}{i^2} \approx \frac{\pi^2}{6} \tag{4.69}$$

it follows that

$$D^2(Z_t) \leqq \frac{\pi^2}{6\beta^2} \quad \text{that is} \quad D(Z_t) \leqq \frac{\pi}{\beta\sqrt{6}}. \tag{4.70}$$

While to calculate the moments of the maximum exceedance Z_t is rather cumbersome and tedious (even if only the expected value and standard deviation is needed) the calculation of distribution quantiles (e.g., of median, lower and upper quantiles, etc.) can be executed easily. Any α-quantile can be obtained by using formula

$$F_t(x) = \frac{e^{-\lambda t e^{-\beta x}} - e^{-\lambda t}}{1-e^{-\lambda t}} = \alpha.$$

It is because formula (4.54) involves that

$$e^{-\lambda t e^{-\beta x}} = \alpha + (1-\alpha)e^{-\lambda t}$$

which yields that

$$e^{-\beta x} = -\frac{\ln[\alpha + (1-\alpha)e^{-\lambda t}}{\lambda t}.$$

From this we get

$$\tilde{x} = -\frac{1}{\beta} \ln\left[\frac{-\ln[\alpha + (1-\alpha)e^{-\lambda t}}{\lambda t}\right]. \tag{4.71}$$

The mode $\hat{X}$ of maximum exceedance Z_t is calculated by means of density function

$$F_t'(x) = f_t(x) = \frac{\lambda t \cdot \beta}{1-e^{-\lambda t}} e^{-[\lambda t e^{-\beta x} + \beta x]} \quad (x \geqq 0). \tag{4.72}$$

A simple calculation yields that

$$f_t'(x) = \frac{\lambda t \beta^2 e^{-(\lambda t e^{-\beta x} + \beta x)}}{1 - e^{-\lambda t}} (\lambda t e^{-\beta x} - 1). \tag{4.73}$$

Hence

$$\check{x} = \frac{\ln \lambda t}{\beta} \quad \text{if} \quad \lambda t > 1. \tag{4.74}$$

The point of inflection of the density function (when $\lambda t > 1$) are at points

$$x_{1,2} = -\frac{1}{\beta} \ln \frac{3 \pm \sqrt{5}}{2\lambda t}. \tag{4.75}$$

In the course of practical applications (in calculations for gauges in the Tisza river, for three month intervals) λt was found to be less than 1. It is easy to realize that in this case, for positive x values, the density function $f_t(x)$ is monotonously decreasing. Namely,

$$f_t'(x) < 0 \quad \text{if} \quad \lambda t e^{-\beta x} < 1 \quad \text{that is if} \quad x > \frac{\ln \lambda t}{\beta}.$$

As when $\lambda t < 1$ then $\ln \lambda t < 0$ and since $\beta > 0$ the mode is

$$\check{x} = 0.$$

4.2.5. KOLMOGOROV—SMIRNOV TYPE LIMITING DISTRIBUTIONS

As it was seen, according to Glivenko's theorem the statistic

$$D_n = \sup_x |F_n(x) - F(x)|$$

tends to zero with probability 1 if n tends to infinity. This means also the convergence of $F_n(x)$ to the theoretical cumulative distribution function $F(x)$. Kolmogorov investigated the problem how fast this convergence is, i.e., with large n, how large absolute difference can be expected between the empirical and theoretical distribution functions. It was shown already by Eq. (2.5) that the order of magnitude of D_n was approximately $1/\sqrt{n}$. Kolmogorov proved that the random variable $\sqrt{n} D_n$ fluctuated around a bounded value and he obtained for the limiting distribution of this random variable the following expression:

$$\lim_{n \to \infty} P(\sqrt{n}\, D_n < z) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 z^2} = K(z). \tag{4.76}$$

Values of the limiting distribution $K(z)$ can be found in Table 6.

The above result of Kolmogorov applies to cases where the number of sample elements is large. However, in the practice of hydrology the records are short frequently. This is the reason why a table is annexed here which contains the critical values of the random variable D_n for relatively small values of n.

As to the exact distribution of the random variable $\sqrt{n}\, D_n$ (with finite n) reference is made to work [A.9].

The distribution of the random variables

$$D_n^+ = \sup_x [F_n(x) - F(x)]$$

and

$$D_n^- = \sup [F(x) - F_n(x)]$$

(one-sided deviations) were analyzed by Smirnov who arrived at the following limiting distribution theorem:

$$\lim_{n\to\infty} P(\sqrt{n}\, D_n^+ < z) = \lim_{n\to\infty} (\sqrt{n}\, D_n^- < z) = 1 - e^{-2z^2} = S(z) \quad (z \geqq 0). \tag{4.77}$$

Table for applying the Kolmogorov one-sample test for two-sided alternative hypothesis

n	0.95	0.99	n	0.95	0.99
8	0.4543	0.5419	21	0.2827	0.3443
9	0.4300	0.5133	22	0.2809	0.3367
10	0.4093	0.4889	23	0.2749	0.3295
11	0.3912	0.4677	24	0.2693	0.3229
12	0.3754	0.4491	25	0.2640	0.3166
13	0.3614	0.4325	26	0.2591	0.3106
14	0.3489	0.4176	27	0.2544	0.3050
15	0.3376	0.4042	28	0.2499	0.2997
			29	0.2457	0.2947
16	0.3273	0.3920	30	0.2417	0.2899
17	0.3180	0.3809	35	0.2243	0.2690
18	0.3094	0.3706	40	0.2101	0.2521
19	0.3014	0.3612	45	0.1984	0.2380
20	0.2941	0.3524	50	0.1884	0.2260

If the observed value of $D_n = \max_x |F_n(x) - F(x)|$ attains or surmounts the value given in the table then the hypothesis that the distribution function of the random variable is $F(x)$ has to be rejected.

The application of Smirnov's limiting distribution doesn't need any special table but a table of natural logarithms.

The mode, expected value and median of the limiting distribution of random variable $\sqrt{n}\, D_n^+$ can be found through simple calculations.

The mode $\hat{z}$ can be obtained from the density function

$$S'(z) = s(z) = 4ze^{-2z^2} \tag{4.78}$$

by solving the equation

$$s'(z) = 4e^{-2z^2} - 16z^2 e^{-2z^2} = 0.$$

With $\tilde{z}=1/2$

$$f(\tilde{z}) = 2e^{-\frac{1}{2}} \approx 2 \cdot 0.6 = 1.2.$$

The expected value $E(\sqrt{n}\, D_n^+)$ will be obtained by means of partial integration

$$E(\sqrt{n}\, D_n^+) = \int_0^\infty z \cdot 4ze^{-2z^2}\, dz = -ze^{-2z^2}\Big|_0^\infty + \int_0^\infty e^{-2z^2}\, dz \tag{4.79}$$

applying the substitution $z=u/2$:

$$\frac{1}{2}\int_0^\infty e^{-\frac{u^2}{2}}\, du = \frac{\sqrt{2\pi}}{4} \approx 0.627.$$

Furthermore, as to the median $\tilde{z}_{1/2}$ its derivation is

$$e^{-2z^2} = 0.5, \quad \tilde{z}_{1/2} \approx 0.6.$$

It can be seen that the rank of magnitudes is

$$\tilde{z} < \tilde{z}_{1/2} < E(\sqrt{n}\, D_n^+)$$

that is mode$<$median$<$mean. (Note that in many cases of continuous one-peaked distributions the sequence of these three numerical characteristics is either mean$<$ $<$median$<$mode or the reverse, as with the distribution above.)

The maximum of the relative deviation

In certain cases beyond the deviation between the empirical and theoretical distribution functions the relative deviation, i.e., the ratio of the maximum deviation and of the theoretical distribution function should also be examined.

The limiting distribution of the relative deviations that is the asymptotic behavior of statistics

$$R_n^+(a) = \sqrt{n} \sup_{a \leqq x} \frac{F_n(x) - F(x)}{F(x)}$$

and

$$R_n(a) = \sqrt{n} \sup_{a \leqq x} \frac{|F_n(x) - F(x)|}{F(x)}$$

was investigated by Rényi [A.22]. (Here a denotes a number for which $F(a)>0$ but at the same time small.) According to Rényi's theorem

$$\lim_{n\to\infty} P[R_n^+(a) < z] = \sqrt{\frac{2}{\pi}} \int_0^{z\sqrt{F(a)[1-F(a)]}} e^{-\frac{t^2}{2}}\, dt, \quad z \geqq 0 \tag{4.80}$$

and

$$\lim_{n\to\infty} P[R_n(a) < z] = \frac{4}{\pi} \sum_{i=0}^{\infty} \frac{(-1)^i}{2i+1} e^{-\frac{(2i+1)^2}{8} \frac{1-F(a)}{2z^2}}, \quad z \geqq 0. \tag{4.81}$$

For these statistics tabulated values can be found, e.g., in [A.22].

It is worth mentioning that the analogous relative deviations for the empirical density function $f_n(x)$ and the theoretical density function $f(x)$ were examined by Tusnády, G. [B.40]; he proved that if $f(x)$ satisfied certain conditions (these can be met, e.g., for the density function of the exponential distribution) and if $f_n(x)$ denoted the empirical density function defined by Eq. (2.17) then, with

$$\delta_n = \sup_{x_0^+ < x < x_m^+} \left| \frac{f_n(x) - f(x)}{f(x)} \right|$$

and

$$\Delta_n = \delta_n \sqrt{2k \log m} - \left(2 \log m - \frac{1}{2} \log \log m \right) \tag{4.82}$$

one could obtain that

$$\lim P(\Delta_n < z) = e^{-\frac{1}{\pi} e^{-z}}. \tag{4.83}$$

This is nothing else but the well-known Type I distribution of extremes.

Concerning the order of magnitudes of the quantities k, m and n the following restrictions should be satisfied:

$$k = [n^k] \quad \text{and} \quad m = \left[\frac{m - 2n}{n^k} \right]$$

$\left(\text{where } 0 < \lambda < 1;\ 0 < k < \frac{2}{3} \lambda;\ \text{see [B. 40]} \right)$.

CHAPTER 5

5.1. THEORY OF STATISTICAL ESTIMATION

5.1.1. PROBLEM OF ESTIMATION

In the statistical practice often occurs that the type of distribution function of a random variable X is known but it contains unknown parameters. This means that in the expression

$$P(X < x) = F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{-\infty}^{x} f(t;\, \theta_1, \theta_2, \ldots, \theta_k)\,dt$$

we have a known $f(x; \theta_1, \ldots, \theta_k)$ with partly unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$.

As it was stated in Section 4.1.4, when dealing with flood waves the X magnitude of exceeding a certain water level c was an exponentially distributed random variable so that the density function of its distribution was

$$f(x) = \alpha e^{-\alpha x}$$

an unknown parameter α which is to be estimated from the statistical sample X_1, $X_2, \ldots, X_n$. Here X_i=peak value minus c.

If the number of exceedance above c during a specified season of year, say in the first quarter, is denoted by ν then, as it will be seen in Section 6.3.3 this ν follows a Poisson distribution:

$$P(\nu = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

where λ is also an unknown parameter which should be estimated by a suitable statistic.

Theoretical considerations frequently suggest that a random variable X is normally distributed with a density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

where m and σ are unknown parameters which should again be estimated from a statistical sample.

Following from the foregoing one of the basic problems of statistics is the following.

Denote by $f(x;\theta)$ the density function of a random variable X (where θ stands for a real parameter or a parameter vector); how to determine the unknown value of θ on the basis of a statistical sample of size n

(I) $$X_1, X_2, \ldots, X_n.$$

Another formulation of this question could be: what kind of statistical function ought to be formed from the sample elements to obtain a statistic whose calculated value would give a good approximation of the parameter θ? Is it possible, at all, to create such a statistic from the sample? It will be seen that it is possible and, what is more, not only by a single method.

As it was seen the sample elements were independent random variables with identical distribution; this means that the distribution of each sample element can be described by using the same distribution function

$$P(X_i < x) = P(X < x) = F(x;\theta), \quad i = 1, 2, \ldots, n.$$

This involves that if a statistical function

$$\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$$

is constructed from the sample (I) then $\hat{\theta}_n$ will be a random variable as well; consequently, $\hat{\theta}_n$ also has some kind of distribution, expected value, standard deviation, etc. A simple example of such a $\hat{\theta}$ statistical function is the sample mean

$$\hat{\theta}_n(x_1, x_2, \ldots, x_n) = \frac{x_1+x_2+\ldots+x_n}{n} = \bar{X}$$

through which the expected value $E(X)=\theta$ of a random variable X is approached (estimated). $\bar{X}$ itself is, of course, a random variable as well, a variable whose expected value equals exactly the parameter $E(X)=\theta$ and whose standard deviation is much more less than σ, the standard deviation of the random variable X, notably $D(\bar{X})=\frac{\sigma}{\sqrt{n}}$. This means that if n is large there is a high probability that the value $\bar{X}$ will be close to the expected value θ.

As it was seen (Section 1.1.2) the $P(A)=p$ probability of a certain event A was approached by the relative frequency k/n of A where k was the number of outcomes of event A in a sequence of n independent experiments. If the random variables X_1, $X_2, \ldots, X_n$ are the indicator variables of the outcomes (events) of this sequence and out of them there are k and $(n-k)$ that take the value 1 and zero, respectively, then

$$\frac{X_1+X_2+\ldots+X_n}{n} = \frac{k}{n} = \hat{p}.$$

This means that to estimate the probability p of an event A is a special case of estimating an expected value.

Now this raises the question how to obtain a statistic $\hat{\theta}_n=\hat{\theta}_n(X_1, \ldots, X_n)$ giving a good approximation to the parameter θ. Several methods do exist to obtain estimators for a given parameter θ. Out of them two methods are described below; the one is the method of moments and the other is the so-called maximum likelihood estimation. Both procedures will be illustrated for given distributions. What a "good" estimator means will be treated in 5.1.3.

a) *Method of moments*

A method of estimating the parameters $\theta_1, \theta_2, \ldots, \theta_n$ occurring in the distribution $F(x; \theta_1, \theta_2, \ldots, \theta_n)$ of a certain random variable X is the so-called method of moments. This method consists of making equal the theoretical moments — expressed as functions of the parameters concerned — and the corresponding empirical moments. In this way an equation system containing the parameters can be obtained which is to be solved for the unknown parameter values.

Consider first the case of the exponential distribution. So the cumulative distribution function of a random variable X is $F(x)=1-e^{-\alpha x}$. As it was seen in Section 2.2.12 in case of exponential distribution

$$M_1(x) = E(X) = \frac{1}{\alpha} = \int_0^\infty xf(x)\,dx.$$

The method described yields that

$$M_1(X) = \frac{1}{\alpha}\int_0^x x\,dF_n(x) = \sum_{i=1}^n \frac{1}{n} X_i^* = \bar{X} \tag{5.1}$$

and so $\frac{1}{\hat{\alpha}}=\bar{X}$ that is $\hat{\alpha}=\frac{1}{\bar{X}}$.

Consider now the case of normal distribution. The density function of a random variable X is

$$f(x) = \frac{1}{\sigma 2\sqrt{\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

where

$$M_1(X) = E(X) = m, \quad \sigma^2 = M_2(X) - M_1^2(X) \tag{5.2}$$

$$M_2(X) = E(X^2) = \sigma^2 + M_1^2(X) = \sigma^2 + m^2. \tag{5.3}$$

Our equations will be

$$M_1(X) = \int_{-\infty}^\infty x\,dF(x) \approx \int_{-\infty}^\infty x\,dF_n(x) = \sum_1^n X_i^* \cdot \frac{1}{n} = \bar{X} = m$$

and

$$M_2(X) = \int_{-\infty}^\infty x^2\,dF(x) = \int_{-\infty}^\infty x^2\,dF_n(x) = \sum_1^n X_i^2 \cdot \frac{1}{n} = \sigma^2 + m^2$$

and hence

$$\hat{m} = \bar{X} \tag{5.4}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_1^n x_i^2 - \bar{x}^2 = \frac{\sum (x_i - \bar{x})}{n} = S_n^2. \tag{5.5}$$

Let be considered now the gamma distribution. If the density function of a random variable X is

$$f(x;\, \alpha, p) = \frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x} \quad (x > 0)$$

then

$$M_1(X) = E(X) = \frac{\alpha^p}{\Gamma(p)} \int_0^\infty x^p e^{-\alpha x}\, dx = \tag{5.6}$$

$$= \frac{1}{\alpha\Gamma(p)} \int_0^\infty (\alpha x)^p e^{-\alpha x}\, d(\alpha x) = \frac{1}{\alpha} \frac{\Gamma(p+1)}{\Gamma(p)} = \frac{p}{\alpha}$$

and

$$M_2(X) = E(X) = \frac{\alpha^p}{\Gamma(p)} \int_0^\infty x^{p+1} e^{-\alpha x}\, dx = \tag{5.7}$$

$$= \frac{1}{\alpha^2\Gamma(p)} \int_0^\infty (\alpha x)^{p+1} e^{-\alpha x}\, d(\alpha x) =$$

$$= \frac{1}{\alpha^2} \frac{\Gamma(p+2)}{\Gamma(p)} = \frac{1}{\alpha^2} \frac{(p+1)\Gamma(p+1)}{\Gamma(p)} = \frac{p^2+p}{\alpha^2}.$$

As the expected value of a distribution may be estimated by the sample mean $\bar{X}$ and the variance by the estimate of variance S_n^2 it follows that

$$M_1(X) = \frac{p}{\alpha} = \bar{X}$$

and

$$M_2(X) - M_1^2(X) = \frac{p}{\alpha^2} \approx S_n^2.$$

From these two equations we obtain

$$\hat{\alpha} = \frac{\bar{X}}{S_n^2} \tag{5.8}$$

and

$$\hat{p} = \frac{\bar{X}^2}{S_n^2}. \tag{5.9}$$

The conclusion is that the parameters contained in the gamma distribution can be estimated by the arithmetic mean and the variance.

b) *Maximum likelihood estimation*

This method and its application is presented through an example.

Suppose that the distribution of a random variable X is exponential with a density function $f(x)=\theta e^{-\theta x}$ where θ is an unknown parameter.

The observations on the random variable X are now written into the formula of the joint density function which is denoted by

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_1^n x_i}. \tag{5.10}$$

The probability that the random vector $(X_1, X_2, \ldots, X_n)$ falls in a small interval $\left(x_1 \pm \frac{\Delta x_1}{2}, x_2 \pm \frac{\Delta x_2}{2}, \ldots, x_n \pm \frac{\Delta x_n}{2}\right)$ of the n-dimensional space is

$$L(x_1, x_2, \ldots, x_n; \theta)\Delta x_1 \Delta x_2 \ldots \Delta x_n. \tag{5.11}$$

Obviously, with different values of θ this probability will have different values as well. Now a fundamental principle of mathematical statistics is that when there are several options to select an unknown parameter the one is chosen by which the probability of the event occurring actually is higher or the highest. This means that such a θ value is looked for which maximizes the probability (2.64). As the value of the maximum doesn't depend on $\Delta x_1 \cdot \Delta x_2 \cdot \ldots \cdot \Delta x_n$ the relation

$$L(x_1, x_2, \ldots, x_n; \theta) = \theta^n e^{-\theta \sum_1^n x} = \max_{\theta}$$

is to be solved.

To simplify calculations, instead of the function $L(x_1, x_2, \ldots, x_n; \theta)$ its logarithm, i.e., $\ln L(x_1, x_2, \ldots, x_n; \theta)$ can be examined which, being a monotonic function of it, will, of course, take its maximum at the same θ value as does the function itself:

$$\ln L(x_1, x_2, \ldots, x_n; \theta) = n \ln \theta - \theta \sum_1^n x_i.$$

Consider now the location of the maximum for which the derivative must be equal to zero:

$$\frac{d \ln L}{d\theta} = \frac{n}{\theta} - \sum_1^n x_i = 0$$

that is

$$\hat{\theta} = \frac{n}{\sum_1^n x_i} = \frac{1}{\bar{x}}. \tag{5.12}$$

Since $\frac{d^2 \ln L}{d\theta^2} = -\frac{n}{\theta^2} < 0$, at $\hat{\theta} = \frac{1}{\bar{x}}$ there is a maximum, indeed, and the same holds for the function $L(x_1, x_2, \ldots, x_n; \theta)$ as well. In case of exponential distribution the statistic $\hat{\theta} = \frac{1}{\bar{x}}$ is called the maximum likelihood estimator of the parameter θ.

When a random variable X is normally distributed with a density function $f(x)=$ $=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$ the joint density function of the sample element is

$$L(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_1^n \frac{(x_i-m)^2}{\sigma^2}}.$$

The logarithm of the likelihood function is

$$\ln L(x; m, \sigma) = \ln \frac{1}{(2\pi)^{n/2}} - n \ln \sigma - \frac{1}{2} \sum_1^n \frac{(x_i-m)^2}{\sigma^2}.$$

Suppose now that the value of parameter σ is fixed, so

$$\frac{\partial \ln L}{\partial m} = \frac{1}{n} \sum_1^n x_i - m = 0$$

that is

$$\hat{m} = \frac{\sum_1^n x_i}{n} = \overline{X}. \tag{5.13}$$

The conclusion is that when the maximum likelihood estimation is applied for a normal distribution then sample mean $\overline{X}$ will be obtained as the estimate of the expected value. As to the variance, after a substitution $m=\overline{X}$, one will obtain that

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_1^n (X_i-\overline{X})^2 = 0$$

that is

$$\sum_1^n (X_i-\overline{X})^2 = n\sigma^2.$$

Hence

$$\hat{\sigma}^2 = \frac{\sum_1^n (X_i-\overline{X})}{n} = S_n^2. \tag{5.14}$$

This means that by using the maximum likelihood estimation for normal distribution the estimate of variance will be obtained as the sample variance.

Now an example is given for the application of the maximum likehood estimation when the probability distribution is discrete.

Let be X distributed according to the Poisson law that is

$$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda > 0, \quad k = 0, 1, 2, \ldots.$$

It will be seen in Section 6.3.3/a that, if the number of exceedances above a given water level c and within a given period (say, in the first quarter) is denoted by X, this variable follows in many cases a Poisson distribution. Choosing from the hydrograph-

ic yearbook n years, the number of exceedances, in the first quarters throughout, the statistical sample

$$X_1 = k_1,\ X_2 = k_2,\ \ldots,\ X_n = k_n$$

will be obtained. Now the likelihood function is

$$(5.14)' \qquad L(x_1, x_2, \ldots, x_n;\ \lambda) = P(X_1 = k_1)P(X_2 = k_2)\ldots P(X_n = k_n) =$$

$$= \prod_{i=1}^{n} \left(\frac{\lambda^{k_i}}{k_i!} e^{-\lambda} \right).$$

Hence

$$\ln L = \sum_{1}^{n} k_i \ln \lambda - \lambda n - \sum_{1}^{n} \ln k_i$$

$$\frac{d \ln L}{d\lambda} = \sum_{i=1}^{n} \frac{k_i}{\lambda} - n = 0$$

$$n\lambda = \sum_{1}^{n} k_i$$

$$(5.15) \qquad \hat{\lambda} = \frac{\sum_{1}^{n} k_i}{n} = \overline{X}.$$

So by using the maximum likelihood method for the Poisson distribution the sample mean $\overline{X}$ will be obtained as the estimate of parameter λ.

Finally, let us examine what kind of estimate will be obtained by using the maximum likelihood method for the $P(A)=p$ probability of an event A.

If out of n independent experiments the outcome A occurs k times $(k \neq 0)$ while the outcome $\overline{A}$ $(n-k)$ times the likelihood function has the form

$$L = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\ln L = k \ln p + (n-k) \ln (1-p) + \ln \binom{n}{k}$$

$$\frac{d \ln L}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k-np}{p(1-p)} = 0.$$

Hence

$$(5.16) \qquad \hat{p} = k/n$$

which means that the maximum likelihood estimate of the unknown probability p is the relative frequency. It will be seen that the estimates derived by the maximum likelihood method possess certain favourable properties.

Finally, it is worth mentioning that the method of least squares is also a fundamental method of statistical estimation. This method will be presented in Chapter 8.

There are different viewpoints to make judgement on the "goodness" of estimation. In general, the intention is to get estimators with the following properties:

a) *Unbiasedness*

The statistic $\hat{\theta}=\hat{\theta}(X_1, X_2, \ldots, X_n)$ is called an unbiased estimator of a parameter θ when the expected value of the random variable $\hat{\theta}$ is equal to θ that is when

$$E(\hat{\theta}) = \theta,$$

for all possible values of θ.

Now it will be proved that the sample mean $\bar{X}=\dfrac{X_1+X_2+\ldots+X_n}{n}$ is an unbiased estimate of $E(X)=m$, the expected value of a random variable X, in case of all distributions possessing a first moment. To realize this statement it is sufficient to consider that each of the independent sample elements $X_1, X_2, \ldots, X_n$ is identically distributed and the relation $P(X_i<x)=P(X<x)$ $(i=1, 2, \ldots, n)$ holds; so, by virtue of Eq. (1.28) which relates to the expected value, it follows that

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\cdot n\cdot m = m. \tag{5.17}$$

This relationship also implies that the relative frequency k/n is an unbiased estimate of the probability $P(A)=p$ of the event A. Namely, if n experiments are performed and to each one an X_i indicator variable of the event A is attached then

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1-p,$$

$$E(X_i) = 1\cdot p+0\cdot(1-p) = p.$$

Then

$$\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n} = \frac{k}{n},$$

$$E(\bar{X}) = \frac{1}{n}\sum_{1}^{n} E(X_i) = \frac{1}{n}\cdot n\cdot p = p.$$

Also in cases where m is estimated by using a weighted arithmetic mean $\left(\sum_{1}^{n} p_i=1\right)$

$$\tilde{m} = p_1X_1+p_2X_2+\ldots+p_nX_n \tag{5.18}$$

an unbiased estimate of $E(X)=m$ will be obtained as well.

It is because

$$E(\tilde{m}) = \sum_{1}^{n} p_iE(X_i) = m\sum_{1}^{n} p_i = m. \tag{5.19}$$

Despite this fact to use the arithmetic mean is more advantageous than to apply any other weighted arithmetic mean. This will be shown in the section dealing with the efficiency of estimators.

If the statistic

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n}\sum_1^n (X_i - X)^2$$

is used to estimate the σ^2 variance of a random variable X this estimator fails to meet the requirement of unbiasedness since

$$E(S_n^2) \neq \sigma^2.$$

Namely, if the statistic S_n^2 is written in the form

$$S_n^2 = \frac{1}{n}\sum_1^n [(X_i - m) - (\overline{X} - m)]^2 =$$

$$= \frac{1}{n}\sum_1^n (X_i - m)^2 - (\overline{X} - m)^2$$

and since

$$E(X_i - m)^2 = D^2(X_i) = \sigma^2$$

and

$$E(\overline{X} - m)^2 = D^2(\overline{X}) = D^2\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) =$$

$$= \frac{1}{n^2} nD^2(X_i) = \frac{\sigma^2}{n}$$

is follows that

$$E(S_n^2) = \sigma^2 - \frac{\sigma^2}{n} = \left(1 - \frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2. \tag{5.20}$$

So the value around which the random variable S_n^2 fluctuates is $\frac{n-1}{n}\sigma^2$ instead of σ^2. The unbiasedness can be reached easily since the corrected variance

$$S_n^{*2} = \frac{n}{n-1} S_n^2 = \frac{\sum_1^n (X_i - \overline{X})^2}{n-1}$$

is already an unbiased estimator of σ^2:

$$E(S_n^{*2}) = \frac{n}{n-1} E(S_n^2) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2. \tag{5.21}$$

As $\lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$ the estimator S_n^2 becomes asymptotically unbiased if n large. Estimators of this type are called asymptotically unbiased estimators. So the maximum likelihood estimation provides asymptotically unbiased estimates for σ^2.

b) *Efficiency*

An important requirement against unbiased estimators is that their fluctuation around parameter θ be as small as possible, i.e., the variance $E(\theta-\hat{\theta})^2$ be the possible smallest. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators for θ and $D^2(\hat{\theta}_1) \leqq D^2(\hat{\theta}_2)$ then it is said that $\hat{\theta}_1$ is a more efficient estimator of θ than $\hat{\theta}_2$, for all possible values of θ.

If among the unbiased estimators of the parameter θ such a $\hat{\theta}_0$ exists whose variance is smaller for all θ values than the variance of any other estimator, $\hat{\theta}_0$ is called the uniformly most efficient unbiased estimator. It can be proved that if an estimator, producing the least variance, exists then this is the only estimator of this type. The index of efficiency of the estimator $\hat{\theta}_0$ is the quotient

$$E_\theta(\hat{\theta}) = \frac{\inf_{\theta_i} D_\theta^2(\hat{\theta}_i)}{D_\theta^2(\hat{\theta}_0)} \tag{5.22}$$

whose value always falls in between 0 and 1 and which is equal to 1 only when $\hat{\theta}_0$ is the best unbiased estimator.

The efficiency of estimators are not always related to the variances of all existing unbiased estimators but sometimes to such a narrower class of estimators which sometimes contains the uniformly most efficient estimator. For instance, if the $E(X)$ expected value of a random variable X is θ then, among the linear estimators of the form

$$\hat{\theta}_1 = \sum_{i=1}^{n} X_i p_i \quad \left(\sum_1^n p_i = 1, p_i \geqq 0\right),$$

the arithmetic mean $\overline{X}$ is the most efficient estimator for which

$$p_i = 1/n \quad (i = 1, 2, \ldots, n).$$

By virtue of the Schvarz inequality

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leqq \sum_1^n a_i^2 \cdot \sum_1^n b_i^2.$$

holds.

Let now be $a_i = p_i$, $b_i = 1$ $(i = 1, 2, \ldots, n)$, then

$$1 = \left(\sum_1^n p_i\right)^2 \leqq n \sum_1^n p_i^2$$

that is $\sum_1^n p_i^2 \geqq \frac{1}{n}$ with equality if and only if $p_i = \frac{1}{n}$, $i = 1, 2, \ldots, n$. Evidently,

$$D^2(\overline{X}) = D^2\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) =$$

$$= \frac{1}{n} \sum_1^n D^2(X_i) = D^2(X_i)$$

and

$$D^2(\hat{\theta}_1) = D^2(x_1 p_1 + \ldots + x_n p_n) = D^2(X_i) \sum_1^n p_i^2 \geqq D^2(X_i). \tag{5.23}$$

So the arithmetic mean, and only this, is among the linear estimators having the smallest variance.

At the same time, as far as the expected value is concerned, the arithmetic mean is not for all kinds of distributions the most efficient estimator. If, e.g., within a certain $[a, b)$ interval a random variable X has the rectangular distribution that is when the distribution function of X is

$$F(x) = P(X < x) = \begin{cases} 0 & \text{if } x \leqq a \\ \dfrac{x-a}{b-a} & \text{if } a < x \leqq b \\ 1 & \text{if } x > b \end{cases}$$

and $X_1^* < X_2^* < \ldots < X_n^*$ is the ordered form of a sample of size n related to X then the statistic

$$M_{1/2} = \frac{X_n^* + X_1^*}{2}$$

i.e. the mean of the largest and smallest sample elements is an unbiased estimator of the expected value $E(X) = \dfrac{a+b}{2}$ that is

$$E(M_{1/2}) = \frac{a+b}{2};$$

and, on the other hand,

$$D^2(M_{1/2}) = \frac{(b-a)^2}{2(n+1)(n+2)} = 0\left(\frac{1}{n^2}\right) \tag{5.24}$$

while

$$D^2(X) = \frac{(b-a)^2}{12n} = 0\left(\frac{1}{n}\right). \tag{5.25}$$

So while the standard deviation of the arithmetic mean has the order of magnitude $1/\sqrt{n}$ that of $M_{1/2}$ (midrange) is $1/n$ only; this means that for uniform distributions midrange is a much more efficient estimator of the expected value than the sample mean.

It is worth mentioning here that for the probability $P(A) = p$ of a certain event A the relative frequency k/n is not only an unbiased estimator thereof but at the same time it has minimum variance among all possible unbiased estimators.

c) *Strongly consistent and consistent estimators*

In general the more the number of sample elements, the higher the accuracy expected in the estimation of distribution parameters.

If $\hat{\theta}_n=\hat{\theta}_n(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of the parameter θ and the variance of $\hat{\theta}_n$ tends to zero as n tends to infinity that is if

$$E[(\hat{\theta}_n-\theta)^2] \to 0 \quad n \to \infty \tag{5.26}$$

then $\hat{\theta}_n$ is said to be a *strongly consistent estimator* of the *parameter* θ.

The statistic $\hat{\theta}_n$ is a *consistent estimator* of θ if the relation

$$\lim_{n\to\infty} P(|\hat{\theta}_n-\theta| > \varepsilon) = 0 \tag{5.27}$$

holds for all $\varepsilon>0$.

By virtue of the Chebyshev inequality — for an unbiased estimator θ_n — the following relation holds:

$$P(|\hat{\theta}_n-\theta| > \varepsilon) < \frac{E(\hat{\theta}_n-\theta)^2}{\varepsilon^2}. \tag{5.28}$$

This means that if with increasing n $E[(\theta_n-\theta)^2]$ tends to zero or, in other words, when the estimator $\hat{\theta}_n$ is strongly consistent then it is consistent as well.

If relation (5.27) is satisfied $\hat{\theta}_n$ is said to converge stochastically to θ. So an estimator converging stochastically to the estimated prarameter is called consistent estimator.

As it was seen the arithmetic mean

$$\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$$

was an unbiased estimator of the expected value $E(X)$ and, on the other hand,

$$D^2(\bar{X}) = \frac{D^2(X)}{n} \to 0 \quad \text{if} \to n$$

so that the arithmetic mean is a strongly consistent estimator of the expected value for all distributions whose standard deviation is finite.

If, however, for a variable X the fourth moment also exists, the corrected estimate of variance, S_n^{*2}, is an unbiased and strongly consistent estimator of σ^2.

A consistent estimator is not necessarily unbiased but it is always asymptotically unbiased. Therefore, when large samples are handled, consistency is a more important property for estimators than unbiasedness.

d) *Cramer—Rao inequality*

As it was seen the variance of a highly consistent estimator tends to zero as n tends to infinity. Now the question arises whether, in the case of a fixed samlpe size n, an estimator can be found whose variance is very small or, conversely, there exists a lower limit for the variance of the unbiased estimator. This question is answered by the so-called Cramer—Rao inequality which states the existence of such a lower limit.

Suppose that the distribution function of a random variable X contains one single real parameter, say θ, and let the statistic $\hat{\theta}_n$ be an unbiased estimator of the parameter

$g(\theta)$. Furthermore, let

$$f(x_1, x_2, \ldots, x_n; \theta) \tag{5.29}$$

be the joint density function of the sample elements $X_1, X_2, \ldots, X_n$. Now the Cramer—Rao inequality, which may also be called the uncertainty relation of mathematical statistics, is valid under certain regularity conditions only, for the proof and for non-regular cases we refer to. The Cramer—Rao inequality expresses that with fixed finite n the variance of any estimator cannot be smaller than the reciprocal value of Fisher's information quantity, times $g'(\theta)^2$:

$$D(\hat{\theta}_n) \geqq \frac{g^2(\theta)^2}{I_n} \tag{5.30}$$

where

$$I_n = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \left(\frac{d \ln f}{d\theta}\right)^2 dx_1 \ldots dx_n = E\left(\frac{1}{f}\frac{df}{d\theta}\right)^2$$

is called the Fisher's information quantity.

We have to remark that relation (5.30) under the regularity conditions is true for a sample with non-independent elements. For independent samples $I_n = nI_1$ is true, i.e., $O\left(\frac{1}{n}\right)$ is the best order of magnitude.

Eq. (5.30) is of great theoretical and practical importance: if, when making estimation by means of a certain unbiased $\hat{\theta}_n$ and the Cramer—Rao lower limit is reached then we have a minimum variance estimator.

Two examples are presented below, in both cases for the statistic $\hat{\theta}_n$ the relation

$$D^2(\hat{\theta}_n) = \frac{[g'(\theta)]^2}{I_n}$$

is satisfied.

a) Denote by X the indicator variable of event A, having probability p:

$$X = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{if event } \bar{A} \text{ occurs} \end{cases}$$

and

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

If out of n experiments event A having probability p occurs k times the joint "density function" of sample elements $X_1, X_2, \ldots, X_n$ is

$$f(X_1, X_2, \ldots, X_n; p) = p^k (1-p)^{n-k}.$$

Since

$$\ln f = k \ln p + (n-k) \ln (1-p)$$

the derivate is

$$\frac{d \ln f}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k - np}{p(1-p)}$$

so that

$$I_n = E\left[\left(\frac{k-np}{p(1-p)}\right)^2\right] = \frac{1}{p^2(1-p)^2} E(k-np)^2 =$$

$$= \frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)}.$$

If now the relative frequency $\hat{p}=\frac{k}{n}$ is applied as an estimator of parameter p this yields that

$$E\left(\frac{k}{n}\right) = \frac{1}{n} E(k) = \frac{np}{n} = p,$$

$$D^2\left(\frac{k}{n}\right) = \frac{p(1-p)}{n} = \frac{1}{I_n}.$$

This relationship indicates that the most efficient estimator for the unknown probability is the relative frequency.

b) Let now X have Poisson distribution with parameter λ and suppose that in the course of n observations concerning X the observed values were $X_1=k_1$, $X_2=k_2$, ..., ..., $X_n=k_n$ (where each of the numbers $k_1, k_2, \ldots, k_n$ is one of the values 0, 1, 2, ...).

The joint "density function" of the sample elements is

$$f(X_1, X_2, \ldots, X_n; \lambda) =$$

$$= \frac{\lambda^{k_1}}{k_1!} \cdot \frac{\lambda^{k_2}}{k_2!} \cdots \frac{\lambda^{k_n}}{k_n!} e^{-n\lambda}$$

$$\ln f = -n\lambda + \sum_{i=1}^{n} k_i \ln \lambda - \sum_{i=1}^{n} \ln k_i!$$

$$\frac{d \ln f}{d\lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} k_i$$

$$E\left[\left(\frac{d \ln f}{d\lambda}\right)^2\right] = \frac{1}{\lambda^2} \sum_{1}^{n} D^2(k_i) = \frac{n}{\lambda^2} = \frac{n}{\lambda} = I_n.$$

If statistic $\bar{\lambda} = \frac{k_1+k_2+\ldots+k_n}{n} = \bar{X}$ is applied to estimate the parameter λ then

$$D^2(\lambda) = \frac{1}{n^2} \sum_{i=1}^{n} D^2(k_i) = \frac{n\lambda}{n^2} = \frac{\lambda}{n} = \frac{1}{I_n}.$$

This result indicates that in case of Poisson distribution, when estimating the expected value, λ, there is no estimator with better efficiency than the arithmetic mean.

The question as to what is the type of distributions with which it may be expected that the lower limit in the Cramer—Rao inequality will be attained is answered in the next section.

e) *Sufficient estimators*

The main aim of mathematical statistics is to obtain all information from the sample $X_1, X_2, \ldots, X_n$ representing a random variable X on the distribution of X. Formulating this from the viewpoint of theory of estimation in other words this means that our intention is to form from the sample elements a statistical function comprising in itself all such information on the estimated parameter which is contained in the sample. A statistical function possessing these properties is called sufficient estimator.

As to the parameter in question, in certain cases the opportunity to find a sufficient estimator to it is given; this will be demonstrated below by a few examples. Since the reader may raise the question how to recognize whether a certain statistic actually does or doesn't contain all information on the parameter concerned, in the course of analysing the examples an attempt will be made to answer this question.

Suppose that a random variable X is normally distributed with an expected value $E(X)=m$ and a fixed standard deviation $D(X)=\sigma$. To estimate the unknown expected value m from the sample $X_1, X_2, \ldots, X_n$ statistic $\overline{X}=\frac{\sum X_i}{n}$ is used which is, as it was seen previously, an unbiased and strongly consistent estimator of the expected value m.

Now the question faced is whether the information on m would not be more by considering the numerical value of each sample element, that is by utilizing the location of the sample along the line, instead of forming a single numerical value from the sample elements, as the same sample mean $\overline{X}$ can belong to an innumerable quantity of samples, see Figure 45.

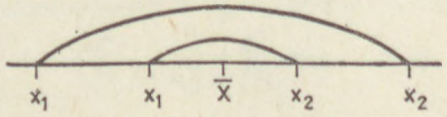


Figure 45

As it was seen, with normal distribution the distribution of the sample mean $\overline{X}$ was normal as well; in our case $E(\overline{X})=m$ and $D(\overline{X})=\frac{\sigma}{\sqrt{n}}$ so that the density function of $\overline{X}$ is

$$f_n(x)=\frac{1}{\sigma}\sqrt{\frac{n}{2\pi}}\,e^{-\frac{n(X-m)^2}{2\sigma^2}}. \tag{5.31}$$

As the individual sample elements X_i are independent and their distribution is $N(m;\sigma)$, their joint density function is

$$f(x_1, x_2, \ldots, x_n;\ m, \sigma)= \tag{5.32}$$

$$=\frac{1}{(\sigma\sqrt{2\pi})^n}\,e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-m)}.$$

Considering the following equation (which can be conceived through a simple calculation)

$$\sum_1^n (x_k-m)^2 = \sum_1^n (x_k-\bar{x})^2+n(\bar{x}-m)^2 \tag{5.33}$$

the joint density function of the sample elements can be written in the form

$$f(x_1, x_2, \ldots, x_n;\ m, \sigma) = \tag{5.34}$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_1^n (x_k-\bar{x})^2} \cdot$$

$$\cdot e^{-\frac{n}{2\sigma^2}(\bar{x}-m)^2} = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-\frac{n}{2\sigma^2}(\bar{x}-m)^2} \cdot$$

$$\cdot \frac{1}{(\sigma\sqrt{2\pi})^{n-1}\sqrt{n}} e^{-\frac{1}{2\sigma^2}\sum_1^n (x_k-\bar{x})^2} .$$

The joint conditional density function of sample elements $X_1, X_2, \ldots, X_n$, under the condition that $\bar{X}=\bar{x}$ is

$$f(x_1, x_2, \ldots, x_n|\bar{X} = x) = \tag{5.35}$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^{n-1}\sqrt{n}} e^{-\frac{1}{2\sigma^2}\sum_1^n (x_k-\bar{x})^2} .$$

The value of this conditional density function doesn't depend on m, consequently, it doesn't contain information on m! This means that having the sample mean $\bar{X}=\bar{x}$ this contains all information concerning m and therefore sample mean $\bar{X}$ is a sufficient statistic on the expected value of m.

Eq. (5.34) indicates that the joint density function of the sample elements can be divided into a product of two factors one of which depends on the sample elements $X_1, X_2, \ldots, X_n$ only, but doesn't depend on the parameter m, while the other one, through statistic $\bar{X}$, depends on the sample elements X_i only and the parameter m also appears in it.

Such a factorization can be executed in all cases where a sufficient statistic can be found to the parameter concerned; hence the conditions to the existence of a sufficient statistic may be formulated as follows:

If the distribution of a random variable X depends on a parameter θ and if $T=T(x_1, x_2, \ldots, x_n)$ is a statistic, and, the joint density function of sample elements can be factorized in the form

$$f(x_1, \ldots, x_n;\ \theta) = \prod_1^n f(x_i;\ \theta) = \tag{5.36}$$

$$= f(x_1, \ldots, x_n|T = t)g(T;\ \theta)$$

then T is a sufficient statistic for the parameter θ.

To check the sufficiency of a statistic T usually the relationship

$$f(x_1, \ldots, x_n | T = t) = \frac{f(x_1, x_2, \ldots, x_n; \theta)}{g(T; \theta)} \tag{5.37}$$

is used: the joint density function of the sample elements is calculated, then the density function $g(T; \theta)$ of T is determined and, finally, the quotient contained in Eq. (5.37) is formed. If the density function obtained in this way depends no longer on θ then the statistic T is a sufficient estimator for θ.

For later purposes (see Section 6.3.3) let now be considered the estimation of expected value for an exponentially distributed random variable X by means of the sample mean $\bar{X}$. Let the density function of X be $f(x)=\lambda e^{-\lambda x}$ and let $X_1, X_2, \ldots, X_n$ be a sample of size n for X. Then the joint density function of the sample elements is

$$f(x_1, x_2, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \tag{5.38}$$

$$= \lambda^n e^{-\lambda \sum_1^n x_i} = \lambda^n e^{-\lambda n \bar{x}}.$$

The density function of the sample mean $\bar{X}$ can be derived easily by using the characteristic function. As the individual sample elements, X_i, are independent and have the same exponential distribution that is for all variables X_i the characteristic function is

$$\varphi_{x_i}(t) = \frac{1}{1 - \frac{it}{\lambda}}$$

and since for the sum of independent random variables the characteristic function is the product of the characteristic functions it follows that for a random variable $n\bar{X}=X_1+X_2+\ldots+X_n$ the characteristic function is

$$\varphi_{n\bar{X}}(t) = \frac{1}{\left(1 - \frac{it}{\lambda}\right)^n}. \tag{5.39}$$

Formula (5.39) involves that the density function of $n\bar{X}$ is

$$f(x) = \frac{\lambda^n}{\Gamma(n)} x^{n-1} e^{-\lambda x}. \tag{5.40}$$

It follows from Eq. (5.40) that for the sample mean $\bar{X}$ the density function is

$$g(x) = nf(nx) = \frac{n\lambda^n}{\Gamma(n)} (nx)^{n-1} e^{-n\lambda x}. \tag{5.41}$$

Hence

(5.42) $$f(x_1, \ldots, x_n|\overline{X} = \bar{x}) =$$

$$= \frac{\lambda e^{-n\lambda x}}{\frac{n\lambda^n}{\Gamma(n)}(n\bar{x})^{n-1}e^{-n\lambda x}} = \frac{(n-1)!}{n^n \bar{x}^{n-1}}.$$

$$(\Gamma(n) = (n-1)!)$$

As it is seen the conditional density function obtained depends no longer on λ so that in the case of exponential distribution $\overline{X}$ is a sufficient statistic to the expected value $E(X)=1/\lambda$ (and, at the same time, to the standard deviation, too).

As to discrete probability distributions, the situation is the same. Consider the case of the Poisson distribution. Also here $\overline{X}$ is a sufficient statistic for the parameter λ.

Let the sample be $X_1=k_1$, $X_2=k_2$, ..., $X_n=k_n$. Now

$$P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) =$$

$$= \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!} e^{-\lambda} = e^{-n\lambda} \frac{\lambda^{\Sigma k_i}}{\prod_{i=1}^{n} k_i!} =$$

$$= e^{-n\lambda} \frac{\lambda^{n\bar{k}}}{\prod_{1}^{n} k_i!}.$$

The distribution of statistic $X_1+X_2+\ldots+X_n=n\bar{k}$ will be determined now by means of the generating function. The generating function for Poisson distribution is:

(5.43) $$G_{X_i}(x) = \sum_{k=0}^{\infty} x^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} e^{\lambda x} = e^{\lambda(x-1)}.$$

By virtue of Eq. (5.43)

$$G_{nX}(x) = [e^{\lambda(x-1)}]^n = e^{n\lambda(x-1)}$$

which is also the generator function of Poisson distribution but with parameter $n\lambda$ so that

$$P(n\overline{X} = n\bar{k}) = \frac{(n\lambda)^{nk}}{(n\bar{k})!} e^{-n\lambda},$$

(5.44) $$P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n|\overline{X} = \bar{k}) =$$

$$= \frac{(n\bar{k})!}{k_1!\, k_2! \ldots k_n!\, n^{n\bar{k}}},$$

which depends no longer on λ.

It can be proved that if statistic $\hat{\theta}_n$ is an unbiased estimator of $g(\theta)$ and if the relationship $D^2(\hat{\theta}_n)=g'^2(\theta)/I_n$ holds then $f(x_1, x_2, \ldots, x_n; \theta)$ can certainly be factorized and $\hat{\theta}_n$ is a sufficient statistic.

So in the case of exponential distribution $\hat{\theta}_n=\overline{X}$ is such an unbiased estimator for $g(\lambda)=1/\lambda$ for which the Cramer—Rao limit will be reached; in this way $\overline{X}$ is a sufficient statistic for $1/\lambda$.

5.1.4. INTERVAL ESTIMATION. CONFIDENCE INTERVALS

In our investigations when estimating a certain unknown θ parameter of a distribution this was done by a single value, the numerical value of $\hat{\theta}=\hat{\theta}(X_1, X_2, \ldots, X_n)$, constructed from the sample elements. This method of estimation is called point estimation as the actual value of the parameter θ is a point on the line and our endeavour is to "hit" or at least approach "well" this point by a $\hat{\theta}$ value calculated from the sample. The $\hat{\theta}$ statistic — let it be supposed to be an unbiased estimator — is a random variable whose values are fluctuating around the true value of the estimated parameter θ. In practice, when an estimation is performed on the basis of a sample containing n elements, by using a certain statistical function $\hat{\theta}=\hat{\theta}(X_1, X_2, \ldots, X_n)$, information is desired on the reliability of this estimation: one may wish to know what a maximum distance may occur by high probability between $\hat{\theta}$ and its true value.

When dealing with a large sample and if $\hat{\theta}$ is an unbiased and consistent estimator, the value of an estimate $\hat{\theta}$ will be close to θ; however, even after $\hat{\theta}$ has been quantified the exact value of θ is not known. Therefore it is desirable to define such a lower and an upper limit, α_1 and α_2, respectively, by which it is assured that the unknown parameter θ will fall by high probability into the interval $[\alpha_1, \alpha_2]$; in this case with these limits and with a predetermined small value ε a relationship given as

$$P(\alpha_1 \leqq \theta \leqq \alpha_2) = 1-\varepsilon$$

will hold. An interval $[\alpha_1, \alpha_2]$ possessing this property is called confidence interval of level $(1-\varepsilon)$. Values of α_1 and α_2 are, of course, also calculated from the statistical sample so that they are random variables as well:

$$\alpha_1 = \alpha_1(X_1, X_2, \ldots, X_n);$$

$$\alpha_2 = \alpha_2(X_1, X_2, \ldots, X_n).$$

Now a few examples are given for confidence intervals.

a) *Confidence interval for the expected value of normal distribution*

Let X be a random variable from normal distribution with unknown expected value m and given standard deviation σ_0. If the expected value m is estimated by the sample mean $\overline{X}$ then the distribution of X is also normally distributed with expected value m

and standard deviation $\sigma_0/\sqrt{n}$; so if, e.g., a level $\varepsilon=0.05$ that is $1-\varepsilon=0.95$ is chosen then

$$P\left(m-2\frac{\sigma_0}{\sqrt{n}} \leqq X < m+2\frac{\sigma_0}{\sqrt{n}}\right) = 0.95. \tag{5.45}$$

The event within the parenthesis can be written in the following form, too:

$$\overline{X}-2\frac{\sigma_0}{\sqrt{n}} \leqq m < \overline{X}+2\frac{\sigma_0}{\sqrt{n}}$$

that is with

$$\alpha_1 = \overline{X}-2\frac{\sigma_0}{\sqrt{n}} \quad \text{and} \quad \alpha_2 = \overline{X}+2\frac{\sigma_0}{\sqrt{n}}$$

a 95 per cent confidence interval is defined to the unknown expected value m.

Confidence interval to the expected value m of a normal distribution when σ is not known.

A question may arise as to the way to define confidence intervals to a constant m when σ is not known. A self-evident idea is to substitute in this case the unbiased estimator

$$S_n^{*2} = \frac{\sum_1^n (X_i-\overline{X})^2}{n-1}$$

for σ^2 that is to examine the interval $\left(\overline{X}-\lambda\frac{S_n^*}{\sqrt{n}},\ \overline{X}+\lambda\frac{S_n^*}{\sqrt{n}}\right)$. Interval

$$\overline{X}-\lambda\frac{S_n^*}{\sqrt{n}} \leqq m < X+\lambda\frac{S_n^*}{\sqrt{n}}$$

is equivalent to inequality

$$-\lambda \leqq \sqrt{n}\cdot\frac{\overline{X}-m}{S_n^*} < \lambda.$$

It can be proved that random variable

$$\tau = \sqrt{n}\,\frac{\overline{X}-m}{S_n^*}$$

is a random variable from Student distribution with $n-1$ degrees of freedom. By using the table of Student distribution (Table 7.7) it is possible to choose such a λ value with which

$$S_{n-1}(\lambda) = P\left(\left|\frac{n(X-m)}{S_n^*}\right| \geqq \lambda\right) = \tag{5.46}$$

$$= 2\int_\lambda^\infty s_{n-1}(u)du = \varepsilon$$

where

$$s_{n-1}(u) = \frac{1}{\sqrt{\pi(n-1)}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{\left(1+\frac{u^2}{n-1}\right)^{\frac{n}{2}}}.$$

In Table T.7 beside n and p the λ values satisfying inequality (5.46) are also included. However, when $n \geqq 30$ and our choice is $\varepsilon = 0.05$ then λ is equal approximately to 2 that is the Student distribution leads to the same values as does the normal distribution with known variance. Interval $\left(\overline{X} - 22\frac{S_n}{\sqrt{n}}, \overline{X} + 2\frac{S_n}{\sqrt{n}}\right)$ is a 95 per cent confidence interval for the unknown expected value m.

An example to illustrate the foregoing is given below.

In Table T.3 the annual maximum stages of River Danube, as observed at Budapest, are shown. The sample mean is

$$X = \frac{\sum_{i=1}^{70} X_i}{70} = 626.66 \text{ cm}.$$

The estimate of standard deviation amounts to

$$S_n^* = \sqrt{\frac{\sum_{1}^{70} (X_1 - X)^2}{69}} = 87.5 \text{ cm}.$$

Hence, for the expected value of annual maximum stages the interval

$$\left(X - 2\frac{S_n^*}{\sqrt{n}}, X + 2\frac{S_n^*}{\sqrt{n}}\right) = (605.86;\ 647.46)$$

is a 95 per cent confidence interval.

b) *Confidence interval for the variance of normal distribution*

As it was seen in Section 5.1.3 the corrected estimate of variance

$$S_n^{*2} = \frac{\sum_{1}^{n} (X_i - X)}{n-1}$$

was an unbiased estimator of σ^2, variance of a random variable X. It can be proved that if a random variable X derives from normal distribution the random variable $n\frac{S_n^{*2}}{\sigma^2}$ follows χ^2 distribution whose parameter is $(n-1)$. By using the table of χ^2 distribution (Table T.5) such c_1 and c_2 values can be determined with which

$$P\left(n\frac{S_n^{*2}}{\sigma^2} < c_1\right) = \frac{\varepsilon}{2} \quad \text{and} \quad P\left(n\frac{S_n^{*2}}{2} > c_2\right) = \frac{\varepsilon}{2}$$

that is with which

$$P\left(c_1 < n\frac{S_n^{*2}}{2} < c_2\right) = 1-\varepsilon.$$

Hence the $(1-\varepsilon)$ level confidence interval for is

$$\left(\sqrt{n}\frac{S_n^*}{\sqrt{c_2}} < \sigma < \sqrt{n}\frac{S_n^*}{\sqrt{c_1}}\right).$$

Now take an example for illustration!

The maximum annual stages of River Danube observed at Budapest in the period 1901/70 are given in Table T.3. In the example $S_n^*=87.5$ cm. As in the Table of χ^2 distribution critical values are given up to n values not more than 30, the table cannot be used if, e.g., $\varepsilon/2=0.02$ is chosen. In cases, however, when $n>30$ the χ^2-distribution can be approached by normal distribution.

Table T.3

Danube river
Annual maximum stages at Budapest

Year	Annual max. cm	Year	Annual max. cm	Year	Annual max., cm
1901	569	1926	737	1951	606
1902	596	1927	596	1952	667
1903	712	1928	628	1953	530
1904	510	1929	469	1954	804
1905	549	1930	622	1955	672
1906	636	1931	542	1956	689
1907	693	1932	576	1957	658
1908	608	1933	529	1958	682
1909	628	1934	415	1959	677
1910	664	1935	584	1960	598
1911	608	1936	599	1961	551
1912	708	1937	624	1962	582
1913	595	1938	598	1963	529
1914	668	1939	704	1964	566
1915	600	1940	788	1965	845
1916	594	1941	670	1966	709
1917	713	1942	689	1967	597
1918	628	1943	689	1968	532
1919	608	1944	754	1969	468
1920	757	1945	654	1970	670
1921	482	1946	652		
1922	564	1947	705		
1923	784	1948	675		
1924	718	1949	681		
1925	645	1950	378		

As the expected value of a random variable from χ^2-distribution with parameter $(n-1)$ is

$E\left(n\frac{S_n^{*2}}{\sigma^2}\right) = n$ and the standard deviation of the same is

$D\left(n\frac{S_n^{*2}}{\sigma^2}\right) = \sqrt{2n}$, the so-called double σ rule that is relationship

$$P\left(n-2\sqrt{2n} \leqq n\frac{S_n^{*2}}{\sigma^2} < n+2\sqrt{2n}\right) \approx 0.95 \tag{5.47}$$

will hold approximately. Relationship (5.47) is equivalent to relation

$$P\left(1-2\sqrt{\frac{2}{n}} \leqq \frac{S_n^{*2}}{\sigma^2} < 1+2\sqrt{\frac{2}{n}}\right) \approx 0.95. \tag{5.48}$$

So interval

$$\left(\frac{S_n^{*2}}{1+2\sqrt{\frac{2}{n}}} \leqq \sigma^2 < \frac{S_n^{*2}}{1-2\sqrt{\frac{2}{n}}}\right) \tag{5.49}$$

is a 95 per cent confidence interval for σ^2. The corresponding 95 per cent confidence interval for the unknown standard deviation is:

$$\left(\frac{S_n^*}{\sqrt{1+2\sqrt{\frac{2}{n}}}}; \frac{S_n^*}{\sqrt{1-2\sqrt{\frac{2}{n}}}}\right) \approx (76.08;\ 150.80).$$

c) *Confidence interval for the λ parameter of the exponential distribution*

As it was seen in the previous section the X exceedances of stages in the Tisza river above the alarm level c followed exponential distribution with distribution function $F(x)=1-e^{-\lambda x}$ or with density function $f(x)=\lambda e^{-\lambda x}$.

When, by using the maximum likelihood estimation, parameter λ is estimated from the statistical sample $X_1, X_2, \ldots, X_n$ by means of statistic

$$\hat{\lambda} = \frac{\sum_1^n X_i}{n} = \overline{X}$$

then point estimation is used. $\overline{X}$ itself is, of course, also a random variable since it is the sum of n random variables:

$$n\overline{X} = X_1+X_2+\ldots+X_n.$$

The sample elements themselves are independent and they have the same distribution given by a cumulative distribution function which is the same as that of the random variable X:

$$P(X_i < x) = 1-e^{-\lambda x}.$$

So the density function of a random variable λX_i is $f(x)=e^{-x}$. By virtue of formula (2.66) the characteristic function of a random variable λX_i is $\varphi_{\lambda X_i}(t)=\frac{1}{1-it}$ and with Eq. (2.65) the characteristic function of a random variable $n\lambda\overline{X}=\lambda X_1+\ldots+\lambda X_n$ is $\varphi_{n\lambda\overline{X}}(t)=\frac{1}{(1-it)^n}$. Recalling those described in Section 2.1.8, from the formula of the characteristic function it may be read out directly that random variable $n\lambda\overline{X}$ derives from gamma distribution and its density function is

$$g_n(x;\ 1)=\frac{1}{\Gamma(n)}X^{n-1}e^{-x}$$

so that

$$P(n\lambda\overline{X}<x)=\int_0^x\frac{u^{n-1}e^{-u}}{(n-1)!}du=G_n(x;\ 1). \tag{5.50}$$

Furthermore, $E(n\lambda\overline{X})=n$, $D(n\lambda\overline{X})=\sqrt{n}$.

In the knowledge of the distribution of random variable $n\lambda\overline{X}$ such limits, $h_n(\varepsilon/2)$ and $h_n(1-\varepsilon/2)$ can be determined with which

$$P(h_n(\varepsilon/2)\leqq n\lambda\overline{X}<h_n(1-\varepsilon/2))=1-\varepsilon \tag{5.51}$$

that is these limits will form a $(1-\varepsilon)$ level confidence interval for the random variable $n\lambda\overline{X}$. Eq. (5.51) is equivalent to relationship

$$P\left[\frac{1}{n\overline{X}}(\varepsilon/2)\leqq\lambda<\frac{1}{n\overline{X}}h_n(1-\varepsilon/2)\right]=1-\varepsilon.$$

By means of the table of χ^2 distribution instead of the limit values $h_n(\varepsilon/2)$ and $h_n(1-\varepsilon/2)$ their doubles can be determined easily. Namely, it is simple to conceive that the distribution of random variable $2n\lambda\overline{X}$ is χ^2, with parameter $2n$ since by virtue of Eq. (5.50) the density function of $2n\lambda\overline{X}$ is

$$\frac{1}{2}g_n\left(\frac{x}{2};\ 1\right)=\frac{1}{2^n\Gamma(n)}X^{n-1}e^{-x} \tag{5.52}$$

which is not else as the density function of a χ^2 distribution whose parameter is $2n$.

Note that if n is large enough then, in accordance with the validity of the theorem of central limiting distribution (Section 2.3.2), the distribution of $\overline{X}$ can be considered normal and its expected value and standard deviation can be given by the following relationships:

$$E(\overline{X})=E(X)=1/\lambda,\ D(\overline{X})=\frac{D(X)}{n}=\frac{1}{\lambda\sqrt{n}}. \tag{5.53}$$

On this basis confidence intervals with any desired level can also be constructed by using the table of standardized normal distribution.

As an illustration let be considered again the exceedances of River Tisza, observed at Szolnok in the first quarters (Table T.1).

In our example $\overline{X}=0.81$ m and $n=41$. In accordance with the validity of the theorem of limiting distribution the distribution of sample mean $\overline{X}$ may be considered normal. As it was seen in Section 5.1.3 in case of an exponential population $E(\overline{X})=1/\lambda$ and $D(\overline{X})=1/\lambda\sqrt{n}$ so that from the table of standardized normal distribution

$$P\left(-1.64 < \frac{\overline{X}-\frac{1}{\lambda}}{\frac{1}{\lambda\sqrt{n}}} < 1.64\right) \approx 0.9. \tag{5.54}$$

(Due to the large variance of exponential distributions in general a safety of 90 per cent has to be regarded sufficient.)

It is easy to see that Eq. (5.54) is equivalent to relation

$$P\left(\frac{1}{\overline{X}}-\frac{1.64}{\sqrt{n}\overline{X}} \leqq \lambda < \frac{1}{\overline{X}}+\frac{1.64}{\sqrt{n}\overline{X}}\right) \approx 0.9$$

where the limits within the parenthesis mean a 90 per cent confidence interval for parameter λ.

In our example:

$$P(0.93 \leqq \lambda < 1.57) \approx 0.9.$$

The confidence interval obtained is rather wide: this indicates that $n=41$ as the number of sample elements is rather small to construct confidence intervals. If a reduced safety might be sufficient the confidence interval may be tightened. For instance the 70 per cent confidence interval for λ is: (1.05; 1.45).

d) *Confidence interval for the empirical distribution function of exceedances*

It was seen in Section 4.1.3 that the empirical distribution function $F_n(x)$ expressed the relative frequency of event $\{X<x\}$, on the basis of n observations. For the same event its probability is given by the theoretical cumulative distribution function $F(x)$. In Section 5.1.3 it was also demonstrated that the empirical distribution function was an unbiased and strongly consistent estimation of the theoretical cumulative distribution function $F(x)$ since

$$E[F_n(x)] = F(x);\ D[F_n(x)] = \sqrt{\frac{F(x)[1-F(x)]}{n}}.$$

As $\sqrt{F(x)[1-F(x)]} \leqq 1/2$ the empirical distribution function $F_n(x)$ satisfies the inequality

$$D[F_n(x)] \leqq 1/2\sqrt{n}. \tag{5.55}$$

In general, relation

$$\left(F_n(x)-\frac{1}{\sqrt{n}} \leqq F(x) < F_n(x)+\frac{1}{\sqrt{n}}\right)$$

is satisfied at a sufficient level of reliability (according to the Chebyshev inequality this is at least 75 per cent but, due to the fact that binomial distribution may be approximated by normal distribution, the true reliability is higher).

As to the exceedances observed in the Tisza river at Szolnok, since $n=41$ and $\sqrt{n} \approx 6.4$, the confidence interval from Eq. (5.5.6) is

$$F_n(x)-0.16 < F(x) < F_n(x)+0.16.$$

which is a rather wide interval. This indicates again that $n=41$ is a rather small sample size to construct confidence intervals.

CHAPTER 6

6.1. TESTING STATISTICAL HYPOTHESES

On the tests of statistical hypotheses — problems in flood hydrology

A few problems are enlisted occurring often in practice; solutions of these problems need the tools of mathematical statistics.

a) The maximum stages of River Danube at Budapest follow normal distribution. May such a statement be made that the expected value of maximum annual stages at Budapest is equal to 625 cm?

b) Is it true that in the Tisza river at Tokaj the probability of a flood wave with a duration of more than 20 days is less than 0.1?

c) Is the expected value of the annual maximum stages of River Tisza at Tokaj and Szeged, respectively, equal?

d) Is the distribution of the annual maximum stages at Budapest identical for the periods 1900/1940 and 1941/1970, respectively? (Has any change occurred in the distribution of annual maxima?)

e) Is it true that at Szeged the magnitude of exceedances above the flood protection alarm level $c = 650$ cm follows an exponential distribution?

In each of these problems one or two random variables are included. The questions formulated above relate to the distribution of random variables concerned or to its parameters. Each of the problems includes a certain assumption and what is questioned is whether this assumption does or doesn't hold. The assumption relating to the distribution of a certain random variable or to one of the parameters of a distribution is called statistical hypothesis.

In a more exact manner, using mathematical terminology, the questions above can be formulated as follows.

a) For a normally distributed random variable X does or doesn't hold that

$$E(X) = 625 \text{ cm}?$$

b) In the Tokaj section out of n flood waves there were k flood waves with longer duration than 20 days. The question, based on the relative frequency k/n, relates to the inequality

$$p < 0.1$$

where p is the unknown probability.

c) If the annual maximum stage at Tokaj is denoted by X_1 and that at Szeged by X_2 does the equality

$$E(X_1) = E(X_2)$$

hold?

d) Denote by X_1 the annual maximum stages at Budapest in the period 1900/1940 and by X_2 the annual maxima in the same section during the period 1941/1970. Is it true that

$$P(X_1 < x) = P(X_2 < x)? \quad (0 < x < \infty).$$

e) Denote the exceedance observed at Szeged above the stage $c = 650$ cm by X. The problem is whether the distribution of X can or cannot be represented by the distribution function

$$P(X < x) = 1 - e^{-\lambda x}$$

(containing a certain fixed but otherwise arbitrarily chosen parameter $\lambda > 0$)?

The questions are to be answered by affirmative or negative answers i.e. decisions are to be made on the acceptance or rejection of the assumptions concerned. In the course of testing a hypothesis statistically the starting point is the assumption that the hypothesis formulated as above is true. This assumption is the so-called null hypothesis which is denoted by H_0.

Another assumption related to a distribution or to a parameter which is in contrast to the null hypothesis is called alternative hypothesis or alternative and it is denoted by H_1.

Consider the null hypothesis under a). Suppose that from previous experience the type of distribution is known: the distribution of that random variable is normal. Furthermore, suppose that the standard deviation of X is known and its value is: $\sigma_0 = 87$ cm.

Then the task is to decide whether the null hypothesis

$$H_0: \; E(X) = 625 \text{ cm}$$

does or doesn't hold. So the problem is confined to the question whether the only unknown parameter of the distribution is or isn't a given value. In this case the distribution is unambiguously defined by the null hypothesis H_0. Such types of hypotheses are called simple hypothesis. Here H_0 is simple hypothesis not because it relates to the value of a single parameter but because the distribution is unambiguously defined by the assumption; in this case the assumption is that the distribution of X is represented by the density function

$$f(x) = \frac{1}{87\sqrt{2\pi}} e^{-\frac{(x-625)^2}{2 \cdot 87^2}}.$$

If to a certain, otherwise true, null hypothesis more than one distribution, i.e., a certain set of distributions can be taken into account then a composite hypothesis is the case. If, e.g., H_0': $620 \leqq E(X) < 630$ and $\sigma_0 = 87$ cm then H_0' is a composite

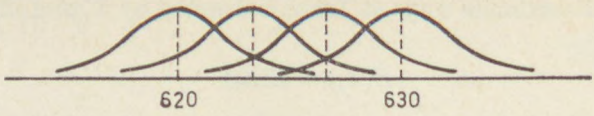


Figure 46

hypothesis because, if satisfied, a whole set of normal distributions with identical standard deviations can be considered, see Figure 46.

Hypothesis $H_0: E(X)=625$ cm is a composite one as well if σ is unknown, i.e., $\sigma\in[0, +\infty]$. In this case the hypothesis will be satisfied for an infinite multitude of normal distributions with given expected value and optional standard deviations, see Figure 47.

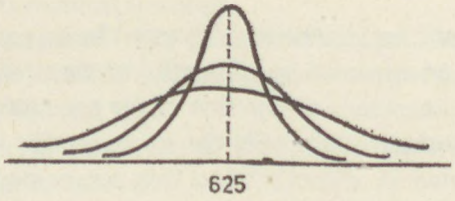


Figure 47

In general terms the problem of simple and composite hypotheses can be illuminated as follows:

Denote by $P(X<x)=F(x;\theta)$ the distribution function of X where θ stands for a parameter or a parameter vector $\theta=(\theta_1, \theta_2, \ldots, \theta_k)$.

Define within the k-dimensional space (parameter space), a specified A_0 region (subset) and regard hypothesis

$$H_0: P(X < x) = F(x; \underline{\theta}) \quad \underline{\theta}\in A_0$$

as a null hypothesis; now H_0 is a simple hypothesis if A_0 is a single point in the k-dimensional space while it is a composite one if more than one point can be found in A_0.

It is a frequent case that, since the null hypothesis is false, in the parameter space a certain A_1 region will need special attention because, e.g., from a practical point of view this region represents the most unfavourable situation. In such a case the null hypothesis is tested against the alternative

$$H_1: P(X < x) = F(x; \underline{\theta}), \quad \underline{\theta}\in A_1$$

which can be equally a simple or a composite hypothesis. It is, of course, assumed that $A_0 \cap A_1 = \emptyset$, i.e., that sets A_0 and A_1 are disjoint.

If there is no specified alternative hypothesis then the common procedure is to test the problem for all of the possible alternative hypotheses, i.e., for the alternatives

$$H_1': P(X < x) = F(x; \underline{\theta}), \quad \underline{\theta} \notin A_0.$$

Parametric and nonparametric problems

In flood hydrology the chosen random variable derives from the practical background (peak value, flood wave duration, maximum flow, etc.) and, consequently, the distribution functions to be found will emerge, in general, from a certain set of distribution functions.

Parametric problem is the case when the parameter space is of finite dimension and the distribution to be found is unambiguously defined by a single point thereof (e.g., dealing with exponential distribution the parameter space is the positive half of the line and the point $\lambda=\lambda_0$ on it indicates which distribution function, $F(x)=1-e^{-\lambda x}$, is the case; in case of normal distribution the point (m_0, σ_0) on the plane defines unambiguously the distribution function concerned, etc.). In such cases the statistical problem relates to one or more parameters. The methods of hypothesis testing related to parameters are called parametric tests. Note that the best known ones relate to the normal distribution. However, it frequently occurs that the type of distribution is also unknown and, e.g., about the stages of a river (at a given site and at a specified point of time) the only thing known is that it is a random variable with continuous distribution. So the set of possible distributions is the set of continuous distributions. In such cases nonparametric problems are dealt with. Such nonparametric problems are, e.g., to decide whether the distribution of two random variables is or isn't the same or to decide whether a given parameter, say the median, of two random variables is or isn't identical. The statistical methods aiming at decisions to be made in nonparametric problems are called nonparametric tests; these include, e.g., the methods of fitting test and the test of homogeneity. In general the nonparametric tests can simply be executed and can "easily" be conceived; the sphere of their application is wide, there is no need to suppose, e.g., the normal or exponential nature of a distribution.

Since in flood hydrology the normal distribution is encountered relatively infrequently, the nonparametric procedures will be discussed in this book in a bit more detail. However, at first a brief overview will be given on the theory of statistical tests and on the parametric tests used most frequently.

6.1.1. GENERALS ON STATISTICAL TEST

In each of the problems enumerated in the previous section under a) through e) a certain question was formulated which could be answered affirmatively or negatively. An affirmative answer represents the acceptance of hypothesis H_0 while a negative answer means its rejection. The task of a statistical test is to provide, on the basis of a statistical sample related to a random variable X included in the problem, opportunity for making decision as to H_0 should be accepted or rejected.

The construction and properties of statistical tests will be described in connection with testing the hypothesis mentioned under a). This example is suitable to draw general conclusions therefrom.

Let the random variable X be normally distributed; its standard deviation is known: $\sigma_0 = 87$ cm. The hypothesis to be tested here is

$$H_0: E(X) = m_0 = 625 \text{ cm}$$

where m_0 is a given value. Consider the statistical sample $X_1, X_2, \ldots, X_n$ consisting of n elements, representing the random variable X.

As it was seen in Section 5.1.3 a rather good estimate of the expected value was the sample mean $\bar{X} = \frac{\sum x_i}{n}$ and, since the case is a normal distribution, $\bar{X}$ is a random variable with normal distribution as well, represented by the unbiased estimator $E(\bar{X}) = m_0$ and standard deviation $D(\bar{X}) = \frac{\sigma_0}{\sqrt{n}}$.

Being the sample mean a random variable, the observed value of $\bar{X}$, due to changes by chance, will probably differ from m_0 even when hypothesis H_0 is true. For making decision it seems to be a rather apparent principle that H_0 may be accepted when $\bar{X}$ is close enough to m_0 while H_0 should be rejected if $\bar{X}$ is far therefrom.

However, the question remains what are the cases where $\bar{X}$ should be regarded to be close to or distant from m_0, respectively. This question can be answered if the distribution of the random variable $\bar{X}$ is known. The starting point is that H_0 is true and then $m_0 = E(\bar{X})$. A well-known property of the normal distribution is that (according to the so-called double σ rule)

$$P\left(m_0 - \frac{2\sigma}{\sqrt{n}} < X < m_0 + \frac{2\sigma}{\sqrt{n}}\right) \approx 0.95. \tag{6.1}$$

This means that, if H_0 is true, $\bar{X}$ will be closer to m_0 than two times the standard deviation by a high (95 per cent) probability and the probability that it would fall outside these limits is 5 per cent only. An equivalent form of the expression above (6.1) is

$$P\left(-2 < \frac{\bar{X} - m_0}{\frac{\sigma_0}{\sqrt{n}}} < 2\right) \approx 0.95. \tag{6.2}$$

Based on the foregoing, as to the acceptance or rejection of hypthesis H_0, the following principle of decision can be established: calculate the statistic $u = \sqrt{n}\,\frac{\bar{X} - m_0}{\sigma_0}$ and if the value of u falls into the range $(-2, +2)$ hypothesis H_0 may be accepted, otherwise it should be rejected. By doing so the decision is made by a probability of 95 per cent or, using the common terminology, at a level of 0.95.

Since, when H_0 is true, the statistic $u = \sqrt{n}\,\frac{\bar{X} - m_0}{\sigma_0}$ will have a *standard normal distribution* with $E(u) = 0$ and $D(u) = 1$, when making decision the level of 0.95 needn't be adhered to. Choice may be made for an optional low value of $\alpha > 0$ and,

by using the table of normal distribution, such an interval $(-u_\alpha, u_\alpha)$ can be defined which will be in concert with relation

$$P\left(-u_\alpha < \sqrt{n}\,\frac{\overline{X}-m_0}{\sigma_0} < u|H_0\right) = 1-\alpha.$$

In this way the level of decision on accepting or rejecting H_0 is $1-\alpha$.

To answer the question under a) let α be chosen as 0.05 (this choice is common in the practice of statistics but sometimes a choice with $\alpha=0.01$ may also be justified).

$$\overline{X} = 626.7\,\text{cm},\ u = \frac{626.7-625}{87}\sqrt{70} \approx 0.17,$$

so that there is no reason to reject H_0 that is H_0 may be accepted.

The statistical test outlined is called *u-test*. In this decision procedure the principle followed is to accept hypothesis H_0 in question of the value of statistics u falls into the interval $(u_{-\alpha}, u_\alpha)$ and to reject H_0 if that value is outside this interval. So from the point of view of decision making the line has been divided into two subsets. Interval $(u_{-\alpha}, u_\alpha)$ is called the region of acceptance while the part of the line outside (i.e. the complementary set of this interval) is called critical region, see Figure 48.

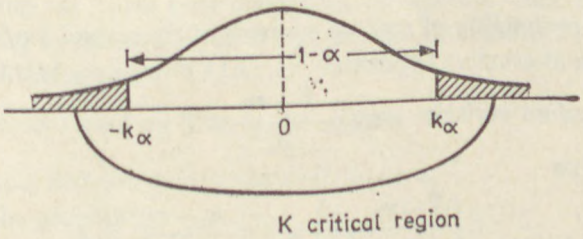


Figure 48

Dealing with statistical tests it can be said that their essence is to choose a critical region. For a statistician the starting point is always an assumption that null hypothesis H_0 is true; then such a critical region K has to be chosen into which, while H_0 is true, the value of the calculated statistic $u=u(X_1, X_2, \ldots, X_n)$ will fall by low probability. If, nevertheless, the actual value of u falls into the critical region K then H_0 should be rejected. The probability by which, while H_0 is true, the value of the test statistic falls into the critical region K is called the size of the test so that in case of

$$P(u \in K) = \alpha$$

the size of the test applied is α.

As it was seen, in the course of applying a u-test the procedure is quite mechanical once the size of the test, the critical region K, have been chosen. If $u \in K$, H_0 is rejected, otherwise it is accepted. When making decisions, mistakes can, of course, also occur. May happen that H_0 is true that is $E(X)=m_0$ and still the value of an u-sta-

tistic (due to chance) falls into the critical region K and, therefore, H_0 is rejected. By doing so an *error of the first kind* is made. So such a mistake which is made when a true hypothesis is rejected is called error of the first kind.

In u-tests the probability of making an error of the first kind is

$$P(|u| > u_\alpha | H_0|) = \alpha$$

that is the probability of such an error is equal to the size of the test. It may also happen that H_0 is false but a certain alternative hypothesis $H_1 : E(X) = m_1 \neq m_0$ is true and still the value of the u-statistic falls into the non-critical region, i.e., into the region of acceptance $(-u_\alpha, u_\alpha)$ and so H_0 is accepted. In this way a so-called error of the second kind is made. The decisions possible in logical sense and both kinds of errors which may be made are summarized in the following scheme:

	Hypothesis H_0	
	accepted	rejected
H_0 is true	Right decision	Error of the first kind
H_0 is false	Error of the second kind	Right decision

Examine the probability of making an error of the second kind in case of u-tests. Suppose that an alternative hypothesis $H_1 : E(X) = m_1 \neq m_0$ is true. Here the expected value of random variable $u = \sqrt{n}\,\dfrac{\overline{X} - m_0}{\sigma_0}$ will no longer be equal to zero since $E(\overline{X}) = m_1$ and so

$$E\left(\frac{\overline{X} - m_0}{\sigma_0}\sqrt{n}\right) = \sqrt{n}\,\frac{m_1 - m_0}{\sigma_0} = \Delta.$$

For a given $E(\overline{X}) = m_1$ the probability of making an error of the second kind is shown in Fig. 49 by the measure of probability shaded.

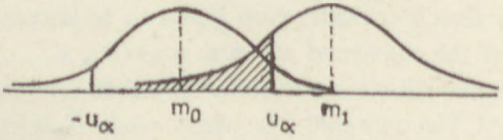


Figure 49

Denote by β the probability of making an error of the second kind, so

$$\beta = P(-u_\alpha < u < u_\alpha | m_1).$$

Obviously, the probability of errors of the second kind will depend on m_1 that is on the magnitude of the expected value belonging to random variable X. Based on Fig. 49 it can easily be realized that if m_1 is close to m_0 (i.e., Δ is small) then H_0 will be

accepted by high probability although it is H_1 that is true; in this way an error of the second kind will be made by high probability. Fortunately, if m_1 is close to m_0 hypothesis H_0 will almost be satisfied so that, in general, its acceptance will not cause a greater trouble.

Concludingly, through the aforementioned choice of the critical region the requirement that the probability of an error of the first kind be equal to the prescribed α can be fulfilled; here α, due to practical considerations, should be chosen as a suitable low value (e.g., 0.05 or 0.01). Dealing with u-tests it is apparent that the less the value of α (i.e., the wider the interval $(-u_\alpha, u_\alpha)$) the higher the value of β that is the probability of making an error of the second kind, beside a given H_1. The situation is the same in general cases as well; for a lower error of the first kind a higher one of the second kind should be paid. The size of test to be applied for testing a given H_0 hypothesis depends on the concrete nature of the problem. Considerations should be made on whether what harm can be caused by making errors of the first or second kind, respectively.

6.1.2. THE POWER FUNCTION

In case of a combined alternative hypothesis H_1 the probability of an error of the second kind depends, as it was seen previously, on whether what the true value of $E(X)=m$ belonging to random variable X is. So the probability of an error of the second kind is a function of m:

$$\beta\colon\ P(|u| < u_\alpha|m) \tag{6.3}$$

if the required level of decision on hypothesis H_0 is $1-\alpha$.

Consider now the probability

$$1-\beta\colon\ P(|u| < u_\alpha|m) \tag{6.4}$$

which represents the probability that the calculated value of statistic u falls into the critical region, provided that $E(X)=m$. This probability is, of course, also a function of m: $\varepsilon_n(m)$. This function is called the *power function* of the test. It is apparent that $\varepsilon_n(m_0)=\alpha$ since the critical region has been chosen in such a manner that this probability by equal to α if $H_0: E(X)=m_0$ is true. Calculate now the value of the power function for a freely chosen value $m \neq m_0$ in case of an u-test. Since with $E(X)=m$ now random variable $\sqrt{n}\,\dfrac{\overline{X}-m}{\sigma_0}$ will have standard normal distribution apply the following transformation:

$$u = \sqrt{n}\,\frac{\overline{X}-m_0}{\sigma_0} = \sqrt{n}\,\frac{\overline{X}-m}{\sigma_0} + \sqrt{n}\,\frac{m-m_0}{\sigma_0} =$$

$$= \sqrt{n}\,\frac{\overline{X}-m}{\sigma_0} + \Delta.$$

Since

$$\varepsilon_n(m) = P(|u| > u_\alpha | m) = 1 - P(-u_\alpha < u < u_\alpha | m) =$$

$$= 1 - P\left(-u_\alpha < \sqrt{n}\,\frac{\overline{X} - m_0}{\sigma_0} < u_\alpha | m\right) =$$

$$= 1 - P\left(-u_\alpha < \sqrt{n}\,\frac{\overline{X} - m}{\sigma_0} + \Delta < u_\alpha | m\right) =$$

$$= 1 - P\left(-u_\alpha - \Delta < \frac{\overline{X} - m}{\sigma_0} < u_\alpha - \Delta | m\right)$$

what is obtained is

$$\varepsilon_n = 1 - [\Phi(u_\alpha - \Delta) - \Phi(-u_\alpha - \Delta)] \tag{6.5}$$

where $\Phi(x)$ denotes the standard normal distribution function.

The shape of power function can be guessed by simple considerations. If $m = m_0$, i.e., $\Delta = 0$ then

$$\varepsilon_n(m_0) = 1 - [\Phi(u_\alpha) - \Phi(-u_\alpha)] = 1 - (1 - \alpha) = \alpha$$

which is, in general, a small value (e.g., $\alpha = 0.05$). But if $m \to \infty$ then

$$\Phi\left(u_\alpha - \frac{m - m_0}{\sigma_0}\sqrt{n}\right) \to \Phi(-\infty) = 0.$$

Furthermore, if $m \to -\infty$ then similarly

$$\Phi\left(-u_\alpha - \frac{m - m_0}{\sigma_0}\sqrt{n}\right) \to \Phi(-\infty) = 0$$

that is

$$\varepsilon_n(m) \to 1 \quad \text{if} \quad m \to \pm\infty$$

which means that the higher the absolute value of $E(X) = m$ the greater the value of the power function that is the closer the value of $\varepsilon_n(m)$ to 1. It is said that if $m \to \pm\infty$ the power of test approaches to unity.

Consequently, the diagram of a power function $\varepsilon(m)$ may be sketched in the form of Figure 50. The figure indicates that the greater the difference between the true expected value $E(X) = m$ and the supposed value, m_0, the higher the probability that the

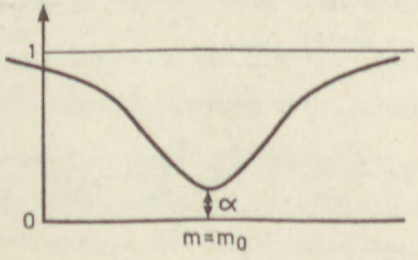


Figure 50

value of the u-statistic falls into the critical region that is the higher the probability of rejecting hypothesis H_0. At the same time in case of $m \to \pm\infty$ the probability $\beta = 1 - \varepsilon(m)$ of an error of the second kind will approach to zero.

Thus, with a given $m \neq m_0$, the higher $\varepsilon_n(m)$ for a given m the better the test; this $\varepsilon_n(m)$ is called *the power of the test* against the given alternative m. However, for m values being close to m_0, $\varepsilon_n(m)$ is close to α that is the power of the test is small. What could be the way to increase the power of a test beside a given m?

Since σ_0 is given and the error of the first kind, α, has been chosen in advance it is the value of n that can be increased. Namely, by considering the formula

$$\varepsilon_n(m) = 1 - \Phi\left(u_\alpha - \frac{m - m_0}{\sigma_0}\sqrt{n}\right) +$$

$$+ \Phi\left(-u_\alpha - \frac{m - m_0}{\sigma_0}\sqrt{n}\right)$$

it can be seen that with $n \to \pm\infty$

$$\Phi\left(u_\alpha - \frac{m - m_0}{\sigma_0}\sqrt{n}\right) \to \Phi(-\infty) = 0$$

$$\Phi\left(-u_\alpha - \frac{m - m_0}{0}\sqrt{n}\right) \to \Phi(-\infty) = 0$$

so that with any given m

$$\lim_{n \to \infty} \varepsilon_n(m) = 1.$$

A test whose power converges to one when $n \to \pm\infty$, for all elements of the alternative hypothesis, is called *consistent test.*

Consequently, dealing with u-tests, if the number of observations will be increased and a certain δ is fixed, the power $\varepsilon_n(m_0 \pm \delta)$ can be approached to unity as close as necessary.

Hitherto, by using a concrete test, the problems discussed were how to determine the probabilities of errors of the first and second kind, respectively, how to construct an alternative with due regard to practical requirements and how to improve the test by increasing the number of sample elements. In the following the way of constructing a parametric test will be formulated in more general terms. The u-test outlined in the foregoing was to serve as illustration only. Below a remark will be made on the notion of critical region.

In case of u-tests the definition of critical region has been formulated as a set of u values for which a hypothesis H_0 should be rejected; this set is the complementary set of interval $\left(m_0 - u_\varepsilon \frac{\sigma_0}{\sqrt{n}}, m_0 + u_\varepsilon \frac{\sigma_0}{\sqrt{n}}\right)$ on the line that is (in the sense of set theory) the union of two infinite intervals, see Figure 51.

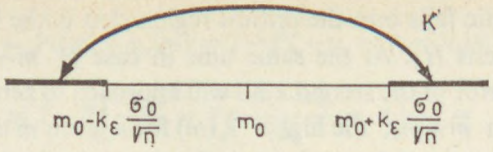


Figure 51

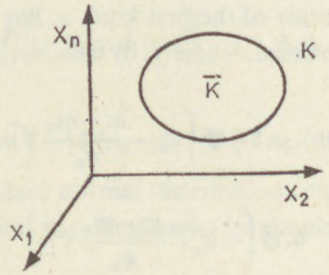


Figure 52

A sample $X_1, X_2, \ldots, X_n$ is a random point in the n-dimensional space, the so-called sample space. Decision on the acceptance or rejection of hypothesis H_0 is to be made through the position of the random point $\vec{X}=(X_1, X_2, \ldots, X_n)$. So, for making decision, a procedure might be to choose such a subset, denoted by K, in the n-dimensional space, which is considered critical region and if $\vec{X}\in K$ hypothesis H_0 would be rejected, see Figure 52. What will be done is essentially the same. What is said is let the critical region K be regarded as a set of all such points $\vec{X} = (X_1, X_2, \ldots, X_n)$ for which statistic

$$u = u(X_1, X_2, \ldots, X_n) = \frac{\overline{X}-m_0}{\sigma_0}\sqrt{n} \tag{6.6}$$

falls into the set $\left(-\infty, m_0-u_\varepsilon\frac{\sigma_0}{\sqrt{n}}\right)\cup\left(m_0+u_\varepsilon\frac{\sigma_0}{\sqrt{n}}, +\infty\right)$.

Since both the n-dimensional space and the subsets thereof are sophisticated, the latter itself, i.e., the relevant two infinite intervals of the line will be called critical region.

6.1.3. UNIFORMLY BEST TEST FOR SIMPLE HYPOTHESES

Now a method will be presented for constructing the best test in the case of simple hypotheses. A test will be called "the best" one if among the tests with size α (i.e., with error of the first kind of probability α) the one will be chosen for which the probability β of making an error of the second kind is the least. The basis for constructing such a test is the so-called Neyman—Pearson lemma (the theorem described below was proved and used first by these two statisticians).

Let the random variable X have a continuous distribution with density function $f(x;\theta)$ where θ is an unknown parameter. (The proof will be scheduled for the real variable X and unknown θ but the theorem also holds when X denotes a vector.) Suppose that a sample of size n for X resulted in the values

$$X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n.$$

Let the null hypothesis be $H_0 : \theta = \theta_0$ while the simple alternative hypothesis: $H_1 = \theta = \theta_1$.

The Neyman—Pearson lemma claims the following: there exists a constant k_α with which

$$K = \left\{(x_1, \ldots, x_n): \frac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)} > k_\alpha\right\}, \tag{6.7}$$

i.e.,

$$\bar{K} = \left\{(x_1, \ldots, x_n): \frac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)} \leqq k_\alpha\right\} \tag{6.8}$$

with $P(K|\theta_0) = \alpha_1$ and K is the best critical region with size α.

The proof of the lemma is relatively simple.

For sake of simplicity let us use the notation

$$L_0 = \prod_{i=1}^{n} f(x_i;\theta_0) \quad \text{and} \quad L_1 = \prod_{i=1}^{n} f(x_i;\theta_1).$$

(The function $L(\theta) = \prod_{i=1}^{n} f(x\ ;\theta)$ is called the likelihood function.)

Since the sample elements are independent random variables with identical distribution, L_0 and L_1 are the joint density functions of the sample elements if H_0 and H_1, respectively, are true.

Let K^* be another critical region having the same size α (see Figure 53).

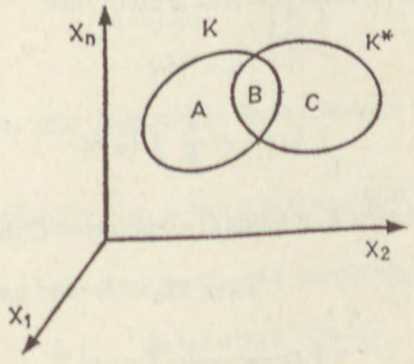


Figure 53

With the above notations

$$\int_K L_0\,dx = \alpha$$

which expresses the probability that the sample point falls into region K is H_0 is true, being the probability of an error of the first kind. Since the size of both K and K^* is α it is involved that

$$\int_K L_0\,dx = \int_{K^*} L_0\,dx = \alpha.$$

As it is seen on the figure $B=K\cap K^*$ so that equality

$$\int_K L_0\,dx = \int_C L_0\,dx \tag{6.9}$$

also holds.

We turn now to the determination of the probabilities of the error of the second kind. These are equal to the probabilities that the sample point falls outside the critical region if H_1 is true. These are equal to 1 minus the probability that the sample point falls within the critical region if H_1 is true, which means that for K and K^*:

$$\beta = 1 - \int_K L_1\,dx;\ \beta^* = 1 - \int_{K^*} L_1\,dx.$$

Consequently:

$$\beta^* - \beta = \int_K L_1\,dx - \int_{K^*} L_1\,dx =$$

$$= \int_{A+B} L_1\,dx - \int_{B+C} L_1\,dx =$$

$$= \int_A L_1\,dx + \int_B L_1\,dx - \int_B L_1\,dx - \int_C L_1\,dx$$

that is

$$\beta^* - \beta = \int_A L_1\,dx - \int_C L_1\,dx.$$

Since set A is a part of set K, at all points of A holds that

$$L_0 \leqq kL_1$$

and so

$$\int_C L_1\,dx \leqq \frac{1}{k}\int_C L_0\,dx.$$

As set C falls outside region K, it holds at all points of C that

$$L_0 \geqq kL_1$$

that is

$$\int_C L_1\,dx \leqq \frac{1}{k}\int_C L_0\,dx.$$

Hence

$$\beta^*-\beta = \int_A L_1\,dx - \int_C L_1\,dx \geqq \frac{1}{k}\int_A L_0\,dx -$$

$$-\frac{1}{k}\int_C L_0\,dx = \frac{1}{k}\left[\int_A L_0\,dx - \int_C L_0\,dx\right].$$

Because, by virtue of relationship (6.9),

$$\int_A L_0\,dx - \int_C L_0\,dx = 0$$

one can obtain that

$$\beta^*-\beta = 0$$

that is the error of the second kind of region K^* is greater than (or equal to) the error of the second kind of region K; this means that no better critical region can be found than K.

Below an example is given for the application of the Neyman—Pearson lemma. Let $f(x;\theta)$ be equal to $\theta e^{-\theta x}(x \geqq 0)$, i.e., let X be distributed exponentially. Let the null hypothesis be: $H_0: \theta=\theta_0$ and the alternative hypothesis $H_1: \theta=\theta_1<\theta_0$.

The likelihood function for $\theta=\theta_0$ is

$$L_0 = \prod_{i=1}^{n} f(x_i;\ \theta_0) = \theta_0^n e^{-\theta_0 \sum_1^n x_i}$$

and for $\theta=\theta_1$

$$L_1 = \prod_{i=1}^{n} f(x_i;\ \theta_1) = \theta_1^n e^{-\theta_1 \sum_1^n x_i}.$$

The critical region K corresponding to equation (6.7) is a set of all those sample points for which

$$\frac{\theta_1^n e^{-\theta_1 \sum_1^n x_i}}{\theta_0^n e^{-\theta_0 \Sigma x_i}} > k$$

that is

$$e^{(\theta_0-\theta_1)\sum_1^n x_i} > k\left(\frac{\theta_0}{\theta_1}\right)^n.$$

Transforming both sides into logarithmic form (Because of the monotonity of the logarithmic function):

$$(\theta_0-\theta_1)\sum_1^n x_1 > \ln k + n(\ln\theta_0 - \ln\theta_1).$$

As $\theta_1<\theta_0$, after dividing both sides by $(\theta_0-\theta_1)$ we obtain for the ciritical region K:

$$\sum_1^n x_i \geqq \frac{\ln k + n(\ln\theta_0 - \ln\theta_1)}{\theta_0-\theta_1}. \tag{6.10}$$

Dividing both sides of this inequality by n we obtain:

$$\bar{x} \geqq \frac{\ln \theta_0 - \ln \theta_1}{\theta_0 - \theta_1} + \frac{\ln k}{n(\theta_0 - \theta_1)} = C. \tag{6.10a}$$

Here the value of k or C should be chosen to get $P(\bar{X} > C | H_0) = \alpha$. But C depends on α and θ_0 only because the distribution of $\bar{X}$ depends on θ_0. This means that we obtained a best test (critical region) between the simple hypothesis $H_0: \theta = \theta_0$ and the composite hypothesis $H_1': \theta < \theta_0$.

Relationship (6.10/a) shows that with $\theta < \theta_0$ the best critical region is the right hand side tail of the distribution, see Figure 54.

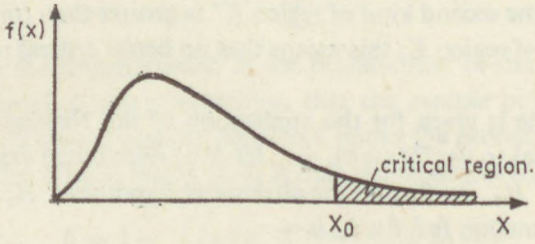


Figure 54

6.2. PARAMETRIC TEST

6.2.1. STUDENT t-TEST

a) *The one-sample case*

Student t test is used to test a hypothesis on the expected value of a normally distributed random variable. The test described in Section 6.1.1 can be applied to check a hypothesis on the expected value of a normally distributed random variable X only when the standard deviation of the distribution is known. In practice this is infrequent, the standard deviation has to be estimated usually from the sample by using the statistic

$$S_n^* = +\sqrt{\frac{\Sigma(X_i - X)^2}{n-1}}.$$

In such cases the so-called t test is applied.

Let the distribution of X be $N(m; \sigma)$ $(-\infty < m < +\infty, \sigma > 0)$ and let $X_1, X_2, \ldots, X_n$ be a sample of size n for X.

To check hypothesis $H_0: E(X) = m_0$ construct the statistic

$$t = \frac{\bar{X} - m_0}{S_n^*} \sqrt{n}. \tag{6.11}$$

The cumulative distribution function of this statistic is

$$(6.12)\quad P(t<x)=\frac{1}{\sqrt{\pi(n-1)}}\cdot\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\int_{-\infty}^{x}\frac{du}{\left(1+\frac{u^2}{n-1}\right)^{n/2}}\qquad(-\infty<x<\infty)$$

the so-called Student distribution with $(n-1)$ degrees of freedom. By using a table such a $t_{\alpha/2}$ value can be picked out for any arbitrary ε level with which it holds that

$$(6.13)\qquad P(|t|>t_{\alpha/2})=\alpha.$$

This means that for the alternative hypothesis $H_1: E(X)=m\neq m_0$ the α-size symmetrical critical region ($t<t_{\alpha/2}$ or $t>t_{\alpha/2}$) is applied.

As it can be seen in formula (6.12) the distribution of t doesn't depend on the unknown σ, so that the probability of an error of the first kind is independent of σ as well; with given $t_{\alpha/2}$ the size of test is α for any σ. (A test of this type is called similar test.) The region of acceptance for a t test is, therefore, the interval

$$-t_{\alpha/2}<\sqrt{n}\,\frac{\overline{X}-m_0}{S^*}<t_{\alpha/2}$$

which has probability $1-\alpha$ under H_0.

Relationship (6.13) is equivalent to

$$P\left(m_0-t_{\alpha/2}\frac{S^*}{\sqrt{n}}<\overline{X}<m_0+t_{\alpha/2}\frac{S^*}{\sqrt{n}}\right)=1-\alpha,$$

if the value of $\overline{X}$ fails to fall into the above interval.

In case of a one-sided alternative H_1 against H_0 that is with $H_1': E(X)=m>m_0$ the critical region is that part of the line which is to the right from t_α defined by $P(t>t_\alpha|H_0)=1-\alpha$.

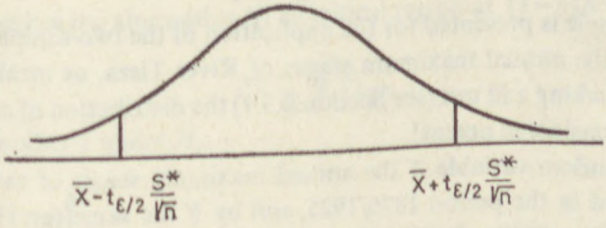


Figure 55

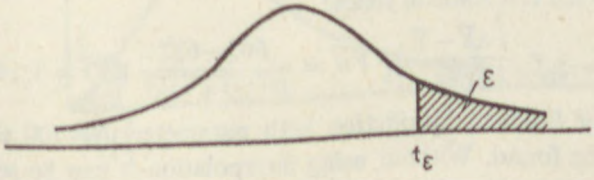


Figure 56

b) *The two-sample case*

When the task is to make comparison between the expected values of two normally distributed random variables and the variances are not known, the corresponding Student test, the so-called two-sample t test, may be constructed only when the unknown variances of both variables are identical.

The equality of variances may be verified by previous experience or theoretical considerations. When these are lacking the equality of variances require justification by applying an F test (Section 6.2.2).

Let X and Y be independent random variables from normal distribution with equal standard deviations $D(X)=D(Y)$ and let the null hypothesis $H_0: E(X)=E(Y)$ be tested on the basis of independent samples: $X_1, \ldots, X_n$ for X and $Y_1, \ldots, Y_n$ for Y. For making decision in this problem the statistic

$$t_{n+m-2} = \frac{\overline{X}-\overline{Y}}{\sqrt{(n-1)S_n^{*2}+(m-1)S_n^{*2}}} \sqrt{\frac{nm(n+m-2)}{n+m}} \tag{6.14}$$

has to be applied.

If H_0 is true this statistic follows a Student distribution with parameter $(n+m-2)$.

If $H_1: E(X)\neq E(Y)$ then, similarly to the one-sample case, a symmetrical critical region is to be chosen i.e. from the table of Student distribution such a t_α value is taken out with which it holds that

$$P(-t_{\alpha/2} < t_{n+m-2} < t_{\alpha/2}) = 1-\alpha.$$

If the t_{n+m-2} statistic calculated for the test falls outside the interval defined in the parenthesis then the hypothesis H_0 will be rejected at level $(1-\alpha)$.

Note that if the number of elements, n, is the same in both samples statistic then (6.14) takes the following simpler form:

$$t_{2n-2} = \frac{\overline{X}-\overline{Y}}{\sqrt{S_X^{*2}+S_Y^{*2}}} \sqrt{n}. \tag{6.15}$$

Now an example is presented for the application of the two-sample t test.

In table T.2 the annual maximum stages of River Tisza, as measured at Szeged, are shown. By making a fit test (see Section 6.3.1) the distribution of annual maximum stages may be considered normal.

Denote by random variable X the annual maximum stages of the River Tisza at Szeged, observed in the period 1876/1925, and by Y the same for 1926/75. Consider the hypothesis $H_0: E(X)=E(Y)$. Alternative H_1 let be taken as $H_1: E(X)\neq E(Y)$. We may assume that $D(X)=D(Y)$ according to the result of the F-test (see 6.2.2).

Calculation of the test statistic yields:

$$t_{2n-2} = \frac{\overline{X}-\overline{Y}}{\sqrt{S_X^{*2}+S_Y^{*2}}} \sqrt{n} = \frac{663-632}{\sqrt{135^2+138^2}} \sqrt{50} = 1{,}14.$$

In the table of Student distribution with parameter $N=100$ the critical value $t_{0,95}=1.984$ can be found. Without using interpolation it can be seen that the test statistic falls into the region of acceptance and so there is no reason to reject H_0.

When the problem is to decide whether two normally distributed random variables — X and Y — haver or haven't equal variances the so-called F test is used. The true expected values of both variables can be neglected now. The formula

$$P(F_{f_1,f_2} < x) = \frac{f_1}{f_2} \frac{\Gamma\left(\frac{f_1+f_2}{2}\right)}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)} \cdot \tag{6.16}$$

$$\cdot \int_0^x \frac{\left(\frac{f_1}{f_2} t\right)^{\frac{f_1}{2}-1}}{\left(1+\frac{f_1}{f_2} t\right)^{\frac{f_1+f_2}{2}}} dt, \quad (x \geqq 0).$$

But in the practice for the two-sided alternative $H_1 : D(X) \neq D(Y)$ the test statistic

$$F^*_{f_1,f_2} = \max\left(\frac{S^{*2}_n}{S^{*2}_m}; \frac{S^{*2}_m}{S^{*2}_n}\right) > 1 \tag{6.17}$$

is applied.

Let the hypothesis H_0 be the equality of variances that is

$$H_0 = D(X) = D(Y).$$

Suppose that for X a sample of n elements: $X_1, X_2, \ldots, X_n$ and for Y a sample of m elements: $Y_1, Y_2, \ldots, Y_n$ is available. The F-statistic is $F_{f_1 f_2} = \frac{S^*_n}{S^*_m}$. This means that the larger estimate of variance is written in the numerator and in this way $f_1 = n-1$ and $f_2 = m-1$ if the first term is greater than the second one, see Figure 57.

Commonly against the alternative H_1 as critical region at $(1-\alpha)$ level the following region is chosen:

$$F^*_{f_1,f_2} \geqq F_{1-\alpha}$$

which has probability α under H_0.

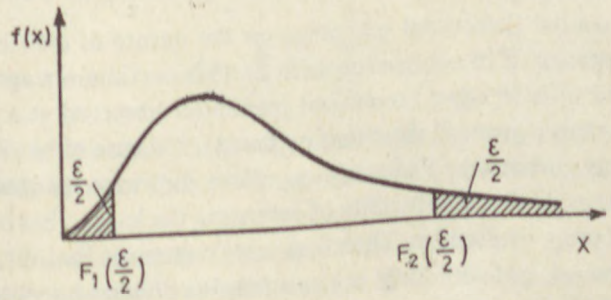


Figure 57

The way of applying the table used for F tests is, when the chosen level is $(1-\varepsilon)$, to compare F^* to $F_{1-\varepsilon}$ contained in the table. When using the table it should be kept in mind that f_1 is always the number of sample elements of the sample being in the numerator, whose estimated variance is larger while f_2 is the number of sample elements being in the denominator.

If $F^* < F_{1-\alpha}$ H_0 is accepted while if $F^* \geqq F_{1-\alpha}$ it is rejected, at $(1-\alpha)$ level.

As an example let be considered the problem whether the variances of annual maximum stages of River Tisza at Szeged, calculated for the intervals 1876/1925 and 1926/75, respectively, are or aren't equal.

Denote by X the maximum stages in the first fifty years and by Y the same for the second. Then

$$S_X^{*2} = 135^2 \quad \text{and} \quad S_Y^{*2} = 138^2,$$

$$F^* = \frac{138^2}{135^2} = 1.04, \quad f_1 = f_2 = 49.$$

If $\alpha = 0.1$ then $F_{1-\alpha} = 1.6$ so that there is no reason to reject H_0 (H_0 is accepted at 90 per cent level).

6.3. TEST OF GOODNESS OF FIT

6.3.1. ON TESTING THE GOODNESS OF FIT

In hydrological research it is a frequent situation that the distribution of a given random variable is not known. In the previous sections dealing with parametric tests the distribution of the random variables was supposed to be normal. This means that in such cases the normality is to be justified. On the other hand, in the practice of hydrology other kinds of distribution occur frequently as well. In flood hydrology, where, e.g., annual maximum stages or the durations of flood waves, etc., are analyzed, in the first place the distribution of the random variables in question are to be determined. Considering, e.g., the physical background of the phenomenon, sometimes an attempt can be made to derive a theoretical form of the distribution and then to create some hypothesis on the type thereof. The hypothesis obtained in this way is then checked by the tools of mathematical statistics, by means of the so-called testing of goodness of fit.

In certain cases the theoretical inherence on the nature of distribution is not too difficult. For instance, if in connection with annual maximum stages the only thing we are interested in is whether the annual maximum observed at a given gauge was higher or lower than a given x_0 the situation faced is the case of binomial distribution defined by some parameter $P(X > x_0) = p$. Here, however, another difficulty will arise, namely, whether the p probability of exceeding the level x_0 has or hasn't changed during the past years or decades. Therefore it is convenient that the record available (which is, in general, unfortunately not too long) is divided into parts and a test of homogeneity (see Section 6.4) is carried out thereupon.

When the physical background of a process is too sophisticated to build up some apparent hypothesis and the sample size is large enough then, using the sample, a density histogram is constructed whose shape may provide some instruction on the nature of distribution which should be checked then by a statistical test.

The most wide-spread two tests used to check fitting are the χ^2-test and the Kolmogorov test. The χ^2-test can be used equally for discrete or continuous distributions but a prerequisite to its application is the large number of sample elements. The Kolmogorov test can be applied for continuous distributions only. In the majority of cases such a hypothesis concerns not a single distribution but a (parametric) set of distributions. Hence in both tests in the form of estimators are used (the parameters of the distribution function has to be estimated also from the sample), reducing thereby the efficiency of these tests. In case of a smaller sample sometimes the sample elements will be transformed suitably and the Kolmogorov test will be applied on the new variables. For tests on normality a useful procedure of transformation is presented in Section 6.3.5.

6.3.2. THE χ^2-TEST

First the theoretical basis of the χ^2 test will be given, then its application to the goodness of fit tests and then, using hydrological examples, the so-called homogeneity and independence tests will be treated.

Let $A_1, A_2, \ldots, A_r$ a complete system of events, i.e., $\sum_1^r A_i = I$ and $A_i A_j = \emptyset$ if $i \neq j$. Consider the null hypothesis H_0: $P(A_i) = p_i$ $(i = 1, 2, \ldots, r; \sum_1^r = p = 1)$.

Suppose that out of n experiments the occurrences of events $A_1, A_2, \ldots, A_r$ were $\nu_1, \nu_2, \ldots, \nu_r$, respectively. The distribution of the random vector variable $(\nu_1, \nu_2, \ldots, \nu_r)$ is multinomial that is if H_0 is true then

$$P(\nu_1 = k_1, \nu_2 = k_2, \ldots, \nu_r = k_r | H_0) = \frac{n!}{k_1!\, k_2! \ldots k_r!} p_1^{k_1} p_2^{k_2} \ldots p_r^{k_r} \tag{6.18}$$

where $k_1 + k_2 + \ldots + k_r = n$.

Let now be formed the following statistic:

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}. \tag{6.19}$$

If hypothesis H_0 is true then $E(\nu_i) = Np_i$ that is in the nominator of the terms the square of deviations between each random variable ν_i and its own expected value can be found.

In Section 2.2.12 it was seen that the sum of the squares of r independent random variables having standard normal distribution is distributed according to the χ^2 distribution with parameter r. In formula (6.19) the sum of the squares of r non-inde-

pendent random variables is contained. It can be proved that if n tends to infinity the χ^2 statistic defined by formula (6.19) has χ^2 distribution with parameter $(r-1)$ that is

$$\lim_{n\to\infty} P(\chi^2 < x \mid H_0) = \frac{1}{2^{\frac{r-1}{2}} \cdot \Gamma\left(\frac{r-1}{2}\right)} \int_0^x t^{\frac{r-3}{2}} e^{-t}\, dt. \tag{6.20}$$

As a consequence, from the table of χ^2 distribution a critical value $\chi^2_{r-1}(\varepsilon)$ can be taken out for certain $\varepsilon>0$ with which

$$P\left(\chi^2 < \chi^2_{r-1}(\varepsilon)\right) = 1-\varepsilon. \tag{6.21}$$

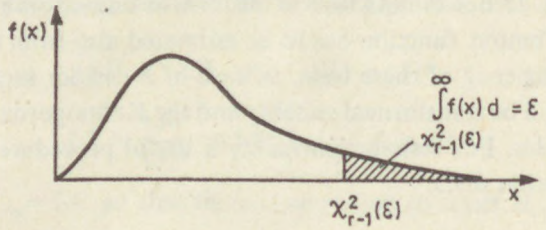


Figure 58

Note that the limiting distribution in formula (6.19) doesn't depend on p_i contained in H_0 so that the same table can be used to different hypotheses (considering, however, the value of r, the number of terms in the complete system of events). When selecting events A_i care should be taken for assuring that $np_i \geqq 10$ holds for the sake of reliability of the test; furthermore, this procedure can be used with large n values only. When the probabilities p_i depend on parameters of number s which were estimated from the sample then the parameter of the χ^2 statistic will be diminished by s, i.e., it will be $r-s-1$.

6.3.3. APPLICATION OF THE χ^2-TEST FOR FLOOD DATA

a) *Testing the distribution of flood wave occurrences*

According to records on the stages of River Tisza at Tokaj between 1903 and 1971 that is during $n=68$ years there were 30 years when water level $c=600$ cm was not exceeded in the first quarter. One exceedance in a quarter was observed in 25 years while two or more in 13 years.

Let now such a hypothesis H_0 be chosen that the number ν of exceedances in the first quarter follows Poisson distribution. For the parameter λ the value 0.8 was obtained from the data. Let be used the notation $A_1=\{\nu=0\}$, $A_2=\{\nu=1\}$, $A_3=\{\nu\geqq 2\}$. From the table of Poisson distribution when $\lambda=0.8$ then

$$P(A_1) = p_1 = 0.4493 \quad Np_1 = 30.55$$
$$P(A_2) = p_2 = 0.3595 \quad Np_2 = 24.45$$
$$P(A_3) = p_3 = 0.1912 \quad Np_3 = 13.00.$$

With these data

$$\chi^2 = \frac{(30-30.55)^2}{30.55} + \frac{(25-24.45)^2}{24.45} + \frac{(13-13)^2}{13} \approx 0.023.$$

From the table of χ^2 distribution with parameter $(r-2)=1$ and with the choice $\alpha=0.05$ the critical value $\chi^2_{0.05}=3.841$ is obtained. Consequently, there is no reason to reject the hypothesis. (Otherwise, as the coincidence between frequencies v_i and the expected values is surprisingly good, this result might be anticipated in advance.)

Note that since also the λ parameter of the hypothesized Poisson distribution has also been estimated from the sample, in the table of χ^2-distribution the critical value has been taken from the row corresponding to a parameter $(r-1-1)$. In practice mostly this kind of tests occur. If the parameters of the distribution are supposed to be known independently of the sample then the procedure is the same and the parameter of statistic χ^2 is $r-1$.

b) *Testing the distribution of exceedances*

Consider the exceedances of the Tisza river at Szeged, above the level $c=650$ cm in the second quarters. The numerical values are shown in Table T.1. Theoretical considerations and experience both suggest that the distribution is exponential. This gives the hypothesis

$$H_0\colon\ F(x) = 1-e^{-\lambda x},\quad x \geqq 0.$$

The value of parameter λ has to be estimated from the sample: as $x=1/\lambda \approx 100$ cm, $\lambda=0.01$. The exceedances are ranked into four categories by means of the quartiles of the distribution function $F(x)=1-e^{-0.01x}$: $\tilde{x}_{1/4}=30$ cm, $\tilde{x}_{1/2}=70$ cm and $\tilde{x}_{3/4}=140$ cm. Now

$$P(X < 30 \text{ cm}) = P(30 \text{ cm} \leqq X < 70 \text{ cm}) =$$

$$= P(70 \text{ cm} \leqq X < 140 \text{ cm}) = P(X \geqq 140 \text{ cm}) = 0.25.$$

The frequencies of the events above are:

$$v_1 = 6,\quad v_2 = 9,\quad v_3 = 9,\quad v_4 = 7.$$

With the given partitioning $np_i=31:4=7.75\approx 8$ as the number of exceedances observed in the second quarters of the period 1900/1970 was $n=31$.

Apply now the χ^2 test! The test statistic (with $r=4$) is:

$$\chi^2 = \sum_{i=1}^{4} \frac{(v_i-np_i)^2}{np_i} = \frac{(6-8)^2}{8} + \frac{(9-8)^2}{8} + \frac{(9-8)^2}{8} + \frac{(7-8)^2}{8} = \frac{7}{8} = 0.875.$$

As what is performed is a goodness of fit test using the estimation of one parameter the result obtained should be compared with the χ^2 value of $r-1-1=2$ degrees of freedom. The critical value of 0.05 per cent is χ^2_2 (crit.) $=599$. As it can be seen the fit is strikingly good. Inspecting the table of χ^2-distribution it is seen that, if H_0 is true, the probability of obtaining by chance a larger deviation than 0.875 is greater than 80 per cent. So there is no reason to reject hypothesis H_0.

Although our result seems to be rather convincing, having a relatively small sample the performance of a Störmer test is still advisable. (This is expedient all the more because in a Störmer test no part is played by the value of parameters.) For the Störmer test see Section 6.3.6.

c) *Testing the distribution of the largest exceedances*

We establish our null hypothesis according to considerations in Section 6.1. This means that the conditional distribution

$$H_0\colon F_t(x) = \frac{e^{-\lambda t e^{-\beta x}} - e^{-\lambda t}}{1 - e^{-\lambda t}}$$

will be the basis of our procedure; accordingly, only those years will be taken in which exceedance happened at all.

Consider again the exceedances in the Tisza river at Szeged (in the second quarters) but now in all quarters only the maximum one will be considered. Estimating both λt and β from the sample the null hypothesis has the form

$$H_0\colon F(x) = \frac{e^{-0.44e^{-0.01x}} - e^{-0.44}}{1 - e^{-0.44}}.$$

The number of sample elements (from Table T.1) is now $n=25$ which is, unfortunately, rather low. The maximum exceedances Z are ranked into four categories by using again the quartiles. The quartiles of the cumulative distribution function $F_t(x)$ are

$$\tilde{x}_{1/4} = 35 \text{ cm}; \quad \tilde{x}_{1/2} = 81 \text{ cm}; \quad \tilde{x}_{3/4} = 155 \text{ cm}.$$

With the given partitioning:

$$P(Z < 35 \text{ cm}) = P(35 \text{ cm} \leqq Z < 81 \text{ cm}) =$$

$$= P(81 \text{ cm} \leqq Z < 155 \text{ cm}) = P(Z \geqq 155 \text{ cm}) = 0.25.$$

From Table T.1 the frequencies of the above events are:

$$\nu_1 = 6, \quad \nu_2 = 5, \quad \nu_3 = 7, \quad \nu_4 = 7.$$

As $np_i = 25/4 \approx 6$ the actual value of the test statistic (with $r=4$) is:

$$\chi^2 = \sum_{i=1}^{4} \frac{(\nu_i - np_i)^2}{np_i} = \frac{(6-6)^2}{6} + \frac{(5-6)^2}{5} + \frac{(7-6)^2}{6} + \frac{(7-6)^2}{6} = \frac{3}{6} = 0.5.$$

Since the number of parameters estimated is two the critical value of the χ^2-distribution belonging to $r-3$ degrees of freedom at 0.05 per cent level is

$$\chi^2_1(\text{crit.}) = 3.84.$$

From the table of χ^2-distribution it is seen that, if H_0 is true, the probability of obtaining by chance a larger deviation in χ^2 than 0.5 is greater than 95 per cent. So there is no reason to reject H_0.

d) *Application of χ^2-test for testing homogeneity*

When large samples are handled the χ^2-test is used customarily to perform a so-called homogeneity test. The purpose of testing homogeneity is to decide whether two random variables, X and Y, are or aren't of the same distribution that is, in other words, whether the sample $X_1, X_2, \ldots, X_n$ and the other, $Y_1, Y_2, \ldots, Y_m$ related to Y, are or aren't drawn from populations with the same distribution. So the null hypothesis is:

$$H_0: P(X < x) = P(Y < x), \quad (-\infty < x < +\infty).$$

Testing homogeneity is of very great importance in the practice of hydrology. If, e.g., the question to be answered is whether in the course of long decades there were or weren't significant changes in the flow regime of a given river the data series is divided into two (or more) sub-records, in accordance with the date(s) of change(s) (such as land use, structures, etc.) and a test is performed to check homogeneity. The test of homogeneity is also of great importance when the integration of data series is the problem: e.g., in order to have a larger sample for improving the reliability of a next test of fit. Due to the importance of this subject other methods of testing homogeneity are also presented in Sections 6.4.2 and 6.4.4.

By using the χ^2-test the procedure of testing homogeneity is as follows. Divide the line into r parts by applying the division points $-\infty = z_0 < z_1 < \ldots < z_r = \infty$. Out of the sample representing a random variable X denote by ν_i the number of observations found in interval (z_{i-1}, z_i) while let μ_i denote the same for a random variable $Y (i = 1, 2, \ldots, r)$. Obviously, $\sum_1^r \nu_i = n$ and $\sum_1^r \mu_i = m$. The following statistic will be applied:

$$\chi^2 = nm \sum_1^r \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\nu_i + \mu_i}. \tag{6.22}$$

It can be proved that if $n \to \infty$ and $m \to \infty$ then the distribution of the statistic 6.22 tends to a χ^2 distribution with parameter $(r-1)$.

The test of homogeneity will be demonstrated by the following example. The question to be answered is whether the annual maximum stages of River Tisza, as observed at Szeged, did or didn't follow the same distribution in the periods 1876/1925 and 1926/75. (See Table T.2.)

Denote by the random variable X the maximum stages in the first fifty years and by Y those in the second. Using the notation $A_1: \{X < 5\ \text{m}\}$, $A_2: \{5\ \text{m} \leqq X < 6\ \text{m}\}$, $A_3: \{6\ \text{m} \leqq X < 7\ \text{m}\}$, $A_4: \{7\ \text{m} \leqq X < 8\ \text{m}\}$, $A_5: \{X \geqq 8\ \text{m}\}$ and, in addition denoting by ν_i the number of sample elements X_k contained in set A_i and by μ_i the number of sample elements Y_j in the same set $(i = 1, 2, \ldots, 5;\ k = 1, 2, \ldots, 50;\ j = 1, 2, \ldots, 50)$ the χ^2 statistic $(n = m = 50, r = 5)$ takes a simple form:

$$\chi^2 = n^2 \sum_1^r \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\nu_i + \mu_i} = \sum_1^r \frac{(\nu_i - \mu_i)^2}{\nu_i + \mu_i}.$$

The data obtained are as follows.

A_i	ν_i	μ_i
A_1	5	10
A_2	11	11
A_3	13	13
A_4	13	10
A_5	8	6

$$\chi^2 = \frac{(5-10)^2}{15} + \frac{(11-11)^2}{22} + \frac{(13-13)^2}{26} + \frac{(13-10)^2}{23} + \frac{(8-6)^2}{14} \approx 2.35.$$

Since in our example $r=5$, for testing the null hypothesis the critical value can be taken from the χ^2 distribution whose parameter is $(r-1)=4$. As for the level chosen $(\alpha=0.05)$ $1-\alpha=0.95$ this value is $\chi^2_4(0.05)=9.48$.

As it can be seen there is no reason to reject hypothesis H_0.

e) *Test of independence by using χ^2-test*

The purpose of testing independence is to check if two random variables — let they be denoted by X and Y — can or cannot be considered independent that is it does or doesn't hold that

$$H_0: P(X < x, Y < y) = P(X < x)P(Y < y), \quad (-\infty < x, y < +\infty).$$

The test of independence is a tool of analysing the stochastic relation between two random variables, an important problem encountered frequently in the technical practice. This subject is discussed in Chapter 7 in more detail. The only intention here is to show the procedure of applying χ^2-test for checking independence and to illustrate this procedure through a concrete example.

Consider a sample of N elements, each consisting of a couple (X, Y): (X_1, Y_1); (X_2, Y_2); ...; (X_N, Y_N). (Here X_i and Y_i are related values, coming from the same i-th observation.) Random variables X and Y may be of quite different nature, e.g., X can be a discrete while Y a continuous variable.

Now the χ^2-test is applied in the following manner. Divide the x axis into r sub-intervals by division points $-\infty = x_0 < x_1 < x_2 < \ldots < x_r = +\infty$, taking into account the possible values of X, and the y axis into s sub-intervals by division points $-\infty = y_0 < y_1 < \ldots < y_s = +\infty$, in accordance with the possible values of Y. Denote by A_i the event $\{x_{i-1} \leqq X < x_i\}$ and by B_j the event $\{y_{j-1} \leqq Y < y_j\}$, $(i=1, 2, \ldots, r;$ $j=1, 2, \ldots, s)$. The paired values (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N) constitute a set of points on a plane.

Denoting by ν_{ij} the joint occurrence of events A_i and B_j, obviously ν_{ij} is equal to the number of points within an oblong with sides A_i and B_j, see Figures 59 and 60.

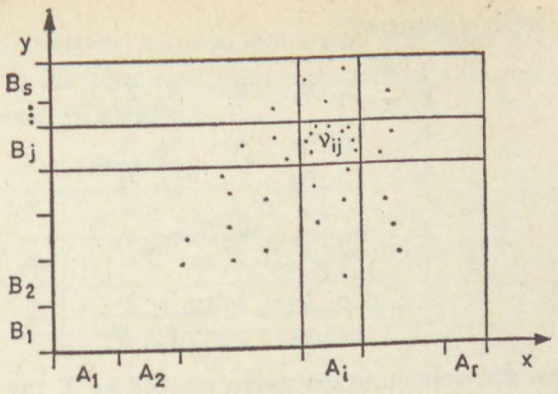


Figure 59

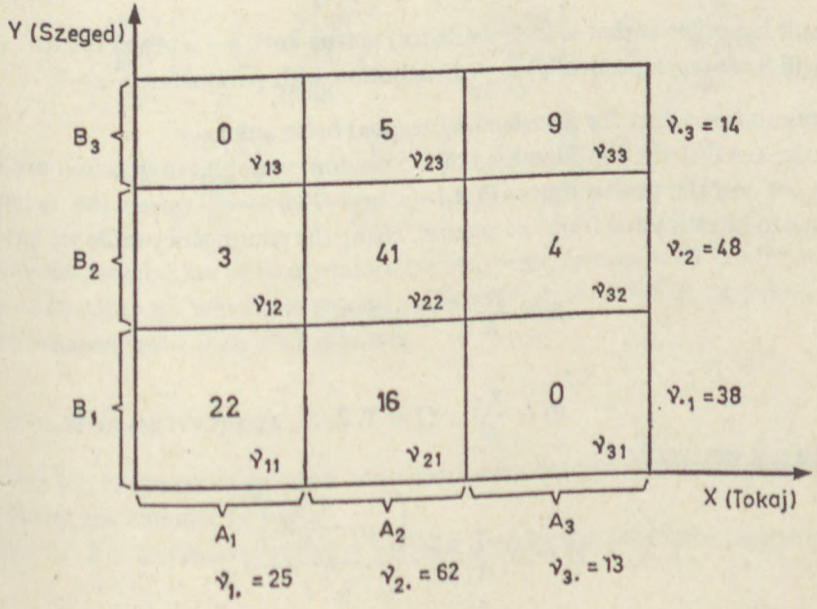


Figure 60

If X and Y are independent the events A_i and B_j are independent as well so that for hypothesis H_0 the hypothesis

$$H_0' = P(A_i, B_j) = P(A_i)P(B_j) = p_i q_j, \quad (i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, s)$$

will be replaced. To show the frequencies of the paired events $A_i B_j$ commonly a so-

called contingency table is prepared:

X \ Y	B_1	B_2	...	B_j	...	B_s	Σ
A_1	ν_{11}	ν_{12}	...	ν_{1j}	...	ν_{1s}	$\nu_{1.}$
A_2	ν_{21}	ν_{22}	...	ν_{2j}	...	ν_{2s}	$\nu_{2.}$
$\vdots$							
A_i	ν_{i1}	ν_{i2}	...	ν_{ij}	...	ν_{is}	$\nu_{i.}$
$\vdots$							
A_r	ν_{r1}	ν_{r2}	...	ν_{rj}	...	ν_{rs}	$\nu_{r.}$
Σ	$\nu_{1.}$	$\nu_{.2}$		$\nu_{.j}$		$\nu_{.s}$	N

In the marginal row and column, respectively, marked by Σ the frequencies corresponding to the so-called marginal distributions can be found. ($\nu_{i.}$ is the frequency of event A_i and $\nu_{.j}$ is that of $B_{i.}$). The test statistic for checking independence is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(\nu_{ij} - N p_i q_i)^2}{N p_i q_j}. \tag{6.23}$$

If the null hypothesis (that is independence) is true and $n \to \infty$ then the test statistic (6.23) will have (asymptotically) a χ^2 distribution with parameter $r \cdot s - 1$.

g) *Testing independence the distribution functions being unknown*

In practice the distribution functions of the random variables in question are usually not known and the probabilities $P(A_i) = p_i$ and $P(B_j) = q_j$ cannot be determined, they have to be estimated from the sample. Using the contingency table we have

$$p_i \approx \frac{\nu_{i.}}{N}, \quad (i = 1, 2, ..., r);$$

$$q_j \approx \frac{\nu_{.j}}{N}, \quad (j = 1, 2, ..., s).$$

Now the test statistic is

$$\chi^2 = N \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\nu_{ij} - \frac{\nu_{i.}\nu_{.j}}{N}\right)^2}{\nu_{i.}\nu_{.j}}. \tag{6.24}$$

This statistic follows (asymptotically) χ^2-distribution with parameter $(r-1)(s-1)$. Statistic (6.24) has a simple form when $r = s = 2$:

$$\chi^2 = N \frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1.}\nu_{2.}\nu_{.1}\nu_{.2}}. \tag{6.25}$$

In this case we have a statistic distributed according to χ^2-distribution with parameter $(r-1)(s-1) = 1$.

As an illustration consider the following example: Check whether the annual maximum stages of River Tisza at Tokaj can be considered independent of those

observed at Szeged. (Between these gauges several tributaries belonging to separate catchments flow into the Tisza river.) Based on Table T.1 the following contingency table and χ^2 values can be obtained:

Y	(Szeged)			
B_3	$\nu_{13}=0$	$\nu_{23}=5$	$\nu_{33}=9$	$\nu_{.3}=14$
B_2	$\nu_{12}=3$	$\nu_{22}=41$	$\nu_{32}=4$	$\nu_{.2}=48$
B_1	$\nu_{11}=22$	$\nu_{21}=16$	$\nu_{31}=0$	$\nu_{.1}=38$
	A_1	A_2	A_3	X (Tokaj)

$$\nu_{1.}=25 \quad \nu_{2.}=62 \quad \nu_{3.}=13$$

$$\chi^2=100\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{\left(\nu_{ij}-\frac{\nu_{i.}\nu_{.j}}{100}\right)^2}{\nu_{i.}\nu_{.j}}=100\left[\frac{\left(22-\frac{25.38}{100}\right)^2}{25.38}+\frac{\left(16-\frac{62.38}{100}\right)^2}{62.38}+\right.$$

$$+\frac{\left(0-\frac{13.38}{100}\right)^2}{13.38}+\frac{\left(3-\frac{25.48}{100}\right)^2}{25.48}+\frac{\left(41-\frac{62.48}{100}\right)^2}{62.48}+\frac{\left(4-\frac{13.48}{100}\right)^2}{13.48}+$$

$$\left.+\frac{\left(0-\frac{25.14}{100}\right)^2}{25.14}+\frac{\left(5-\frac{62.14}{100}\right)^2}{62.14}+\frac{\left(9-\frac{13.14}{100}\right)^2}{13.14}\right]\approx 64.$$

As now the distribution to be considered is the χ^2 distribution with $(r-1)(s-1)=4$ degrees of freedom for which the critical value at level $\varepsilon=0.001$ is 18.465 the hypothesis of independence should be rejected.

6.3.4. KOLMOGOROV-TEST

For testing the goodness of fit when continuous random variables are considered the Kolmogorov-test can also be used.

Denote by X a continuous random variable and let the corresponding sample be

(I) $$X_1, X_2, \ldots, X_n$$

consisting of n elements.

Hypothesis H_0 is now H_0: $P(X<x)=F(x)$ where $F(x)$ is supposed to be completely known. When $F(x)$ has unknown parameters they should be estimated from the sample; in such cases goodness of fit tests with estimation are performed. In this case it is more convenient to use the method which is based on the transformation of sample elements (see 6.3.5).

The testing procedure is as follows. First the elements of sample (I) are ranked in increasing order of magnitude:

(II) $$X_1^* < X_2^* < \ldots X_m^*,$$

then the empirical distribution function $F_n(x)$ is constructed and, finally, the test statistic

$$D_n = \sup_{-\infty<x<\infty} |F_n(x)-F(x)|$$

is calculated.

Since the empirical distribution function $F_n(x)$ depends on the sample elements $X_1, X_2, \ldots, X_n$ (i.e., $F_n(x)$ is a random function), the D_n value of the maximum deviation between the theoretical and empirical distribution function depends on chance as well that is D_n is a random variable.

If H_0 is true then, in accordance with Glivenko's theorem (see 4.1.3), with increasing n the D_n value tends to zero. When n is fixed the order of magnitude of D_n is $1/\sqrt{n}$ and therefore the $\sqrt{n}D_n$ value fails to tend to zero even if n is extremely large. The limiting distribution of random variable $\sqrt{n}D_n$ was determined by Kolmogorov who showed that if H_0 was true then

$$\lim_{n\to\infty} P(\sqrt{n}\,D_n<z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2z^2} = K(z), \quad z>0. \tag{6.26}$$

Values of function $K(z)$ have been tabulated (see Table T.6).

By the aid of a table containing function $K(z)$ such a z_0 can be chosen for which the relation

$$1-K(z_0) = P(\sqrt{n}\,D_n > z_0) = \alpha$$

holds where α is a given value (e.g., $\alpha=0.05$), the size of the given test. If the actual value of $\sqrt{n}D_n$ appears to be greater than z_0, hypothesis H_0 is rejected at $(1-\alpha)$ level.

Note that statistic

$$D_n^+ = \sup_x [F_n(x)-F(x)]$$

is also commonly used for fitting tests; this is an examination of the one-sided maximum deviation when the hypotheses are

$$H_0: P(X<x) = F(x)$$

and

$$H_1: P(X<x) > F(x),$$

i.e., the alternative is one-sided.

It has been shown that

$$\lim_{n\to\infty} P(\sqrt{n}\,D_n^+ < z) = 1-e^{-2z^2} = S(z), \quad (z \geqq 0). \tag{6.27}$$

To calculate the critical values no table is needed.

If a level of $(1-\alpha)$ decision is required for the decision then it is sufficient to choose such a critical value z_0 for which $e^{-2z_0^2}=\alpha$. Hence $-2z_0^2=\ln\alpha$ and

$$z_0 = \sqrt{-\frac{1}{2}\ln\alpha}.$$

The critical region is that part of the line which lies to the right from z_0. If, e.g., $\alpha=0.05$ then $\ln \alpha=-2.99\cong -3$ and $z_0=\sqrt{1.5}=1.2247$.

The expected value of statistic $\sqrt{n}D_n^+$ can also be determined easily as its density function is $s(z)=4ze^{-2z^2}$.

$$E(\sqrt{n}\, D_n^+)=\int_0^\infty z(4ze^{-2z^2})\,dz \approx \frac{\sqrt{2\pi}}{4}=0.627.$$

Now a few examples are shown, using the Kolmogorov test.

a) *Testing of the goodness of fit for the distribution of annual maximum stages of River Danube at Budapest*

The data of annual maximum stages observed in the Danube at Budapest between 1901 and 1970 can be seen in Table T.3.

Let hypothesis H_0 be

$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2 dt}}$$

where the expected value m and variance σ^2 are estimated by the sample mean $\bar{X}$ and by the estimate of variance, respectively. The sample mean is

$$\bar{X}=\frac{\sum_1^{70} X_i}{70}=626.66 \text{ cm}$$

while the estimate of standard deviation comes to

$$S_n=\sqrt{\frac{\sum_1^{70}(X_i-\bar{X})^2}{70}}=87.21 \text{ cm}.$$

After standardizing the sample elements, preparing the empirical distribution function $F_n(x)$ and plotting the standardized normal distribution function the maximum difference between these two functions is to be located. The maximum difference can be found about at $x=0.45$. Here $\Phi(0.45)=0.674$ and $F_n(0.45)=42/70=0.6$ so that

$$\sqrt{70}\, D_{70}=\sqrt{70}\sup_x |F_n(x)-\Phi(x)|=8.4\cdot 0.074=0.6216.$$

According to the table of function $K(z)$ the critical value belonging to the level $\alpha=0.05$ is $z_0=1.35$. As the maximum difference found in the example is considerably smaller than this z_0 there is no reason to reject hypothesis H_0.

b) *Testing the goodness of fit of exceedances to the exponential distribution*

As it was demonstrated in Section 6.3.3 for flood waves (observed in the second quarters at Szeged in the Tisza river) the distribution of exceedances above the level $c=650$ cm fitted "well" the exponential distribution. This was obtained by using the

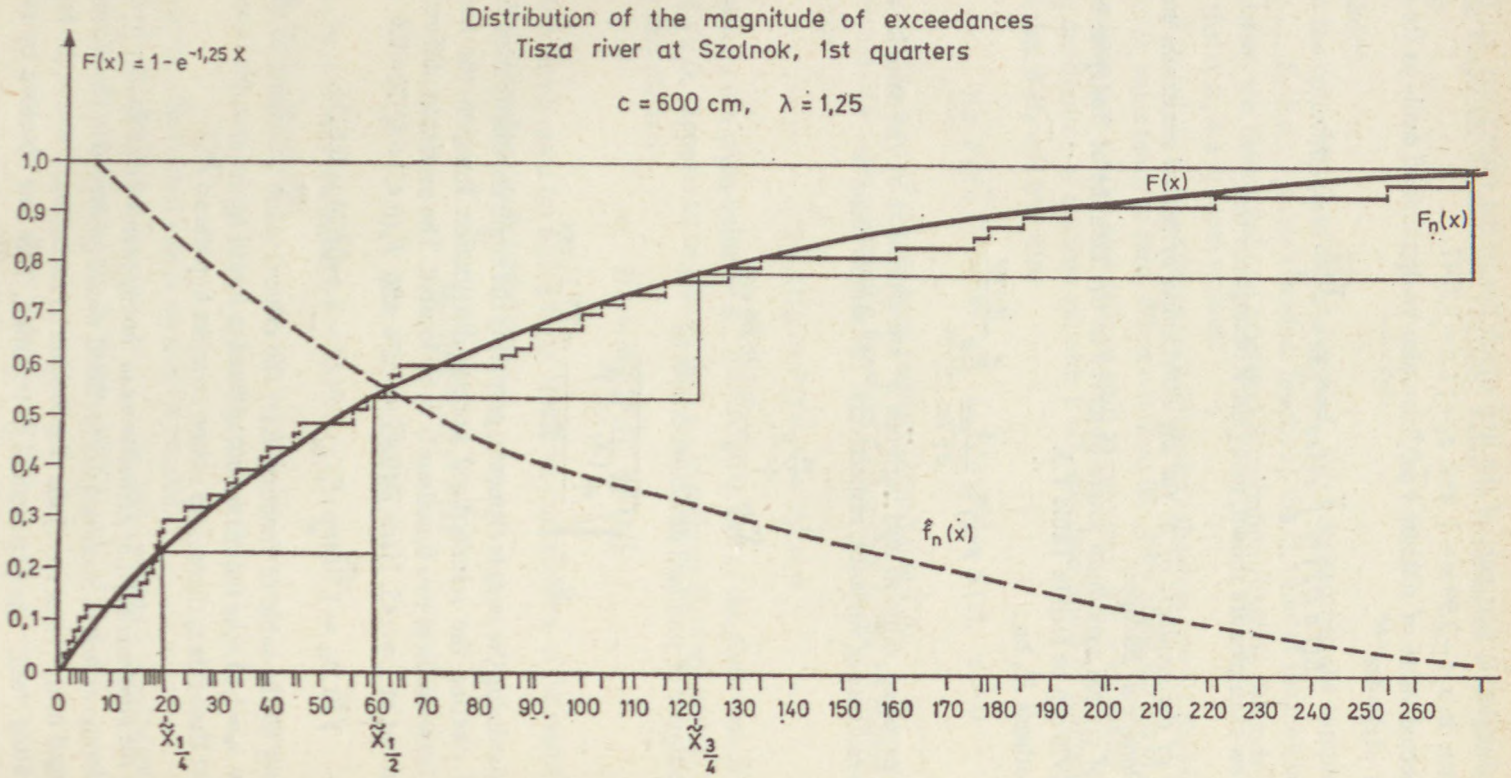

Distribution of the magnitude of exceedances
Tisza river at Szolnok, 1st quarters
c = 600 cm, λ = 1,25
F(x) = 1 - e^-1,25 X
F(x)
Fn(x)
f̂n(x)
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260
X̃1/4
X̃1/2
X̃3/4

Figure 61

χ^2-test. Let us turn now to the Kolmogorov test to check how the exceedances above $c=600$ cm observed in the second quarters at Szolnok fit the exponential distribution. The data of exceedances are shown in Table T.1. A maximum likelihood estimation on the λ parameter of exponential distribution results in $\lambda=1/\overline{X}=1/0.8=1.25$. Let be chosen the hypothesis H_0 so that the distribution function of X exceedances is $F(x)=1-e^{-1.25x}$.

Utilizing the ordered sample prepared from the exceedances contained in the table it is convenient to plot the empirical distribution function as it is shown in Fig. 61.

From the figure the maximum difference is $D_n=\max_x |F_n(x)-F(x)|: z_0=0.08$ and in this case $n=41$. According to Table 6.1 for $n=40$ the critical value for D_n is 0.25 which is about three times the value of the present z_0. So there is no reason to reject hypothesis H_0. Otherwise the fit is convincing merely by visual inspection, too. If the reliability of accepting H_0 is needed then the $\sqrt{n}D_n$ value should be calculated. This is about 0.6 and in the table of Kolmogorov's $K(z)$ function $K(0.6)\approx 0.1457\approx$ ≈ 0.15. So

$$\lim_{n\to\infty} P(\sqrt{n}\, D_n > 0.6) \approx 0.85$$

which means that if H_0 is true, merely by chance, in more than 85 per cent of the cases a higher maximum absolute difference would be observed between $F_n(x)$ and $F(x)$ than the present value.

The table below indicates that similar conclusions can be drawn also when the distributions of exceedances observed at Tokaj, Tiszafüred, Tiszaug and Szeged are concerned, either in the first quarter or in the second.

c) *Testing the goodness of fit to the distribution function*

$H_0: F_t(x)=\dfrac{e^{-\lambda te^{-\beta x}}-e^{-\lambda t}}{1-e^{-\lambda t}}$ in the case of maximum exceedances

Consider the fit of maximum exceedances for different gauges in the Tisza river in the first and second quarters to the distribution function $F_t(x)$ Fig. 62. Note that $F_t(x)$ is a conditional distribution function. If, in a given quarter, the magnitude of maximum exceedances is denoted by Z_t and the number of exceedances in the same quarter by ν then $F_t(x)=P(Z_t<x|\nu>0)$. As ν is distributed according to the Poisson law with parameter λt it holds that

$$P(\nu=0)=e^{-\lambda t}=P(Z_t\leqq 0).$$

The parameters λt and β are estimated from the corresponding samples. (The data of exceedances and the levels of c are given in the table.)

In Table 6.2 $\sqrt{n}\, D_n=\sqrt{n}\sup_x |F_n(x)-F_t(x)|$ where $F_n(x)$ is the empirical distribution function and z_0 is the actual value of the quantity $\sqrt{n}\, D_n$.

It is seen from the table that in all cases convincing fits can be found to $F_t(x)$.

Table 6.1

Fit of the X magnitude of exceedances to the exponential disbrution;

$$H_0\colon\ H(x) = 1-e^{-\lambda x}$$

$$\sqrt{n}\, D_n = \sqrt{n} \max_x |H_n(x)-H(x)|$$

where $H_n(x)$ is the empirical distribution function.

Tokaj: Quarter I:	Quarter II:
$\beta = 0.012$	$\beta = 0.01$
$\sqrt{n}\, D_0 = 0.9030$	$\sqrt{n}\, D_0 = 0.8784$
$P(\sqrt{n}\, D_n > \sqrt{n}\, D_0) = 0.39$	$P(\sqrt{n}\, D_n > \sqrt{n}\, D_0) = 0.42$
Tiszafüred: Quarter I:	Quarter II:
$\beta = 0.02$	$\beta = 0.02$
$\sqrt{n}\, D_0 = 0.7006$	$\sqrt{n}\, D_0 = 0.6890$
$P(\sqrt{n}\, D_n > \sqrt{n}\, D_0) = 0.62$	$P(\sqrt{n}\, D_n > \sqrt{n}\, D_0) = 0.73$
Tiszaug: Quarter I:	Quarter II:
$\beta = 0.01$	$\beta = 0.008$
$\sqrt{n}\, D_0 = 0.7860$	$\sqrt{n}\, D_0 = 0.7156$
$P(\sqrt{n}\, D_n > \sqrt{n}\, D_0) = 0.48$	$P(\sqrt{n}\, D_n > \sqrt{n}\, D_0) = 0.69$
Szeged: Quarter I:	Quarter II:
$\beta = 0.0096$	$\beta = 0.01$
$\sqrt{n}\, D_0 = 0.8438$	$\sqrt{n}\, D_0 = 0.5851$
$P(\sqrt{n}\, D_n > \sqrt{n}\, D_0) = 0.48$	$P(\sqrt{n}\, D_n > \sqrt{n}\, D_0) = 0.88$

6.3.5. TEST OF NORMALITY BASED ON THE TRANSFORMATION OF SAMPLE ELEMENTS (SARKADI TEST)

Given a sample for X, for which neither its variance nor its expected value is known decision should be made on whether it does or doesn't come from normal distribution. Let the sample elements be $X_1, X_2, \ldots, X_n$. Apply the following transformation:

$$Y_i = \frac{X_i - \bar{X}'}{S}\, \psi_{n-2}\left[\frac{|X_{n-1}-X_n|}{S}\right], \quad i = 1, 2, \ldots, n-2 \tag{6.28}$$

where $\bar{X}$ is the arithmetic mean of the sample and $\bar{X}'$ is

$$X' = \frac{\bar{X} + \sqrt{\frac{n}{2}}\,(X_{n-1}+X_n)}{n+\sqrt{2n}} \tag{6.29}$$

while S is calculated from

$$S^2 = \frac{2}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\,\bar{X}^2 - \frac{1}{2}(X_{n-1}-X_n)^2\right]. \tag{6.30}$$

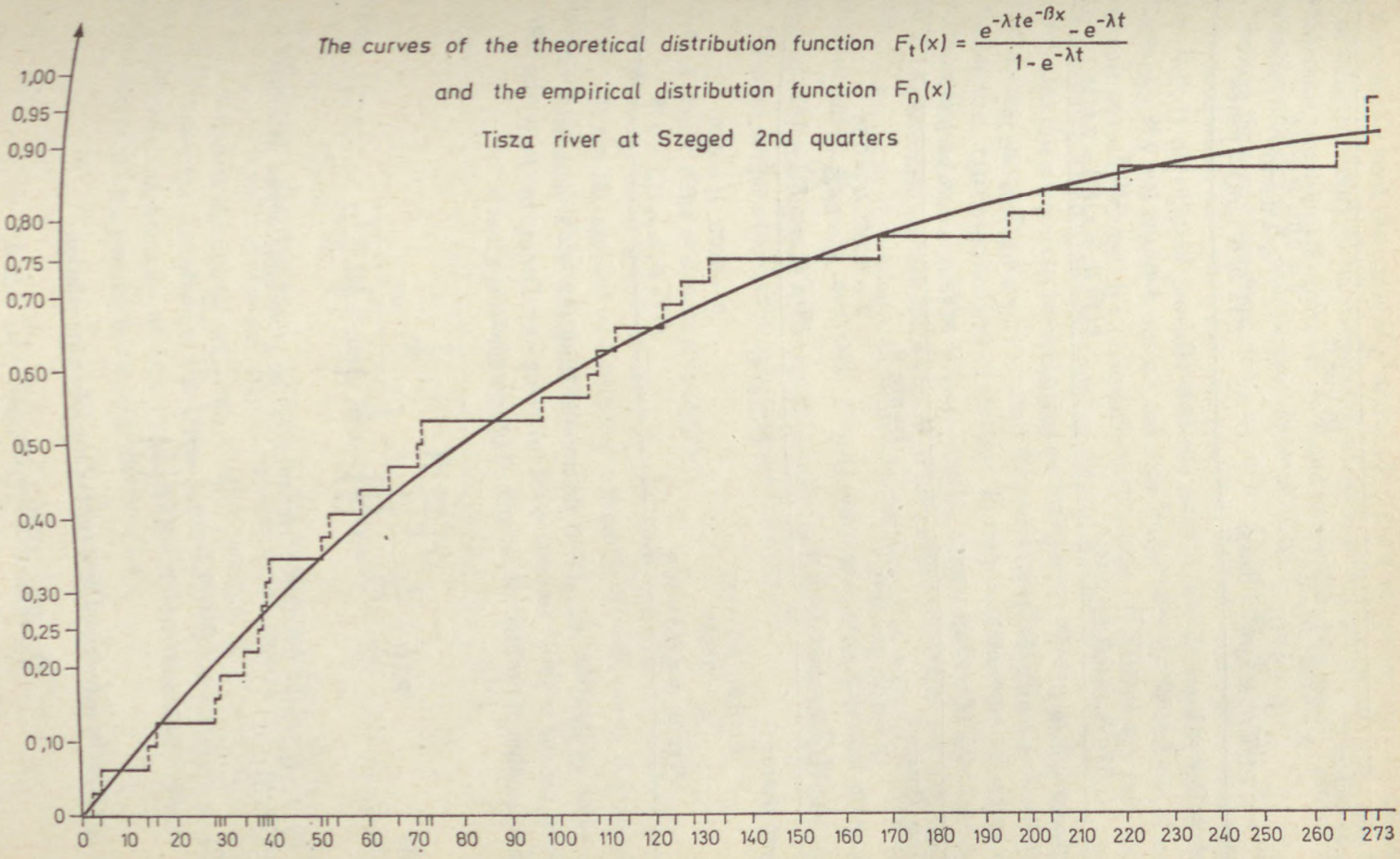

The curves of the theoretical distribution function $F_t(x) = \frac{e^{-\lambda t e^{-\beta x}} - e^{-\lambda t}}{1 - e^{-\lambda t}}$
and the empirical distribution function $F_n(x)$
Tisza river at Szeged 2nd quarters
1,00
0,95
0,90
0,80
0,75
0,70
0,60
0,50
0,40
0,30
0,25
0,20
0,10
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 273

Figure 62

Table 6.2

Tokaj: Quarter I	Quarter II
$\beta = 0.012;\ \lambda t = 0.8$	$\beta = 0.01;\ \lambda t = 0.6$
$z_0 = 0.7411$	$z_0 = 0.5610$
$P(\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.6430$	$P\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.9164$
Szolnok: Quarter I	Quarter II
$\beta = 0.01:\ \lambda t = 0.64$	$\beta = 0.01;\ \lambda t = 0.38$
$z_0 = 0.6833$	$z_0 = 0.6063$
$P(\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.7484$	$P(\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.8674$
Tiszafüred: Quarter I	Quarter II
$\beta = 0.02;\ \lambda t = 0.50$	$\beta = 0.008;\ \lambda t = 0.29$
$z_0 = 0.9617$	$z_0 = 0.7411$
$P(\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.31$	$P(\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.64$
Tiszaug: Quarter I	Quarter II
$\beta = 0.01:\ \lambda t = 0.35$	$\beta = 0.008;\ \lambda t = 0.26$
$z_0 = 0.9110$	$z_0 = 0.623$
$P(\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.38$	$P(\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.83$
Szeged: Quarter I	Quarter II
$\beta = 0.01;\ \lambda t = 0.89$	$\beta = 0.008;\ \lambda t = 0.26$
$z_0 = 0.7025$	$z_0 = 0.6231$
$P(\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.71$	$P(\sqrt{n}\, D_n \geqq z_0 \mid H_0) = 0.78$

The value of function $\psi_{n-2}(x)$ is determined from the $n-2$ parameter Student distribution and of the χ^2 variable in the following way. Denote by $P(t|\nu)$ the distribution function of the Student variable whose parameter is ν that is

$$\Phi_\nu(t) = \frac{1}{\nu\pi} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \int_{-\infty}^{t} \frac{du}{\left(1+\frac{u^2}{\nu}\right)^{\frac{\nu+1}{2}}} \tag{6.31}$$

and by $1-Q_\nu(t)$ the distribution function of the χ^2 variable whose parameter is ν which yields that

$$Q(t|\nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \int_{t}^{\infty} u^{\frac{\nu-1}{2}} e^{-\frac{u}{2}}\, du. \tag{6.32}$$

$\psi_{n-2}(t)$ can be determined by means of the following relation:

$$Q_{n-2}\{[\psi_{n-2}(t)]^2\} = 2\Phi_{n-2}(t) - 1, \quad t \geqq 0.$$

Since in this book the tables of χ^2 and Student distributions contain only a limited set of values necessary for the corresponding tests they are unsuitable to perform these calculations. Appropriate tables can be found in the following works: Biometrical Tables for Statistician [C.6] or Bolsev—Smirnov [C.1].

Through the above transformation a set of $n-2$ variables $Y_1, Y_2, \ldots, Y_{n-2}$ will be obtained which are — as it has been proved by Sarkadi — independent variables with $N(0; 1)$ distribution provided that the original sample comes from normal distribution. In this way the problem has been traced back to the analysis of such a sample where hypothesis is its normality with known parameters. For samples of this type purely a fitting test can be applied, e.g., the Kolmogorov test.

Note that this method can be used to test normality (even when — as in the analysis of variance — normality has to be tested) for several small samples simultaneously. This happens, e.g., in the analysis of variance. In such cases the expected values and variances may be different in the different samples. Consequently, the transformation has to be performed on each small sample. If the number of samples is r with sample sizes $n_1, n_2, \ldots, n_r$ then, after transformation, the number of independent variables obtained will be $\sum_{i=1}^{r} n_i - 2r$, each with $N(0; 1)$ distribution, provided that normality holds in each small sample. After transformation the test of hypothesis can take place for the great sample which is the union of the r small samples.

6.3.6. A TEST FOR EXPONENTIALITY

Given the independent sample elements $X_1, X_2, \ldots, X_n$ the question is whether they may or may not be regarded as distributed exponentially with a certain parameter. In this case the applicable transformation is

$$Y_\nu^* = \frac{\sum_{i=0}^{\nu-1} (n-1)(X_{i+1}^* - X_i^*)}{\sum_{j=1}^{n} X_j^*}, \quad (= 1, 2, \ldots, \nu-1) \tag{6.33}$$

where $X_1^*, X_2^*, \ldots, X_n^*$ is the ordered form of the given sample (that is $X_1^* \leqq X_2^* \leqq \ldots \leqq X_n^*$; $X_0^* = 0$).

If the original distribution is exponential the set of $n-1$ variables $Y_1^* \leqq Y_2^* \leqq \ldots \leqq Y_{n-1}^*$ obtained in this way may be regarded as the ordered form of a sample of $n-1$ elements distributed uniformly in the interval $[0, 1]$. This problem can be tested by using the Kolmogorov test (Section 6.3.4) and this combined method is asymptotically consistent for all continuous alternative hypotheses. Example on the application of this method can be found in Section 4.1.4.

6.4. METHODS FOR TESTING HOMOGENEITY

6.4.1 ON TESTING HOMOGENEITY, IN GENERAL

Let X and Y be continuous random variables with cumulative distribution functions $F(x)$ and $G(x)$, respectively. The question whether the distribution function of these random variables are the same (which means homogeneity) will be answered by nonparametric tests. Here the null hypothesis is

$$H_0: F(x) \equiv G(x)$$

and its alternative:

$$H_1: F(x) \not\equiv G(x).$$

So the meaning of this alternative is that these two cumulative distribution functions are not identical, there is at least one point where they differ (but in this case, due to continuity, they differ along a certain section). If H_0 is not valid that is H_1 is true then $G(x)$ includes a very wide variety of distributions differing from $F(x)$. A question arising now is whether it is possible at all to produce such a statistical test which will provide an efficient method to reject H_0 in all cases where a certain distribution function $G(x)$ differs from $F(x)$. Such nonparametric tests being *consistent* against all alternatives do exist; this means that with sufficiently large number of sample elements (denoted by $N \to \infty$) the probability of rejecting H_0 tends to unity provided that $G(x) \equiv F(x)$. This means that, dealing with sufficiently large samples, the test will reveal even the relatively small differences between the distribution functions. A test having this property is, e.g., the Kolmogorov—Smirnov two-sample test. With fixed number of sample elements the power of such a universal test is, of course, different against different alternative hypotheses. If the difference between $G(x)$ and $F(x)$ is small the disclosure thereof may be expected only when the number of sample elements is very large. Therefore, it is a common task to consider what are the alternatives against which a given test is efficient and less efficient, respectively. Especially the following two types of alternatives are investigated customarily:

$$H_1^{(*)}: G(x) = F(x-\delta)$$

which represents a shift in the expected value while the nature of distribution is unchanged and

$$H_1^{(**)}: G(x) = F\left(\frac{x}{h}\right)$$

which assumes a change in the scale of the variable that is in the variance. Concerning the alternative hypothesis $H_1^{(**)}$, if $h > 1$ then the variance of Y is greater than that of Y while with $h < 1$ the reverse is true.

Certain nonparametric tests are more sensitive to altenatives of type $H_1^{(*)}$ while others to those of type $H_1^{(**)}$. Sometimes the practical experience permits to draw conclusions on the type of such alternatives for which a test is more suitable.

The majority of nonparametric tests for homogeneity is based on ordered samples. To determine the test statistics based on ordered samples the elements of the two samples belonging to X and Y will be ranked in increasing order of magnitude:

(I) $$X_1^* < X_2^* < \ldots < X_n^*$$

and

(II) $$Y_1^* < Y_2^* < \ldots < Y_m^*.$$

(The decision as to which one of the random variables will be denoted by X and which one by Y has to be made before the test.)

A test covering more than one sample can be performed by utilizing the notion of rank attached to the elements of the samples. The definition of rank is the following. Let us have two samples and let us order them into a single sequence in increasing order of magnitude from 1 to $(n+m)$. Let this ordered sample be denoted by

(III) $$Z_1^* < Z_2^* < \ldots < Z_{n+m}^*.$$

Then we look for the place occupied by the smallest element of sample (I), X_1^*. Let this place be denoted by r_1 in the combined sample III; then the number r_1 is the rank of sample element X_1^*. Similarly, let the rank of X_2^* be denoted by r_2, etc. In this way the rank numbers of sample (I) will be represented by $r_1, r_2, \ldots, r_n$. In the combined sample let the ranks of the remained elements $Y_1^*, Y_2^*, \ldots, Y_m^*$ be denoted by $s_1, s_2, \ldots, s_m$. So it is obvious that the rank numbers of both series will give the integers from 1 to $(n+m)$. It is easy to realize that $r_1 < r_2 < \ldots < r_n$ and $r_i \geqq i$, furthermore, that $s_1 < s_2 < \ldots < s_m$ and $s_j \geqq j$. The numbers r_i and s_i themselves are random variables.

If hypothesis H_0 that both samples are of the same distribution is true then the probability of all arrangements of the two samples relative to one another is the same. As the number of possibilities to place X_i elements to $(n+m)$ places is $\binom{n+m}{n}$ the probability of a given sequence is $\dfrac{1}{\binom{n+m}{n}}$. Consequently,

$$P(r_1 = a_1, r_2 = a_2, \ldots, r_n = a_n | H_0) = \frac{1}{\binom{n+m}{n}},$$

for any sequence $1 \leqq a_1 < a_2 < \ldots < a_n \leqq (n+m)$.

If the null hypothesis is true, using combinatorial methods it is possible to determine the probability distribution of certain functions of random variables r_i and s_j. This needs sometimes rather tedious calculations even when a relatively simple function of rank numbers is defined. Therefore the tables of some tests cover small sample sizes only and when the number of sample elements is large limiting distributions are used.

Now the presentation of a few nonparametric tests follows.

This is a two-sample test which can be performed rather simply. It is asymptotically consistent for the alternative hypotheses. Its sensitivity is lower when the medians of distributions $F(x)$ and $G(x)$ are equal. But in turn it is consistent and efficient for hypotheses of the type $H_0: F(x) \equiv G(x)$ and $H_1^{(*)}: G(x) = F(x-\delta)$, $\delta \to 0$.

The test is used mostly to check the null hypothesis

$$H_0: P(X < Y) = P(X > Y) = 1/2.$$

Its one-sided form is applicable against the alternative hypothesis

$$H_1^*: P(X < Y) > 1/2.$$

In a case where $P(X<Y)<1/2$ variables X and Y can be inverted. The two-sided test is applicable to check the alternative hypothesis

$$H_1: P(X < Y) \neq P(X > Y).$$

The Wilcoxon test statistic has the form

$$W_{X,Y} = \sum_{i=1}^{n} (r_i - i) = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2} \tag{6.34}$$

which is, apart from an additive constant, the sum of the rank numbers of the X_i sample elements. It is easy to see that what is substracted from the sum of the ranks is the possible minimum of the sum of their ranks. Namely, if for the ranks it holds that $r_1=1$, $r_2=2, \ldots, r_n=n$ then the sum of the ranks is $\frac{n(n+1)}{n}$. The value of a $W_{X,Y}$ statistic may vary between 0 and nm.

It can be proved (see [A.17]) that if H_0 is true the expected value and the variance are

$$E(W_{X,Y}) = \frac{n \cdot m}{2} \tag{6.35}$$

and

$$D^2(W_{X,Y}) = \frac{nm(n+m+1)}{12}, \tag{6.36}$$

respectively. It has also been shown (see, e.g., [A.17] or [A.10]) that the limiting distribution of $W_{X,Y}$ is normal that is the following relation holds:

$$\lim_{n,m \to \infty} P\left(\frac{W_{X,Y} - E(W_{X,Y})}{D(W_{X,Y})} < x | H_0\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

An exact formula for the distribution of statistic $W_{X,Y}$ is not known. For small samples the tables have been calculated by using the following recursion which can be proved easily:

$$P(W_{n,m} = k) = \frac{n}{n+m} P(W_{n-1,m} = k) + \frac{m}{n+m} P(W_{n,m-1} = k). \tag{6.37}$$

The critical region at $(1-\alpha)$ level is the interval $(W_{X,Y} \leqq w_\alpha)$ where the value of w_α is determined by applying the relationship $P_{X,Y}(W > w_\alpha | H_0) = 1-\alpha$.

For showing the application of the Wilcoxon test an example is given for the one-sided case. Suppose that the problem is to check whether there has been an increase in flood levels in the Tisza river at Tokaj in the second quarters within the period 1945/70. In the period concerned the maximum exceedances ($c = 600$ cm) observed in the second quarters were (see Table T.1):

1945/58: $X_1 = 65$; $X_2 = 13$; $X_3 = 164$; $X_4 = 88$; $X_5 = 71$; $X_6 = 45$;

1962/70: $Y_1 = 194$; $Y_2 = 257$; $Y_3 = 123$; $Y_4 = 132$; $Y_5 = 55$; $Y_6 = 258$.

The ordered samples are:

$$X_1^* = 13, \quad X_2^* = 45, \quad X_3^* = 65, \quad X_4^* = 71, \quad X_5^* = 88, \quad X_6^* = 164$$

$$Y_1^* = 55, \quad Y_2^* = 123, \quad Y_3^* = 132, \quad Y_4^* = 194, \quad Y_5^* = 257, \quad Y_6^* = 258.$$

Now $H_0: P(X<Y) = \frac{1}{2}$, $H_1^*: P(X<Y) > 1/2$. The rank numbers are:

$$r_1 = 1; \quad r_2 = 2; \quad r_3 = 4; \quad r_4 = 5; \quad r_5 = 6; \quad r_6 = 9;$$

$$s_1 = 3; \quad s_2 = 7; \quad s_3 = 8; \quad s_4 = 10; \quad s_5 = 11; \quad s_6 = 12.$$

$$\sum_1^6 r_i = 27, \quad \sum_1^6 s_i = 51.$$

The test statistic is

$$W_{6,6} = \sum_1^6 r_i - \frac{6 \cdot 7}{2} = 6.$$

In Table T.9 for sample sizes $n=6$ and $m=6$ and for the level $1-\alpha=0.975$ the critical value of w_α is 5. As the actual value of the test statistic is greater than 5 there is no reason to reject H_0 so that the assumption that the flood levels didn't rise in the given cross section and in the period concerned is accepted.

As to the two-sided alternative hypothesis $H_1: P(X<Y) \neq P(Y<X)$ the critical region is the set $\{W_{X,Y} \leqq w'_{\alpha/2}$ or $W_{X,Y} \geqq w''_{\alpha/2}\}$.

For two-sided cases the table contains only one critical value: $w'_{\alpha/2}$, since the null distribution of the $W_{X,Y}$ statistic is symmetrical to the expected value, $\frac{nm}{2}$, so that $w'_{\alpha/2}$ and $w''_{\alpha/2}$ are in symmetrical position. Thus, when the value of $W_{X,Y}$ is greater than $\frac{nm}{2}$ the value to be compared to the critical value is $\frac{nm}{2} - W_{X,Y}$.

Now another example is shown on the application of the Wilcoxon test and this relates again to a flood problem. Let now be examined the problem whether the behaviour — the magnitude of exceedances — of floods in the Tisza river at Tokaj is the same in the first quarters (from 1st January to 31st March) as in the second ones (from 1st April to 30th June). Denote by the random variable X the maximum ex-

ceedances in the first quarters and by Y the same in the second. The samples related to X and Y, respectively, are shown in Table T.1.

With these two samples (which are relatively large) their ranking in increasing order of magnitude, integration into one single ordered sample and the establishment of ranks are cumbersome tasks. In fact these can easily be carried out if the sample elements are plotted on scale paper in the sequence of their numerical values, using two different colours or marks for the sample elements X_i and Y_j. In the course of such a representation the sample elements will get automatically into an ordered sequence and to determine the sum of rank numbers will also be easy since this is not influenced by the permutation of X_i elements among themselves.

In the example $m=36$ and $n=29$ so that what is applied is the normal limiting distribution of statistic $W_{X,Y}$:

$$W_{X,Y} = 510, \quad E(W_{X,Y}) = \frac{nm}{2} = \frac{36.29}{2} = 522;$$

$$D(W_{X,Y}) = 76.$$

$$W^* = \frac{W - E(W)}{D(W)} = \frac{-12}{76} = -0.16.$$

From the table of the standardized normal distribution

$$P(-2 \leqq W^* < 2) \approx 0.95$$

so that there is no reason to reject hypothesis H_0. (In the course of performing a test of this type such a problem may arise that certain sample elements have the same value that is there are equal X_i^* and Y_j^* elements and they cannot be ranked into a sequence. Suppose that the sample elements having the indices k, $k+1$, ..., $k+s$ are all equal in the combined sample, disregarding whether they are $X_i^* - s$ or $Y_j^* - s$. In this case each X_i^* from among the equal elements will get a rank number

$$r = \frac{k+(k+1)+\ldots+(k+s)}{s+1} = k+\frac{s}{2}.$$

Note that if the elements of both samples are plotted on the line (applying different marks for X_i^* an d Y_j^* observations) and for all X_i^* variables the Y_j^* values less than these X_i^*, that is the number of the pairs (X_i, Y_j) for which $Y_j < X_i$, are counted then, denoting this number by $W_{X,Y}$, the so-called Mann—Whitney statistic is calculated which is identical with the Wilcoxon statistic. To realize this assertion consider the ordered sample $X_1^* < X_2^* < \ldots < X_n^*$ and the corresponding ranks: $r_1 < r_2 < \ldots < r_n$. As the rank of X_1^* is r_1 the number of Y_j-s to the left therefrom is $r_1 - 1$; the rank of X_2^* is r_2, the number of observations, Y_j, less than X_2^* is $r_2 - 1$ but these include X_1^*, too, so that the number of Y_j-s less than X_2^* is $r_2 - 2$, etc. Obviously, the number

of Y_j-s less than X is $r_n - n$ and so

$$W_{Y,X} = r_1 - 1 + r_2 - 2 + \ldots + r_n - n =$$

$$= \sum_{i=1}^{n} r_i - (1+2+\ldots+n) = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2}.$$

Thus the Mann—Whitney test is essentially the same as the Wilcoxon test the only difference being that in the former a different method is included for the numerical quantification of the tests statistic.

As to the other variants of the Wilcoxon test and its power reference is made to Lehmann's work [A.17].

6.4.3. A COMBINATORIAL METHOD OF TESTING HOMOGENEITY

In the hydrological practice it is a frequent problem that decision should be made on the identity or discrepancy of the distributions of two random variables, X and Y, based on a sample where the number of elements is relatively small.

Suppose that for a random variable X the statistical sample

(I) $$X_1, X_2, \ldots, X_n$$

and for Y another sample

(II) $$Y_1, Y_2, \ldots, Y_n$$

are available. Suppose that the distribution of both random variables is continuous. Denote the probability function of random variable X by $F(x)$ and that of Y by $G(x)$.

The task is to test the hypothesis H_0: $F(x) \equiv G(x)$ by means of testing homogeneity. For cases where the sample sizes are equal that is when $n=m$ Gnedenko and Korolyuk have elaborated the exact distribution of the statistics:

$$B^{+}_{n,n} = \sup_x \{n[F_n(x) - G_n(x)]\}$$

and

$$B_{nn} = \sup_x \{n|F_n(x) - G_n(x)|\}.$$

In practice, however, the sample sizes are different and thus, for the application of the Gnedenko—Korolyuk test, some observations would have to be omitted which would result in the loss of valuable information. To avoid this, the Gnedenko—Korolyuk test will somewhat be generalized in the following way.

Let be $m > n$ but the difference $m-n$ should be $c\sqrt{n+m}$. Rank the samples (I) and (II) in increasing order of magnitude by which the ordered samples

(I*) $$X_1^* < X_2^* < \ldots < X_n^*$$

(II*) $$Y_1^* < Y_2^* < \ldots < Y_n^*$$

are obtained.

Combine now the ordered samples (I*) and (II*) into one single sample in increasing order of magnitude:

(III*) $$Z_1^* < Z_2^* < \ldots < Z_{m+n}^*.$$

Let

$$\vartheta_i = \begin{cases} +1 & \text{if } Z_i^* \in (\text{I}^*) \\ -1 & \text{if } Z_i^* \in (\text{II}^*) \end{cases}$$

and $S_j = \vartheta_1 + \vartheta_2 + \ldots + \vartheta_j$, $(j = 1, 2, \ldots, m+n)$. The partial sums S_j can be visualized by trajectories as shown in Fig. 63.

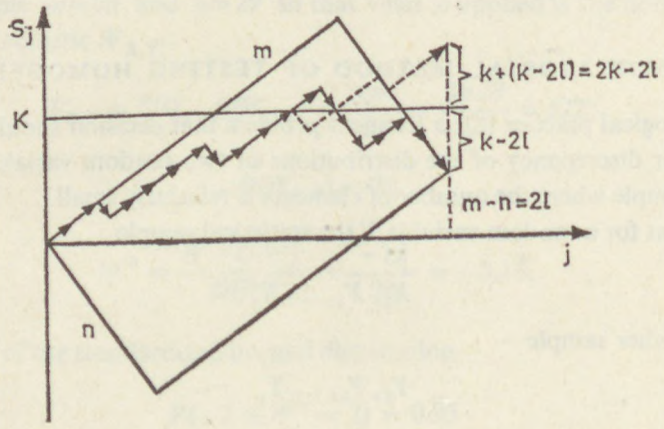


Figure 63

All possible trajectories run within the oblong shown in the figure and, obviously, the number of all trajectories is

$$\binom{m+n}{n} = \binom{m+n}{m}.$$

Choose now a straight line running parallelly to the j axis at a height k. Now such a trajectory which reaches this line, if reflected on the straight line $y=k$ from the first point of touch, will end after $(m+n)$ steps at the point $2k-(m-n)$.

Count the number of trajectories of this kind! For the sake of simplicity suppose that both m and n are even (this means the omission of at most one observation). So let $m+n=2N$, $m-n=2l$. Now the task is to find the number of such trajectories which start from the origin and terminate after $2N$ steps at a point whose elevation is $2k-2l$. If the number of steps upwards is α and the same downwards is β then

$$\begin{aligned} \alpha+\beta &= 2N \\ \alpha-\beta &= 2k-2l \\ \hline \alpha &= N+k-l \\ \beta &= N-k+l. \end{aligned}$$

Consequently, the number of trajectories reaching the line $y=k$ is now

$$\binom{2N}{N+k-l}.$$

Whence

$$P(\max_j S_j > k) = \frac{\binom{2N}{N+k-l}}{\binom{2N}{n}} = \frac{\binom{2N}{N+k-l}}{\binom{2N}{N-l}}. \tag{6.38}$$

To calculate formula (6.38) accurately the table of binomial distribution can be used (at least when $N \leqq 25$), in the following way: From the column under $p=0.5$ in the table of binomial distribution (Table T.10)

$$P_{N+k-l} = \binom{2N}{N+k-l}\left(\frac{1}{2}\right)^{N+k-l}\left(\frac{1}{2}\right)^{N-k+l} = \binom{2N}{N+k-l}\frac{1}{2^{2N}},$$

$$P_{N-l} = \binom{2N}{N-l}\left(\frac{1}{2}\right)^{N-l}\left(\frac{1}{2}\right)^{N+l} = \binom{2N}{N-l}\frac{1}{2^{2N}}.$$

Hence

$$\frac{P_{N+k-l}}{P_{N-l}} = \frac{\binom{2N}{N+k-l}}{\binom{2N}{N-l}}. \tag{6.39}$$

If the required level of decision on hypothesis H_0 is $1-\alpha$, where $\alpha>0$ is a number chosen appropriately, quotient (2.150) could be calculated for $k=1, 2, \ldots$ and such a k value could be selected with which the value of the above quotient will be less than α. However, on the basis of a theoretical consideration given below, the value of k can be anticipated by good approximation and in this way many unnecessary divisions can be avoided.

Now utilize the well-known relationship

$$\lim_{N\to\infty} \frac{\binom{2N}{N+c}}{\binom{2N}{N}} = e^{-\frac{c^2}{N}}. \tag{6.40}$$

Write Eq. (6.39) in the following form:

$$\frac{\binom{2N}{N+k-l}}{\binom{2N}{N-l}} = \frac{\binom{2N}{N+k-l}}{\binom{2N}{N}} \cdot \frac{\binom{2N}{N}}{\binom{2N}{N-l}}.$$

By virtue of the limit value (6.40):

$$\lim_{N\to\infty} \frac{\binom{2N}{N+k-l}}{\binom{2N}{N-l}} = \lim_{N\to\infty} \frac{\binom{2N}{N+k-l}}{\binom{2N}{N}} \cdot \lim_{N\to\infty} \frac{\binom{2N}{N}}{\binom{2N}{N-l}}.$$

Hence

$$P(\max_j S_j \geqq k) \approx e^{-\frac{k^2-2kl}{N}}. \tag{6.41}$$

From this relationship the value of k will be obtained by solving the equation $e^{-\frac{k^2-2kl}{N}}=\alpha$. After a simple calculation:

$$k^2-2kl+N\ln\alpha=0$$

$$k=\frac{2l\pm\sqrt{4l^2-4N\ln\alpha}}{2}=l+\sqrt{l^2-N\ln\alpha}. \tag{6.42}$$

In the practice of statistics $\alpha=0.05$ is chosen in most cases: thus the value of k will be given by the following relationship :

$$k=l+\sqrt{l^2+3N} \tag{6.43}$$

(because $\ln 0.05=-2.99\approx-3$).

The applicability of this formula is shown now through a numerical example.

Let $m=17$; $n=13$; with these $2N=30$, $l=2$. By virtue of formula (6.43)

$$k=2+\sqrt{49}=9.$$

Concludingly

$$\frac{\binom{2N}{N+k-l}}{\binom{2N}{N-l}}=\frac{\binom{30}{22}}{\binom{30}{13}}\leqq 0.05.$$

Indeed, in the table of binomial distribution

$$\frac{P_{22}}{P_{13}}=\frac{\binom{30}{22}}{\binom{30}{13}}=\frac{0.00545}{0.11152}=0.048\approx 0.05$$

can be found in the column of $p=0.5$. So, in our opinion, by using formula (6.4.3) the critical k value will be obtained by a sufficient accuracy, and thus the preparation of any table is unnecessary!

Note that for $m=n$ that is with $l=0$ formula (2.149) includes one of the results reached at by Gnedenko and Korolyuk namely that

$$P(\sup n[F_n(x)-G_n(x)]>k)=\frac{\binom{2N}{N+k}}{\binom{2N}{N}}\approx e^{-\frac{k^2}{N}}. \tag{6.44}$$

Nor is in this case a table needed; the critical k value will be given by formula

$$k=\sqrt{-N\ln\alpha} \tag{6.45}$$

if the level of decision required to hypothesis H_0 is $(1-\alpha)$. If $n=m$ and $\alpha=0.05$ the critical value is

$$k = \sqrt{3N}.$$

From Eq. (6.41) an interesting limiting distribution can be derived in the following way:

Introduce the notation $B^+_{m,n}=\max_x [mF_m(x)-G_n(x)]$ with which

$$P(\max_j S_j < k) = P(B^+_{m,n} < k) = 1-e^{-\frac{k^2-2kl}{N}}.$$

Let now

$$k = z\sqrt{2N}, \quad l = m-n = c\sqrt{2N}$$

which, by virtue of Eq. (2.156), yields a limiting distribution of the Kolmogorov—Smirnov type:

$$P\left(\frac{B^+_{m,n}}{\sqrt{m+n}} < z\right) = 1-e^{-2z^2-4cz}. \tag{6.46}$$

It can be proved that the above test is asymptotically consistent for the alternative hypothesis $H_1: F(x)>G(x)$. To statistic

$$B_{m,n} = \max_x \left| mF_m(x)-nG_n(x)+\frac{m-n}{2} \right| - \frac{m-n}{2}$$

the Gnedenko—Korolyuk test may be generalized as well. (See, e.g., [B. 28].) As to the efficiency of these tests the interested reader will find information in paper [B. 28].

As an illustration for the practical application of this test consider the exceedances of the Tisza river at Tokaj. Let the maximum exceedances observed in the first quarters (from 1st January to 31st March) be represented by the random variable X and those in the second (from 1st April to 30th June) by Y. Denote the distribution function of X by $F(x)$ and that of Y by $G(x)$. The statistical samples related to the random variables X and Y are given in Table 4. In this example $m=35$, $n=29$.

The sequence of the random variables $\vartheta_1, \vartheta_2, \ldots, \vartheta_n$ and the corresponding trajectories are shown in Fig. 63 from which $\max_j S_j=11$ can be read. In this case $m-n=2l=6$ with which, by virtue of formula (2.153), the k value corresponding to the critical level $=0.05$ is

$$k = l+\sqrt{l^2+3N} = 3+\sqrt{105} = 13.2.$$

So there is no reason to reject hypothesis $H_1: F(x)\equiv G(x)$. This means that the distribution of the maximum exceedances at Tokaj observed in the first and second quarters, respectively, are equal, allowing the integration of both samples through which a statistical sample with considerably larger (almost double) number of sample elements will be obtained for the maximum exceedances.

Note that to perform this test the representation of trajectories is unnecessary. By forming the series $\vartheta_i = \pm 1$ (where $i = 1, 2, \ldots, m+n$) a sequence $S_i (i = 1, 2, \ldots)$ can be formed and max S_i can be found very easily as this is shown by the table below.

ϑ_i	1	−1	1	−1	−1	1	−1	−1		1	−1		1	−1	1	1	−1	−1	−1			
S_1	1		0	1		0	−1	0	−1	−2	−1	−2	−1		0	1	0		1		0	−1

	1	1	−1		1	1	−1	1	−1	−1	1
	0	1		0	−1	0		1	0		1	0	−1

ϑ_i	1	1	1	−1	1	1	1	1	−1	−1	1	1	−1	1	1		
S_i	1	2	3		2	3	4	5	6		5		4	5	6		5	6	7

	1	1		1	−1		−1	−1	1	1		1		1	−1		−1	1		−1	−1	1
	8	9	10		9		8		7	8	9	10	11		10		9	10		9		8	9

	−1	−1	−1
	8	7	6

6.4.4. KOLMOGOROV—SMIRNOV TWO-SAMPLE TEST

This test is also used for testing homogeneity that is the problem to be solved through it is to test whether the distributions of random variables X and Y are identical (continuous).

Let the distribution functions be

$$P(X < x) = F(x) \quad \text{and} \quad P(Y < x) = G(x)$$

The null hypothesis is

$$H_0 : F(x) \equiv G(x)$$

Let the sample from X be

(I) $$X_1, X_2, \ldots, X_n$$

and from Y

(II) $$Y_1, Y_2, \ldots, Y_m.$$

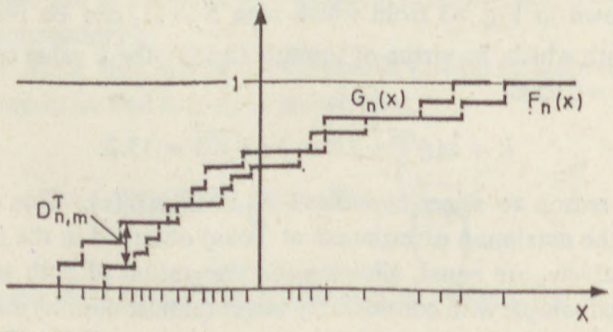


Figure 64

From the ordered form of these samples construct now the empirical distribution functions, $F_n(x)$ and $G_m(x)$, respectively:

Smirnov proved that

$$\lim_{n,m\to\infty} P\left(\sqrt{\frac{nm}{m+n}}\max_x [F_n(x)-G_m(x)] < z|H_0\right) = 1-e^{-2z^2-4cz} \tag{6.47}$$

and

$$\lim_{n,m\to\infty} P\left(\sqrt{\frac{nm}{m+n}}\max_x |F_n(x)-G_m(x)| < z|H_0\right) = \tag{6.48}$$

$$= K(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2z^2}.$$

With the one-sided alternative hypothesis $H_1^*: F(x)>G(x)$ the test statistic is

$$D^+_{n,m} = \max_x [F_n(x)-G_m(x)]$$

and the critical region at $(1-\alpha)$ level is $\{D^+_{n,m}\geqq D_\alpha\}$ where D_α can be calculated from relationship

$$P(D^+_{n,m} < D_\alpha|H_0) = 1-\alpha$$

as it is described in Section 2.5.5. As to the power of the test Vincze's paper [B. 45] can be referred to.

When the question to be answered is only whether the distribution of a random variable X is or is not the same as that of an Y that is when the two-sided alternative is $H_i: F(x)\neq G(x)$ then the $(1-\alpha)$ level critical region $\{D_{n,m}\geqq D'_\alpha\}$ is constructed by using statistic

$$D_{n,m} = \max_x |F_n(x)-G_m(x)|$$

where D'_α can be determined from relationship $P(D_{n,m}<D'_\alpha|H_0)=1-\alpha$. With large n and m the table of Kolmogorov's $K(z)$ function can be used to calculate D'_α. With $\alpha=0.05$, a common choice in the practice of statistics, and with $n, m\geqq 50$ hypothesis H_0 will be rejected at the 95 per cent level if

$$\sqrt{\frac{nm}{n+m}}\, D_{n,m} > 1.35.$$

Now an example related to floods is presented on the application of the two sample Kolmogorov—Smirnov test with two-sided alternative hypothesis.

Examine whether the maximum stages of the Tisza river observed at Szeged follow the same distribution in the periods 1876/1925 and 1926/75, respectively.

Denote by random variable X the annual maximum stages in the period 1876/1925 and by Y the same in the second fifty years. The samples related to X and Y are given in Table T.2. Let

$$H_0: F(x) \equiv G(x) \quad \text{and} \quad H_1: F(x) \not\equiv G(x).$$

The reader will prepare easily both the data ranked into a single sample in increasing order of magnitude and the corresponding empirical distribution functions.

Essentially, either now the representation of empirical distribution functions $F_n(x)$ and $G_m(x)$ in unnecessary, what is required is merely the position of sample elements X_i and Y_i relative to one another. This is because in this example $n=m=50$ so that

$$\sqrt{\frac{nm}{n+m}} D_{n,m} = \sqrt{\frac{n}{2}} D_{nn} = \frac{1}{\sqrt{2n}} \max_x |nF_n(x) - nG_n(x)| =$$

$$= \frac{1}{\sqrt{2n}} \max_i S_i = \frac{9}{10} = 0.9.$$

In the table of Kolmogorov's $K(z)$ function the critical value belonging to level $\alpha=0.05$ is $z_0=1.35$, no reason is, therefore, to reject hypothesis H_0. The application of this test on other gauges of River Tisza have led to similar results which fact may act as a confirmation of the decision made here.

6.5. METHODS FOR TESTING RANDOMNESS

6.5.1. THE WALD—WOLFOWITZ-TEST AND ITS APPLICATION FOR TESTING THE RANDOMNESS OF EXCEEDANCES

There is a strong seasonal variation in the hydrological cycle and, consequently, in the flow regime. From the viewpoint of flood control, due to the development of meteorological conditions, the behavior of most rivers is different in the different seasons. In accordance with this fact it is expedient that the behavior of flood waves in the different seasons are examined separately; in our case the examination of flow regime performed separately for the first, second, etc. quarters seems to be appropriate. Therefore, in the further discussions it is the quarters that will be chosen as interval $[0, T)$. By doing so the seasonal variations (by the very fact that separate seasons are considered) will in fact be eliminated since as far as the rivers examined here are concerned their flow regime may be considered homogeneous.

Let the chosen quarters of the given years be regarded as $[0, T)$ intervals whose number is k (and which include observations on exceedances at given gauges). In connection with the flood waves observed in temporal sequence the exceedances in the i-th period are denoted by the random variables $X_{i1}, X_{i2}, \ldots, X_{iv_i}$. If the values may be regarded as independent random variables from the same distribution the sequence

(I) $$X_{11}, X_{12}, \ldots, X_{1v_1}; X_{21}, X_{22}, \ldots, X_{2v_2};$$
$$X_{k1}, X_{k2}, \ldots, X_{kv_k}$$

may be considered a sample representing a random variable X which denotes the value of exceedances in the chosen period. For sake of simplicity let the series be

re-denoted in the form

(I′) $$X_1, X_2, \ldots, X_n$$

where, apparently, $n = \sum_{i=1}^{k} v_i$.

Thus the first problem to be analysed is whether the elements in series (I′) may or may not be considered independent random variables obtained from the same distribution. If so, the estimation of the distribution, expected value, etc. of X from this sample may be justified.

A sequence consisting of independent random variables from the same distribution is called commonly "random sequence" and this has the important property that although the X_i values are given in temporal sequence still they may be regarded as those forming a random sequence.

Different statistical methods called "testing randomness" — and, therefore, belonging to the sphere of hypothesis testing — can be used to check whether the elements of series (I′) form a random series.

First a H_0 hypothesis assuming that the elements of sequence (I′) constitute a random sequence that is a statistical sample, is set up. Thus this means that random variables $X_1, X_2, \ldots, X_n$ are independent and that for the cumulative distribution functions $P(X_i < x) = F_i(x)$, $(i = 1, 2, \ldots, n)$, a relation stating that $F_1(x) = F_2(x) = \ldots = F_n(x)$ will be valid.

The procedure then is that described in Section 2.4.1 : a statistic providing the basis of test is chosen and — supposing that H_0 is true — its distribution is determined. Then the critical region is constructed for the given level; if the actual value of the statistic falls in this region hypothesis H_0 will be rejected otherwise it is accepted.

The problem is the type of statistic to be constructed from the elements of sample (I′) which provides a basis to make decision on H_0. To find a suitable statistic is made difficult by the fact that there may be several reasons resulting that H_0 is false. When selecting a statistic all the realistic alternative hypotheses should be considered. Obviously, H_0 will be false if some trend is present in sequence (I′) that is when the elements of the sequence show an increasing or decreasing (or possibly another) tendency. H_0 will be false also when some dependence, stochastic interrelation prevails among the elements of the sequence. (It may, e.g., occur that there is a dependence between the magnitude of an exceedance and the preceding one.)

Wald and Wolfowitz [B. 42] have elaborated a so-called exact proof to test randomness, which is based on the serial correlation. Below a brief description of their method is given and then the application thereof is shown by testing the data series of exceedances observed at Szeged in the Tisza river.

Consider again the sequence

(I′) $$X_1, X_2, \ldots, X_n$$

and what is to be tested is the hypothesis H_0 related to this series. Calculate the arithmethic mean $\bar{X} = \frac{1}{n} \sum_{1}^{n} X_i$ and subtract this $\bar{X}$ from each element of the series that

is let be formed

$$X_i' = X_i - \bar{X}, \quad (i = 1, 2, \ldots, n). \tag{6.49}$$

Now calculate statistic

$$R = \sum_{i=1}^{n-1} X_i' X_{i+1}' + X_n' X_1'. \tag{6.50}$$

As it was shown by Wald and Wolfowitz the standardized random variable

$$R^* = \frac{R - E(R)}{D(R)} \tag{6.51}$$

is asymptotically $N(0; 1)$ distributed which means that if $|R^*| > 1.96$ hypothesis H_0 should be rejected at the 95 per cent level.

Statistics $E(R)$ and $D(R)$ are estimated from sample (I') in the following way:

For the sums of powers introduce the notation

$$S_r = \sum_{i=1}^{n} X_i'^r \tag{6.52}$$

while for the expected value and variance of R apply the following estimators:

$$E(R) \sim -\frac{S_2}{n-1} \tag{6.53}$$

$$D^2(R) \sim \frac{S_2^2 - S_4}{n-1} + \frac{S_2^2 - 2S_4}{(n-1)(n-2)} - \frac{S_2^2}{(n-1)^2} \tag{6.54}$$

As an example consider the sequence of exceedances observed in the Tisza river at Szeged above the level $c = 650$ in the second quarter (from 1st April to 30th June) each year in the period 1901/1970.

Table 6.3

Year	i	X_i	X_i'	Year	i	X_i	X_i'	Year	i	X_i	X_i'
		cm				cm				cm	
1901	1	29	−71	1922	11	124	24	1944	22	4	−96
1902	2	14	−86	1924	12	220	120	1952	23	2	−98
1907	3	108	8	1932	13	273	173	1956	24	39	−61
1912	4	72	−28	1937	14	53	−47	1958	25	37	−63
	5	34	−66								
1914	6	128	28	1940	15	197	97		26	66	−44
1915	7	110	10		16	40	−60	1962	27	170	70
1916	8	73	−27		17	28	−72	1964	28	114	14
1919	9	266	166	1941	18	204	104	1965	29	98	−2
1920	10	16	−84	1942	19	38	−62	1967	30	134	34
					20	51	−49	1970	31	309	209
					21	60	−40				

In the example:

$$\bar{X} = \frac{\sum_{1}^{31} X_i}{31} = 100.35 \text{ cm} \approx 100 \text{ cm}$$

$$R = \sum_{1}^{30} X_i' X_{i+1}' + X_{31}' X_1' = 1186$$

$$E(R) \sim -\frac{S_2}{30} = -7207$$

$$D(R) \approx \sqrt{\frac{S_2^2 - S_4}{30} + \frac{S_2^2 - 2S_4}{30.29} - \frac{S_2^2}{30^2}} \approx 3.7447$$

$$R^* = \frac{R - E(R)}{D(R)} \approx 0.2241.$$

As $|R^*| \approx 1.96$ there is no reason to reject H_0. So, by using the Wald—Wolfowitz test, the conclusion is that in the given case the exceedances above the chosen c level are independent random variables from the same distribution that is they constitute a statistical sample.

6.5.2. TESTING RANDOMNESS ON THE BASIS OF RUN STATISTIC

To check a H_0 hypothesis related to the randomness of a series (I') $X_1, X_2, \ldots, X_n$ the application of the so-called run statistic is also common (see, e.g., Lehmann [A. 17] pp. 313 to 315). The procedure may be, e.g., that first the empirical median of series (I'), that is the M_e value related to which the half of observations is smaller and the other half is greater, is singled out and then each value of X_i is compared to median M_e; if $X_i - M_e > 0$ a figure 1 while if $X_i - M_e < 0$ a zero is written. If H_0 is true, the *series* obtained in the above manner, consisting of zeroes and ones, will behave as such a Bernoulli sequence where for each place the probability of occurrence of both zeroes and ones is equally 1/2. Then the number of all runs in the obtained sequence is determined (see Section 1.2.4).

The procedure is illustrated through an example. Consider the sequence of exceedances observed in the Tisza river at Szolnok in the first quarters from 1903 to 1970 (Table 6.4).

Comparing X_i values to median $M_e = X_{38} = 59$ cm the following sequence will be obtained:

(II) 0 0 0 1 0 1 1 1 1 1 1 1 0 1 0 1 1 1 0 0 0 0

1 1 0 0 0 0 1 0 1 0 0 0 1 1 1 0 0 1

In the sequence the number of 1-s and 0-s is the same: $m = 20$. The number of all runs is: $r = 18$.

Table 6.4

Year	X_i	r_i	Year	X_i	r_i
1903	4	4	1922	134	32
1908	29	13	1926	178	35
1912	2	2	1931	18	9
1914	100	27	1937	150	33
1915	38	16	1940	5	5
1916	178	34		66	24
	88	25	1941	222	38
1919	85	24	1942	128	31
1920	104	28	1945	1	1
	116	30	1946	17	7
1947	24	12	1963	91	26
	22	14	1964	3	3
1948	184	36	1965	17	8
1953	201	37		39	17
1955	20	11	1966	255	39
	45	18	1967	281	40
1957	46	19	1968	63	23
	13	6	1969	59	21 M_e
1958	108	29	1970	19	10
1962	35	15		56	20
				65	22

Wald and Wolfowitz have shown (see [B. 42]) that statistic

$$r^* = \sqrt{2}\,\frac{r-m}{\sqrt{m}} \tag{6.55}$$

has $N(0;1)$ distribution. In the example

$$r^* = \sqrt{2}\,\frac{18-20}{\sqrt{20}} \approx \frac{-2}{3.16} < \frac{-2}{3} = -0.66.$$

Since the actual value of r^* is within the 95 per cent confidence interval $(-1.96; 1.96)$ there is no reason to reject H_0.

As to the power of run statistics and the construction of different run tests when a specified class of alternative hypotheses is dealt with the reader's attention is drawn to the work of Lehmann [A. 17].

As an alternative to hypothesis H_0 which expresses the randomness of series (I') may frequently be supposed the existence of some trend, e.g. an increase in floods. If so, obviously, the assumption $F_1(x)=F_2(x)=\ldots=F_n(x)$ cannot be satisfied.

When the randomness of sequence (I') is to be checked to reveal a trend the use of the following simple nonparametric test is proposed.

Rank the observations $X_1, X_2, \ldots, X_n$ into a sequence in increasing order of magnitude, obtaining thereby an ordered sample,

(I″) $$X_1^* < X_2^* < \ldots < X_n^*.$$

If here $X_1 = X_j^*$ then the rank of X_1 is $r_1 = j$, if $X_2 = X_k^*$ the rank of X_2 is $r_2 = k$, etc. Now, by substituting the indices for the observations made in temporal sequence and by writing below each index the rank attached to the respective observation in the ordered sample, the following table will be obtained:

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ r_1 & r_2 & \ldots & r_n \end{pmatrix}.$$

If in sequence (I′) a trend of rise prevails then, in general, a higher rank number will belong to a higher index and therefore the use of statistic

(6.56) $$d = (r_1 - 1)^2 + (r_2 - 2)^2 + \ldots + (r_n - n)^2 =$$

$$= \sum_1^n r_i^2 - 2\sum_1^n i r_i + \sum_1^n i^2 = 2\sum_1^n i^2 - 2\sum_1^n i r_i$$

seems to be reasonable. Obviously, the more definite the increasing trend in sequence (I′) the less the actual value of d. It can be proved that, if H_0 is true, with large n the distribution of d is asymptotically normal and its expected value and variance are

$$E(d) = \frac{n^3 - n}{6}$$

and

$$D^2(d) = \frac{n^2(n+1)^2(n-1)}{36},$$

respectively (see Lehman [A. 17] p.p. 292).

As an example consider again the data given in Table 6.4 where the ranks for all X_i values are also included.

The calculations lead now to the following results:

$$d = 2\sum_{i=1}^{41} i^2 - 2\sum_{i=1}^{41} i r_i = 10\,782$$

$$E(d) = \frac{41(41^2 - 1)}{6} = 11\,480$$

$$D^2(d) = \frac{41^2 \cdot 42^2 \cdot 40}{36} = 329\,476, \quad D(d) \equiv 575$$

$$d^* = \frac{d - E(d)}{D(d)} = -\frac{698}{575} = -1.2.$$

Since the distribution of d^* is asymptotically $N(0; 1)$ and its actual value falls in the 95 per cent confidence interval there is no reason to reject H_0 due a trend.

a) *Testing the null hypothesis* $P(A)=p$

The problem is to test the null hypothesis

$$H_0: P(A) = p_0$$

set up for an event A, by using observations. Here p_0 is a predetermined, fixed value specified independently of the outcome of experiments either by theoretical considerations or on the basis of previous experience.

Suppose that out of n observations the event A occurred k times while the event $\bar{A}(n-k)$ times. If the question should be answered at $(1-\alpha)$ level then, in the possession of a table covering the binomial distribution, such k_1 and k_2 values can be chosen with which the $\varkappa$ frequency of event A will satisfy the following relation

$$P(k_1 < \varkappa < k_2 | H_0) = 1-\alpha$$

under the conditions

$$P(\varkappa \leqq k_1 | H_0) = P(\varkappa \geqq k_2 | H_0) = \alpha/2.$$

So the critical region is

$$X_k = \{\varkappa \leqq k_1 \text{ or } \varkappa \geqq k_2\}.$$

However, there are only relatively few p and n values included in the table of binomial distribution. When n is large and p is not too small the normal approximation to binomial distribution may be used. So the distribution of the random variable

$$\frac{\varkappa - np}{\sqrt{np(1-p)}}$$

is approximately $N(0; 1)$ and, with a u_α value to be determined from the equation $2\Phi(u_\alpha)-1=1-\alpha$, the relation

$$P\left(-u_\alpha < \frac{\varkappa - np}{\sqrt{np(1-p)}} < u_\alpha | p\right) \approx 1-\alpha$$

will be valid by a good approximation. Hence, at $(1-\alpha)$ level, the following critical region is obtained: $X'_k = \{\varkappa \leqq np_0 - u_\alpha \sqrt{np_0(1-p_0)} \text{ or } \varkappa \geqq np_0 + u_\alpha \sqrt{np_0(1-p_0)}\}$.

The approximation with normal distribution may be considered good if for a given value of p the relation $n \geqq \frac{9}{p(1-p)}$ is satisfied. Consequently, e.g.,

If p	Then minimum n
0.40 or 0.60	38
0.30 or 0.70	43
0.20 or 0.80	56
0.15 or 0.85	71
0.10 or 0.90	100
0.05 or 0.95	189

If the exact value of p is not known but it is known to fall within a given interval then from among n values belonging to the end points the larger one should be taken into account.

When $np<1$ or $n(1-p)<1$ the procedure may be to approach the binomial distribution by means of the Poisson distribution; in this case the corresponding limits of probability have to be taken from the table of Poisson distribution. Here, too, such k_1 and k_2 limits may be specified for the critical value with which it holds that

$$P(\varkappa \leqq k_1|np) = P(\varkappa \geqq k_2|np) = \alpha/2$$

where np is the parameter of the corresponding Poisson distribution.

Finally, it should be noted that for testing a null hypothesis $H_0: p>p_0$ the test will be uniformly the most powerful if $X_k=\{k\leqq k_0\}$ is chosen as critical region. Here p_0 and k_0 are fixed values with which, to meet the requirement that the level of test hould be $(1-\alpha)$, th e condition $P(\varkappa<k_0|p_0)$ must be fulfilled.

s

b) *Testing the equality of two probabilities*

In practice the question of the — unknown — probabilities belonging to two events are equal or not will arise rather frequently. Many times, as far as our knowledge is concerned, both the complete phenomenon and the two events in question had taken place in identical circumstances but still a possibility might emerge that one of the influencing factors was not the same in both cases since, e.g., some change came about meanwhile. In these cases what is checked is whether the magnitude of such a change in the respective condition was significant enough to influence the value of the probability. Such a question might be, e.g., whether the probabilities that flood levels at two gauges of a given river will exceed a certain c level are the same or not.

Accordingly, let the events concerned be denoted by A and B, the probabilities by $P(A)=p_1$ and $P(B)=p_2$ and the null hypothesis by $H_0: p_1=p_2$; the latter, as the common probability is unknown, is a composite hypothesis.

One of the solutions to the problem is — when the number of observations is large — the application of a χ^2 test as it was discussed in Section 6.3.2. Frequently, however the experiments involved are costly so that desirably only a (relatively) small sample ought to be covered. It was R. A. Fisher who elaborated an exact method to this problem, being the examination of the so-called 2×2 contigency table. Other problems also lead to this formula or method; these will be discussed later.

For cases falling in between — that is for medium-size samples — an approximation by means of normal distribution is applied; this procedure will also be returned to later in this section.

Considering now the basic problem suppose that for two phenomena n_1 and n_2 experiments were carried out and for the first pehenomenon event A was observed k_1 times while for the second one event B k_2 times. Write the numbers of observations

in the following table:

k_1	n_1-k_1	n_1
k_2	n_2-k_2	n_2
$K=k_1+k_2$	n_1+n_2-K	n_1+n_2

If the hypothesis $p_1=p_2$ is true then, under the condition that $k_1+k_2=K$, the conditional probability of this division is

$$P(k_1|K, n_1 n_2) = \frac{n_1!\, n_2!\, K!\,(n_1+n_2-K)!}{k_1!\,(n_1-k_1)!\, k_2!\,(n_2-k_2)!\,(n_1+n_2)!}.$$

This distribution constituted the basis to compile the Finney—Latscha—Bennett—Hsu table [C. 2] containing the critical values needed to make decision on hypothesis H_0. In this table for cases where $3 \leqq n_2 \leqq n_1 \leqq 30$ the critical value of a one-sided test is given for the nominal levels $(1-\alpha)=0.95$; 0.975; 0.99 and 0.995 while when $n_2 \leqq \leqq n_1 \leqq 40$ the levels included are $(1-\alpha)=0.95$ and 0.99 only. In the first part of this table the exact values of the errors of the first kind are also given.

The way of using this table is as follows. If against hypothesis H_0 the one-sided alternative $p_1>p_2$ is set up then, since $n_1>n_2$, a relation $k_1 \geqq k_2$ may be supposed (in an opposite case $k_1 n/_1 < k_2/n_2$ so that the alternative should be rejected immediately); for the given n_1, n_2 and k_1 values the table provided such a critical value, k_2^*, which will lead to a difference being just significant that is the critical region will be $x_K=\{k_2 \leqq k_2^*\}$. As K is regarded as a fixed value the critical region for k_1 in the table means that

$$\sum_{k=k_1}^{\min(K, n_2)} P(k|K, n_1, n_2) \leqq \alpha$$

and

$$\sum_{k=k_1-1}^{\min(K, n_1)} P(k|K, n_1, n_2) > \alpha$$

so that, for this k_1, $k_2^*=K-k_1$ and $P(k_2 \leqq k_2^*)=\alpha$.

If, in a case where $n_1>n_2$, against hypothesis H_0 the assumption to be tested were $p_1<p_2$ then the assumption tested would be $(1-p_2)>(1-p_1)$ (that is k_i and n_i-k_i would be interchanged).

In case of two-sided hypotheses a critical region has to be constructed (through the procedure described in the table in detail) that is such (k_1, k_2) related values have to be found — with given $k_1+k_2=K$ — for which k_1 is either too small or too large; the latter involves the procedure described above while in the former case the large values of the difference (n_1-k_1) are considered.

If $n_1 \geqq 40$ the approximation by normal distribution may take place in the following way.

The expected value and variance of the aforementioned distribution are

$$E(k_1|K, n_1, n_2) = K\frac{n_1}{n_1+n_2} = m$$

and

$$D^2(k_1|K, n_1, n_2) = \frac{n_1 n_2 K(n_1+n_2-K)}{(n_1+n_2)^2(n_1+n_2-1)} = \sigma_\alpha^2,$$

respectively. In case where $k_1/n_1 > k_2/n_2$ the quantity

$$u = \frac{k_1 - 0.5 - m}{\sigma}$$

will be normally distributed; to the given null hypothesis the probability of a critical region $x_K = \{K > k_1\}$ (that is the probability of an error of the first kind) for the one-sided test is equal to $1-\Phi(u)$. Here Φ stands for the distribution function of the standardized normal distribution while the subtraction of 0.5 provides an adjustment needed because a discrete variable is replaced by a continuous normal distribution.

Finally, when the total number of sample elements, n_1+n_2, is extremely large then, as it was mentioned, the χ^2 test may be applied.

6.7. ELEMENTS OF THE THEORY OF STATISTICAL DECISION FUNCTIONS

As it was seen, all statistical procedures were finished by some decision. When the problem is to make an estimation the decision to be made is the acceptance of a certain value instead of the true value of a parameter sought. To accept or reject a hypothesis is a typical case of making choice between two alternatives. When a confidence interval is defined the scope of decision is a given interval in which the true parameter should fall. All these decisions are based on statistical samples and therefore they are called statistical decisions. In all cases the tool of decision is a function constituted by the sample elements, i.e., by a statistic. To choose suitable statistics is the fundamental task of mathematical statistics. This problem was discussed in the sections covering the estimation theory and hypothesis testing; the choice of a suitable statistic was justified occasionally by the "mathematically" favourable properties of such a statistic. For instance, when the variance of a random variable was to estimate the proposal was to use the corrected estimate of variance, S_n^{*2}, which is an unbiased and strongly consistent estimate of the variance. However, by using the sample size $X_n^* - X_1^*$ the variance can be estimated in a much simpler manner and with a minimum amount of calculation. For testing hypotheses tests involving small errors of the first and second kind were recommended which, in turn, require sometimes a very large number of observations. Many times economic requirements are also to be considered when choosing a statistical procedure. In the practice of hydrology, including flood hydrology, a decision, which has been made in the course of a procedure of parameter estimation or hypothesis testing, may determine at the same time certain practical measures, "decisions", too, which, due to chance, may be correct or erroneous. A faulty decision may cause a considerable economic loss. Since the basis

of decisions is an actual value (or a statistic composed of actual values) of some random variable observed, the decisions involve certain economic risk. How to quantify the risk of a decision? How to select such a decision whose risk is the possible minimum? These questions can be answered by the aid of the statistical decision theory.

6.7.1. STATISTICAL DECISION PROCEDURE. LOSS FUNCTION AND RISK FUNCTION

An attempt to illustrate the notion of statistical decision and the course of a decision procedure is made below, by using an example.

Suppose that from previous studies it is known that the flood peaks of a certain river follow exponential distribution whose density function is $f(x)=\vartheta e^{-\vartheta x}$, $(x\geqq 0)$, where ϑ is the unknown parameter to be estimated that is the decision to be made relates to the numerical value of ϑ. The result of such a decision is a real number: $\vartheta=d$. When, as a result of estimation, a verdict that $\vartheta=d$ is returned, a statistical decision is made. In the example d can be any of the points along the positive half of the real line. So the set of all possible decisions is $\mathcal{D}=\{0\leqq d<\infty\}$. The set $\mathcal{D}$ is called decision space. Consider now what is the basis to make a decision that $\vartheta=d$. The value of parameter ϑ cannot be observed directly (if could be, estimation would not be needed). Instead, observations can be made on the value of a random variable X representing the peak, usually many times. Denote by $X=(X_1, X_2, \ldots, X_n)$ the sample representing a random variable X. Since the distribution of the random variable X depends on a parameter ϑ each observed value of X contains some information on ϑ. The vector variable X, that is the sample, contains, of course, even more information on ϑ than does one single observation. It is this information that will be utilized to make a decision on ϑ that is the value of X will be decisive to the choice that which one from the decisions $\vartheta=d$ will be made. So d is a function of X: $d=\delta(X)$. The rule $\delta(X)$ that is the instruction specifying what a number d should be coupled with the observed value of X is called decision function. As in the example ϑ is the reciprocal expected value of the exponential distribution, a reasonable decision rule may be constituted, e.g., by the following variants:

$$\delta_1(X)=\frac{1}{\sum_1^n X_i}=\frac{1}{\overline{X}},$$

$$\delta_2(X)=\frac{1}{\sum_1^n p_i X_i} \quad \text{where} \quad \sum_1^n p_i=1, \quad p_2\geqq 0,$$

$$\delta_3(X)=\frac{1}{X_1^*+X_n^*}.$$

For sake of simplicity later on the random variable providing the basis of decision will be denoted by X even when it is a vector variable. If $X=(X_1, X_2, \ldots, X_n)$ denotes a sample then the information included therein will be compressed into a certain — possibly a sufficient — statistic $t(X)=t(X_1, \ldots, X_n)$ which itself is also a random variable and which will serve as a basis for making decision on ϑ.

Based on the above example the decision procedure — although somewhat abstractly but with a rather general validity — may be formulated as follows. Given the sample space $\mathfrak{X}$, observations are made on an element $X \in \mathfrak{X}$ thereof and a choice is made for a decision d from the set of the possible decisions $\mathfrak{D}$ defined by the practical problems. The set $\mathfrak{D}$ is called decision space. Selection from the decisions takes place on the basis of a certain rule. This rule is called decision function and is denoted by $\delta(X)$. A decision rule $\delta(X)$ states: the decision $d \in \mathfrak{D}$ to be chosen is the one belonging to an observed X that is $d=\delta(X)$.

Loss function and risk function.

With given X the different decision functions $\delta_i(X)$ provide in general different numerical values for d so that a question if which of them should be chosen may well arise. To answer this question a check on the consequences going together with the decision is necessary. In certain cases estimations can be made on the magnitude of a damage or loss caused by a wrong decision when ϑ is not equal to d. Obviously, the greater the difference between the numerical value d belonging to a certain decision and the true value ϑ the greater the loss which can be caused if it is a wrong decision. The sizing of levees or other flood control activities can be mentioned as examples where the starting point is the distribution of flood peaks. Since when X is known the numerical value of a decision, $\vartheta=d$, will depend solely on whether what a $\delta(X)$ decision function is chosen, the loss caused by the decision is a function of the decision function, $\delta(X)$. Let the loss be denoted by $L[\vartheta, \delta(X)]$ if ϑ stands for the true value of the parameter and if the decision function used is $\delta(X)$ and let $X=x$.

Such losses could be, e.g.:

$$L[\vartheta, \delta(X)] = [\vartheta - \delta(X)]^2$$

or

$$L[\vartheta, \delta(X)] = |\vartheta - \delta(X)|.$$

As X is a random variable, the same decision function, due to chance, may provide different numerical values for d so that with a given decision function different losses may be obtained. Therefore, when making judgement on the efficiency of a given decision function, for the approximation of the true value of ϑ the average loss involved by a given decision function should be taken into account.

The average loss caused by a given decision function, $\delta(X)$, is

$$R_{\delta(X)}(\vartheta) = E_\vartheta[L/\vartheta, d(x)] = \int_{\mathfrak{X}} L[\vartheta, \delta(x)]\, dF(x; \vartheta).$$

This R is function of ϑ and is called the risk function of the decision function. Concludingly, a risk function is the conditional expected value of the decision function

under the condition that δ is the true value of the parameter. (Here $F(x;\vartheta)$ is the distribution function of a random variable X with a hypothesized value of ϑ.)

Supposedly, the foregoing has made it clear that the problem of parameter estimation is a case of statistical decisions. As it was mentioned in Section 4.1 the theory of estimation and hypothesis testing had been integrated by Abraham Wald into the theory of decision functions.

Examine now the case of hypothesis testing as a problem of decision making!

Consider the following example. It was mentioned in Section 6.3.4 that the maximum stages of River Danube followed the normal distribution, $N(\vartheta;\sigma_0)$. Suppose that σ_0 is known and a decision on parameter ϑ will be made through choices from the hypotheses $H_0: \vartheta \leqq \vartheta_0$ and $H_1: \vartheta > \vartheta_0$. Now a statistician faces two possible decisions: either H_0 or H_1 is accepted. Acceptance of H_0 is the decision d_1 while that of H_1 is d_2. So the decision space consists of two points: $\mathcal{D}=\{d_1, d_2\}$.

When the basis of decision is a sample, $X=(X_1, X_2, \ldots, X_n)$, the sample space $\mathfrak{X}$ is an n-dimensional Euclidean space, R_n. The set Θ defined by the possible values of parameter ϑ (the so-called parameter space) is $\Theta=\{-\infty, +\infty\}$. The set of the possible $\delta(X)$ decision functions consists of all functions which transform the sample space, $\mathfrak{X}$, into set $\mathcal{D}=\{d_1, d_2\}$ and which have the properties that the relation $P_\vartheta[d(x)=d_1]$ will be unambiguously defined for all $\vartheta \in \Theta$.

Let the loss function be

$$L(\vartheta; d_1) = \begin{cases} l_1 & \text{if } \vartheta > \vartheta_0 \\ 0 & \text{if } \vartheta \leqq \vartheta_0 \end{cases}$$

$$L(\vartheta; d_2) = \begin{cases} 0 & \text{if } \vartheta > \vartheta_0 \\ l_2 & \text{if } \vartheta \leqq \vartheta_0 \end{cases} \tag{6.57}$$

where l_1 and l_2 are positive numbers*. (In general the loss function is selected so as to obtain zero loss for the correct decision.)

In this case the risk function can be calculated easily:

$$R[\vartheta, d_{(x)}] = \begin{cases} l_1 P_\vartheta[\delta(X) = d_1] & \text{if } \vartheta > \vartheta_0 \\ l_2 P_\vartheta[\delta(X) = d_2] & \text{if } \vartheta \leqq \vartheta_0. \end{cases} \tag{6.58}$$

Two types of error can be made in the course of decision making. If $\vartheta > \vartheta_0$ the probability of making an error by choosing a decision d_1 is $P[\delta(X)=d_1]$, since a decision d_2 ought to be chosen, provided that ϑ is the true value of the parameter. Similarly, if $\vartheta \leqq \vartheta_0$ the probability of making an error by choosing a decision d_2 is

$$P_\vartheta[\delta(X) = d_2] = 1 - P_\vartheta[\delta(X) = d_1],$$

* It is easy to see that loss (6.57) fails to measure the consequence caused by the magnitude of the difference between $\delta_{(x)}$ and ϑ so that for representing the loss this choice will not be appropriate in all cases.

since a decision d_1 ought to be chosen, provided that ϑ is the true value of the parameter.

The above example was to demonstrate that the problem of hypothesis testing may also be discussed in the frame of statistical decision theory. To readers interested in the details of this subject-matter books [A. 7] and [A. 25] are recommended.

6.7.2. PRINCIPLES OF CHOOSING A SUITABLE DECISION FUNCTION. BAYES'S PRINCIPLE FOR DECISION MAKING

It was outlined in the foregoing that when a decision was to be made on the ϑ parameter of a random variable X whose distribution function was $F(x; \vartheta)$ and a decision function $\delta(X)$ was used then the risk function,

$$P\delta_{(X)}(\vartheta) = E_\vartheta[L(\vartheta, d(x)] = \int_{\mathfrak{X}} L[\vartheta, \delta(x)]\, dF(x; \vartheta),$$

was function of parameter ϑ; this means that, depending on the true value thereof, the risk of a given decision function $\delta(X)$ may vary widely.

Now if the risk of two different decision functions, say $\delta_0(x)$ and $\delta_1(x)$, is quantified and with a given ϑ^*

$$R_{\delta_0(x)}(\vartheta^*) < R_{\delta_1(x)}(\vartheta^*) \tag{6.59}$$

is found then it is said that with ϑ^* the estimate $\delta_0(x)$ is better than $\delta_1(x)$. The decision function $\delta_0(x)$ is called uniformly better than $\delta_1(x)$ if the relation

$$R_{\delta_0(x)}(\vartheta) \leqq R_{\delta_1(x)}(\vartheta) \tag{6.60}$$

holds for all ϑ and the relation (6.59) is satisfied for at least one ϑ^*.

If $R_{\delta_0(x)}(\vartheta) = R_{\delta_1(x)}(\vartheta)$ holds for all ϑ then the decision functions $\delta_0(x)$ and $\delta_1(x)$ are called equivalent.

When in a given decision problem such a decision function, $\delta_0(x)$, exists with which the relation (6.60) are satisfied for any other decision functions, $\delta_1(x)$, then $\delta_0(x)$ is said to be the uniformly best decision function.

If in a given decision problem it is found that $\delta_0(x)$ is a uniformly better decision function than $\delta_1(x)$ and the application of the latter cannot be taken into account then $\delta_1(x)$ is called inadmissible decision function. However, when no uniformly best decision function can be found in a given decision problem then the best thing what can be done is to use an admissible decision function. If both $\delta_1(x)$ and $\delta_2(x)$ are admissible decision functions this means that for certain ϑ values $\delta_1(x)$ is better than $\delta_2(x)$. Which one of these decision functions should be chosen in such a case?

If ϑ were known the choice between $\delta_1(x)$ and $\delta_2(x)$ would be easy but in this case no decision would be needed. In the knowledge of at least the probability by which ϑ falls into a given interval the decision would be easier.

Suppose that parameter ϑ is a random variable and that the *a priori* distribution of ϑ is known, e.g., from previous experience. Denote by $\tau(\vartheta)$ the distribution function of

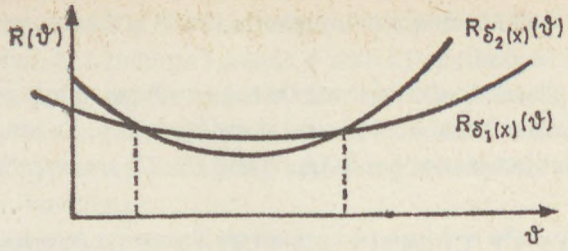


Figure 65

ϑ in the parameter space Θ. Now, having a given *a priori* distribution, $\tau(\vartheta)$, such a decision function is looked for with which the so-called Bayes risk

$$r(\tau, \delta) = \int_{\Theta} R_{\delta(x)}(\vartheta)\, d\tau(\vartheta), \tag{6.61}$$

will be minimum, where the risk function in the integrand is

$$R_{\delta(x)}(\vartheta) = \int_{\mathfrak{X}} L[\vartheta, \delta(x)]\, dF(x|\vartheta).$$

The decision $\delta(x)$ with which risk (6.61) is minimum is called a Bayes decision belonging to the *a priori* distribution, $\tau(\vartheta)$. A Bayes decision is therefore a decision of minimum Bayes' risk.

To calculate Bayes' risk as given in (6.61) consider the following continuous analogons of the theorem of total probability:

$$F(x) = \int_{\Theta} F(x|\vartheta)\, d\tau(\vartheta)$$

and

$$\tau(\vartheta) = \int_{\mathfrak{X}} \tau(\vartheta|x)\, dF(x).$$

By virtue of these, the expression

$$r(\tau, \delta) = \int_{\Theta} \Big[\int_{\mathfrak{X}} L[\vartheta, \delta(x)]\, dF(x|\vartheta)\, d\tau(\vartheta) = \int_{\Theta} \Big[\int_{\mathfrak{X}} L[\vartheta, \delta(x)]\, d\tau(\vartheta|x)\, dF(x)$$

yields the Bayes' risk to be minimized (supposing that the sequence of integration within the double integral above may be interchanged).

Instead of the denotation used in the above Stieltjes integral a denotation using density functions, the Riemann integral, is more common:

Let

$$\frac{d\tau(\vartheta)}{d\vartheta} = g(\vartheta); \quad \frac{d\tau(\vartheta|x)}{d\vartheta} = g(\vartheta|x);$$

$$\frac{dF(x)}{dx} = f(x)$$

then

$$r(\tau, \vartheta) = \int_{\mathfrak{X}} \Big[\int_{\Theta} L[\vartheta, \delta(x)] g(\vartheta|x)\, d\vartheta \Big] f(x)\, dx. \tag{6.62}$$

So to minimize the quantity $r(\tau, \delta)$ expressed by the double integral the internal one,

$$\int_{\Theta} L[\vartheta, \delta(x)]g(\vartheta|x)\,d\vartheta = E_{\vartheta}[L(\vartheta, d)|X = x],$$

has to be minimized.

The conditional density function $g(\vartheta|x)$ is called the *a posteriori* density function of the random variable ϑ. So a Bayes decision is such a $\delta(x)$ decision function which will minimize the *a posteriori* (conditional) loss.

Consider now the case where

$$L[\vartheta, \delta(x)] = (\vartheta - d)^2, \quad \text{where} \quad d = \delta(x) \quad \text{if} \quad X = x.$$

Here Bayes decision will be a decision with which

$$E_{\vartheta} = [L(\vartheta, \delta(x)|X = x)] = \int_{\Theta} (\vartheta - d)^2 g(\vartheta|x)\,d\vartheta = \min.$$

Then

$$\frac{\partial E_{\vartheta}[L(\vartheta, d)|X = x]}{\partial d} = -2\int_{\Theta} [\vartheta - d]g(\vartheta|x)\,d\vartheta = 0.$$

Hence

$$d = \frac{\int_{\Theta} \vartheta g(\vartheta|x)\,d\vartheta}{\int_{\Theta} g(\vartheta|x)\,d\vartheta} = \int_{\Theta} \vartheta g(\vartheta|x)\,d\vartheta.$$

Consequently, when calculated with the given *a posteriori* density function, the conditional expected value of parameter ϑ is $\delta(x) = d = E(\vartheta|X = x)$.

If in a parameter estimation the loss function is

$$L[\vartheta, \delta(x)] = |\vartheta - \delta(x)|$$

then the Bayes decision will be the median of the *a posteriori* distribution of parameter ϑ. Namely, in this case the expression

$$E_{\vartheta}[L(\vartheta, d)|X = x] = \int_{\Theta} |\vartheta - d| g(\vartheta|x)\,d\vartheta \tag{6.63}$$

has to be minimized, where $d = \delta(x)$. Suppose that $\Theta = (-\infty, \infty)$ then

$$\int_{-\infty}^{\infty} |\vartheta - d| g(\vartheta|x)\,dx = \int_{-\infty}^{d} (d - \vartheta) g(\vartheta|x)\,d\vartheta + \tag{6.64}$$

$$+ \int_{d}^{\infty} (\vartheta - d) g(\vartheta|x)\,d\vartheta = d\int_{-\infty}^{d} g(\vartheta|x)\,d\vartheta -$$

$$- \int_{-\infty}^{d} \vartheta g(\vartheta|x)\,d\vartheta + \int_{d}^{\infty} \vartheta g(\vartheta|x)\,d\vartheta - d\int_{d}^{\infty} g(\vartheta|x)\,d\vartheta = \min.$$

For the *a posteriori* expected value of ϑ introduce now the denotation $\int_{-\infty}^{\infty} \vartheta g(\vartheta|x)d\vartheta =$ $=m$ and take into account that $\int_{-\infty}^{\infty} g(\vartheta|x)d\vartheta = 1$, so Eq. (6.64) may be written as

$$\int_{-\infty}^{\infty} |\vartheta - d|\, g(\vartheta|x)\, d\vartheta = d \int_{-\infty}^{d} g(\vartheta|x)\, d\vartheta - \tag{6.65}$$

$$- \int_{-\infty}^{d} \vartheta g(\vartheta|x)\, d\vartheta + m - \int_{-\infty}^{d} \vartheta g(\vartheta|x)\, d\vartheta -$$

$$-d\left[1 - \int_{-\infty}^{d} g(\vartheta|x)\, d\vartheta\right] = \min.$$

The minimum of Eq. (6.65) with respect to d may be at the same location where the derivate with respect to d is zero. Produce, therefore, the derivate of Eq. (6.65) with respect to d:

$$\int_{-\infty}^{d} g(\vartheta|x)\, d\vartheta + dg(d|x) - dg(d|x) -$$

$$-dg(d|x) - 1 + \int_{-\infty}^{\infty} g(\vartheta|x)\, d\vartheta + dg(d|x) = 0.$$

Hence

$$\int_{-\infty}^{d} g(\vartheta|x)\, d\vartheta = \frac{1}{2}.$$

Consequently, d is the median of the *a posteriori* distribution of parameter ϑ.

Note that in the possession of the *a posteriori* distribution of ϑ the Bayes risk can be calculated and the Bayes decision can be found even when ϑ is not a parameter of a random variable X. If both X and ϑ are random variables with interdependence, stochastic relation between them, and both the conditional distributions related to one another and the distribution of one of them is known then the procedure described above is applicable. Following an example is presented on the application of Bayes's decision principle in the hydrology of floods.

Example on the application of Bayes's decision principle

Suppose that dealing with a flood wave observed at a gauge of a certain river the decision (action) to be made expediently is d_1 if the travel time is shorter than two weeks and d_2 if it is longer (where d_1 and d_2 may be decisions on certain protection works, e.g., levee reinforcement, transportation of materials needed for protection, deployment of flood fighting forces, etc.). During a flood event the actions mentioned should, of course, be taken as soon as possible. In Chapter 8 it will be seen that there is a rather close interrelation between the peak value of a flood wave and the travel time: knowing the peak level conclusion can be drawn on the travel time of

a flood wave. As, roughly, the peak is attained at the half of the travel time it is expedient that the information contained in the peak value be utilized to decide whether d_1 or d_2 should be chosen as a decision. It is said that if the travel time is shorter than two weeks the nature is in state ϑ_1 while if it is longer the state of nature is ϑ_2. In this simplified model the set of the possible states of nature is $\Theta=\{\vartheta_1, \vartheta_2\}$ and the set of possible decisions is $\mathcal{D}=\{d_1, d_2\}$.

Let the loss matrix be the following:

	d_1	d_2
1	0	1
2	2	0

that is $L(\vartheta_1, d_1)=0$, $L(\vartheta_1, d_2)=1$, $L(\vartheta_2, d_1)=2$, $L(\vartheta_2, d_2)=0$. (In reality numbers 1 and 2 represent certain amounts, say some million forints.) With consideration to the peak values let $X=1$ if the peak value is lower than c metres and let $X=2$ otherwise.

In the given example four decision rules can be set up, let they be the following:

$$\delta_1(X=1)=d_1, \quad \delta_1(X=2)=d_1$$

$$\delta_2(X=1)=d_1, \quad \delta_2(X=2)=d_2$$

$$\delta_3(X=1)=d_2, \quad \delta_3(X=2)=d_1$$

$$\delta_4(X=1)=d_2, \quad \delta_4(X=2)=d_2.$$

These decision rules can be interpreted in the following way: since, in general, to a higher peak a longer travel time will belong, it may be said that nature, through the random variable X, informs us on its actual state: ϑ, that is with $X=1$ an outcome $\vartheta=\vartheta_1$ may be expected when the correct decision is d_1. However, a swindle of nature may also occur when with $X=1$ the outcome is $\vartheta=\vartheta_2$ and the decision to be made would be d_2. Furthermore, it is also possible that with $X=2$ $\vartheta=\vartheta_1$ when the action d_1 ought to be taken, etc. Now if the decision function $\delta_1(X)$ is applied, independently of peak value X the action (decision) chosen will always be d_1 while with the decision function $\delta_4(X)$ always the decision d_2 will be made, i.e., the information given by nature through the random variable X on its future state ϑ will be disregarded. The decision function $\delta_2(X)$ reflects our belief that what is said by nature is true while $\delta_3(X)$ represents that nature is always misleading. In fact the situation is that, if our action taken relative to ϑ is rendered dependent on the random variable X, in certain cases nature will tell the truth and in other cases it will be misleading.

For instance, as to the Tisza river, on the basis of flood waves observed so far (first quarters from 1900 to 1970, Table T.1) the distribution of the random variable

X — given its true value as a condition — was found to be

$$
\begin{aligned}
&P(X=1|\vartheta=\vartheta_1)=3/4,\\
&P(X=2|\vartheta=\vartheta_1)=1/4,\\
&P(X=1|\vartheta=\vartheta_2)=1/3,\\
&P(X=2|\vartheta=\vartheta_2)=2/3.
\end{aligned}
\tag{6.66}
$$

Calculate now the risk of the decision rules $\delta_1(X)$, $\delta_2(X)$, $\delta_3(X)$, $\delta_4(X)$! The risk is the conditional expected value of losses where the condition is the true state of nature, $\vartheta=\vartheta_2$.

$$
\begin{aligned}
R(\vartheta_1,\delta_1)&=L[\vartheta_1,\delta_1(X=1)]P(X=1|\vartheta=\vartheta_1)+\\
&+L[\vartheta_1,d_1(X=2)]P(X=2|\vartheta=\vartheta_1)=\\
&=0\cdot 3/4+0\cdot 1/4=0.\\
R(\vartheta_1,\delta_2)&=L[\vartheta_2,\delta_1(X=1)]P(X=1|\vartheta=\vartheta_2)+\\
&+L[\vartheta_2 d_1(X=2)]P(X=2|\vartheta=\vartheta_2)=\\
&=2\cdot 1/3+2\cdot 2/3=2.\\
R(\vartheta_1,\delta_2)&=L[\vartheta_1,d_2(X=1)]P(X=1|\vartheta=\vartheta_1)+\\
&+L[\vartheta_1,d_2(X=2)]P(X=2|\vartheta=\vartheta_1)=\\
&=0\cdot 3/4+1\cdot 1/4=1/4.\\
R(\vartheta_2,\delta_2)&=L[\vartheta_2,\delta_2(X=1)]P(X=1|\vartheta=\vartheta_2)+\\
&+L[\vartheta_2,\delta_2(X=2)]P(X=2|\vartheta=\vartheta_2)=\\
&=2\cdot 1/3+0\cdot 2/3=2/3.\\
R(\vartheta_1,\delta_3)&=L[\vartheta_1,\delta_3(X=1)]P(X=1|\vartheta=\vartheta_1)+\\
&+L[\vartheta_1,\delta_3(X=2)]P(X=2|\vartheta=\vartheta_1)=\\
&=1\cdot 3/4+0\cdot 1/4=3/4.\\
R(\vartheta_2,\delta_3)&=L[\vartheta_2,\delta_3(X=1]P(X=1|\vartheta=\vartheta_2)+\\
&+L[\vartheta_2,d_3(X=2)]P(X=2|\vartheta=\vartheta_2)=\\
&=0\cdot 1/3+2\cdot 2/3=4/3.\\
R(\vartheta_1,\delta_4)&=L[\vartheta_1,\delta_4(X=1)]P(X=1|\vartheta=\vartheta_1)+\\
&+L[\vartheta_1,\delta_4(X=2)]P(X=2|\vartheta=\vartheta_1)=\\
&=1\cdot 3/4+1\cdot 1/4=1.\\
R(\vartheta_2,\delta_4)&=L[\vartheta_2,\delta_4(X=1)]P(X=1|\vartheta=\vartheta_2)+\\
&+L[\vartheta_2,\delta_4(X=2)]P(X=2|\vartheta=\vartheta_2)=\\
&=0\cdot 1/3+0\cdot 2/3=0.
\end{aligned}
$$

On the basis of results obtained a choice from the decision rules δ_1, δ_2, δ_3 and δ_4 would be a hard one as they have rather different risks depending on whether the true state of nature is $\vartheta=\vartheta_1$ or $\vartheta=\vartheta_2$. So no decision can be made on whether in the choice of actions the observed value of the random variable X (peak value) should be disregarded or not and, if considered, be it trusted or not. In any case, inspecting the distribution what is seen is that the random variable X has given a good information on ϑ in three fourth of the cases when $\vartheta=\vartheta_1$ and in two third of the cases when $\vartheta=\vartheta_2$. If the probability by which ϑ takes the value of ϑ_1 or ϑ_2 were known the choice from the decision rules δ_1, δ_2, δ_3 and δ_4 would be easier. Remaining at the above example, as far as the flood waves of the Tisza river observed at Tokaj are concerned, the distribution of ϑ determined through the relative frequencies has been obtained as

$$P(\vartheta=\vartheta_1)=4/5;\ P(\vartheta=\vartheta_2)=1/5. \tag{6.67}$$

On this basis the average risks belonging to each of the decision functions $\delta_i(X)$, $(i=1, 2, 3, 4)$, can be quantified as follows:

$$r(\vartheta,\delta_1)=R(\vartheta_1,\delta_1)P(\vartheta=\vartheta_1)+R(\vartheta_2,\delta_1)P(\vartheta=\vartheta_2)=$$

$$=0\cdot\frac{4}{5}+2\cdot\frac{1}{5}=\frac{6}{15}$$

$$r(\vartheta,\delta_2)=R(\vartheta_1,\delta_2)P(\vartheta=\vartheta_1)+R(\vartheta_2,\delta_2)P(\vartheta=\vartheta_2)=$$

$$=\frac{1}{4}\cdot\frac{4}{5}+\frac{2}{3}\cdot\frac{1}{5}=\frac{5}{15}$$

$$r(\vartheta,\delta_3)=R(\vartheta_1,\delta_3)P(\vartheta=\vartheta_1)+R(\vartheta_2,\delta_3)P(\vartheta=\vartheta_2)=$$

$$=\frac{3}{4}\cdot\frac{4}{5}+\frac{4}{3}\cdot\frac{1}{5}=\frac{13}{15}$$

$$r(\vartheta,\delta_4)=R(\vartheta_1,\delta_4)P(\vartheta=\vartheta_1)+R(\vartheta_2,\delta_4)P(\vartheta=\vartheta_2)=$$

$$=1\cdot\frac{4}{5}+0\cdot\frac{1}{5}=\frac{12}{15}$$

The average risk, $r(\vartheta,\delta_i)=E_\vartheta[R(\vartheta,\delta_i)]$, is called the Bayes' risk of decision δ_i. As it can be seen in the example the decision rule involving the least Bayes' risk is $\delta_2(X)$.

In Bayes's decision theory distribution (6.67) is called the *a priori* distribution of ϑ. It may occur, of course, that the *a priori* distribution of ϑ is not available. In this case Bayes's principle for decision making cannot be applied. Instead, to select an appropriate decision function, the procedure may be as follows:

Find the maximum risk of the decision function $\delta_i(X)$, $(i=1, 3, 2, 4)$, that is find the risk

$$\max_j L[\vartheta_j,\delta_i(X)],\quad (j=1, 2).$$

In the example

$$\max_j L[\vartheta_j, \delta_1(X)] = L[\vartheta_2, \delta_1(X)] = 2$$

$$\max_j L[\vartheta_j, \delta_2(X)] = L[\vartheta_2, \delta_2(X)] = 2/3$$

$$\max_j L[\vartheta_j, \delta_3(X)] = L[\vartheta_2, \delta_3(X)] = 4/3$$

$$\max_j L[\vartheta_j, \delta_4(X)] = L[\vartheta_1, \delta_3(X)] = 1.$$

Now the chosen decision function, $\delta_i(X)$, will be the one with which the maximum risk is minimum:

$$\min_i \max_j L[\vartheta_j, \delta_i(X)] = L[\vartheta_2, \delta_2(X)] = 2/3.$$

This principle is called *minimax principle*. It is seen that it is the decision function $\delta_2(x)$ that gives an optimum (has the lowest risk) also in the sense of the minimax principle, similarly as in the case of Bayes's principle.

PART III

STOCHASTIC RELATIONS BETWEEN RANDOM VARIABLES

CHAPTER 7

7.1. CORRELATION ANALYSIS

7.1.1. MEASURING STOCHASTIC RELATIONS

Both in research and practice of hydrology the examination of relations between two or more random variables is a fundamental task. This problem is also of great importance in the hydrology of floods. A stage or flow observed at a gauge of a certain river depends greatly on the depth of precipitation fallen on the catchment, on the stages and flows of tributaries, on temperature, on runoff conditions, etc.

The methods serving the examination of relations between random variables is summarized commonly under the title of correlation and regression theory. This sphere of problems includes tasks of rather different nature whose common feature is that the qualitative and quantitative properties of relations between quantities are examined. The so-called correlation theory deals primarily with the closeness of relations and intends above all to decide whether a relation exists between the random variables or they are independent of (or at least uncorrelated to) one another. If some relation does exist the next question is whether this relation is loose or close, or possibly, it is functional. The latter is regarded to be the closest relation.

The main objective of regression theory is to construct such a "functional relation" between two random variables.

7.1.2. THE CORRELATION COEFFICIENT

The joint behavior and interrelation of two random variables is described completely by a bivariate joint distribution function or density function. However, on the one hand, in practice this function may be considered known in very rare cases only and, on the other hand, even if it is known, it is desirable to have a small number of characteristics which informs us sufficiently on the relation between the random variables concerned.

The notion of correlation coefficient associated with two random variables, X and Y, was described in Section 2.1.7 as

$$\varrho = \frac{E\{[X-E(X)][Y-E(Y)]\}}{D(X)D(Y)} = \frac{E(XY)-E(X)E(Y)}{D(X)D(Y)} = E(X^*Y^*) \tag{7.1}$$

where

$$X^* = \frac{X - E(X)}{D(X)} \quad \text{and} \quad Y^* = \frac{Y - E(Y)}{D(Y)}$$

are the respective standardized variables.

If $\varrho = 0$ the variables X and Y are called *uncorrelated.* ϱ will be equal to zero when X and Y are independent. So when certain random variables are independent they are at the same time uncorrelated but in the reverse case, when X and Y are uncorrelated, their independence, in general, does not hold.

As it was pointed in Section 2.1.7 if the joint distribution of X and Y is a two-dimensional normal distribution their uncorrelatedness implies their independence. The situation is the same when X and Y are indicator variables belonging to two events.

The value of the correlation coefficient ϱ may vary between -1 and $+1$. If $|\varrho| = 1$ then a linear relation of the form $Y = aX + b$ prevails between X and Y that is Y is determined definitely by X. The closer $|\varrho|$ to 1 the more linear the character of relation between the two variables.

In cases where the distribution is two-dimensional normal, the value of $|\varrho|$ is a good measure of the closeness of their relation. The greater the departure of $|\varrho|$ from zero the closer the relation between the two random variables concerned. When the joint distribution of (X, Y) is not a two-dimensional normal then the conclusion on the closeness of relation between them is not always so clear. Therefore a correlation coefficient reflects the linearity of a relation rather than the closeness thereof.

Correlation coefficient ϱ is estimated from sample $(X_1, Y_1); (X_2, Y_2); \ldots; (X_n, Y_n)$ by using the following called empirical correlation coefficient

$$r = \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_1^n (X_i - X)^2 \sum (Y_i - Y)^2}} = \frac{m_{11}}{S_1 S_2}, \tag{7.6}$$

where

$$\bar{X} = \frac{\sum X_i}{n}; \quad \bar{Y} = \frac{\sum Y_i}{n};$$

$$S_1^2 = \frac{\sum (X_i - \bar{X})^2}{n}; \quad S_2^2 = \frac{\sum (Y_i - \bar{Y})^2}{n};$$

$$m_{11} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}.$$

In case of large samples the following asymptotic formulae are valid:

$$E(r) \approx \varrho; \quad D^2(r) \approx \frac{(1 + \varrho^2)^2}{n}. \tag{7.7}$$

So the empirical correlation coefficient, r, is an estimation of the theoretical correlation coefficient, ϱ, which is asymptotically unbiased and asymptotically strongly consistent.

In case of two-dimensional *normal distribution* the density function of r is

$$f_n(z;\varrho) = \frac{n-2}{\pi}(1-\varrho^2)^{\frac{n-1}{2}}(1-z^2)^{\frac{n-4}{2}} \cdot \tag{7.8}$$

$$\cdot \int_0^1 \frac{x^{n-2}}{(1-\varrho xz)^{n-1}} \cdot \frac{dx}{\sqrt{1-x^2}}$$

where ϱ is the theoretical correlation coefficient. If $\varrho=0$ the density function (7.8) will be simplified into the following form:

$$f_n(z;0) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}(1-z^2)^{\frac{n-4}{2}}. \tag{7.9}$$

This means that even in the case of independence ($\varrho=0$) it cannot be expected that the r value calculated from the sample will be zero. For cases where $\varrho=0$ it can be proved, that for the two-dimensional normal distribution, statistic

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \tag{7.10}$$

is distributed according to the Student law with parameter $(n-2)$. Consequently, when a joint normal distribution is the case statistic (7.10) is suitable to test the hypothesis of independence.

If the null hypothesis

$$H_0: P(X < x, Y < y) = P(X < x)P(Y < y)$$

is considered then the t value calculated according to (7.10) has to be compared with the interval $(-t_\varepsilon, t_\varepsilon)$ selected from the table of Student distribution for which

$$P(-t_\varepsilon \leqq t < t_\varepsilon) = 1-\varepsilon.$$

If the actual value of t falls outside interval $(-t_\varepsilon, t_\varepsilon)$ the hypothesis of independence will be rejected at level $(1-\varepsilon)$.

When the joint distribution of X and Y is not a two-dimensional normal then for statistical purposes a transformation introduced by R. A. Fischer:

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r} \tag{7.11}$$

is advisable. In the case of large n, under rather general conditions, the distribution

of the random variable Z is approximately normal with expected value and variance

$$E(Z) \approx \frac{1}{2} \ln \frac{1+\varrho}{1-\varrho} + \frac{\varrho}{2(n-1)} \tag{7.12}$$

and

$$D^2(Z) \approx \frac{1}{n-3},$$

respectively. The same transformation as (7.11) may be used when two empirical correlation coefficients are to be compared, as below.

Consider samples of elements n_1 and n_2, respectively, taken for the pairs of random variables (X, Y) and (U, V):

I. $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{n_1}, Y_{n_1})$

II. $(U_1, V_1)(U_2, V_2), \ldots, (U_{n_2}, V_{n_2})$.

Denote the correlation coefficient of the pair of variables (X, Y) by $\varrho_1 = \varrho_1(X, Y)$ and that of (U, V) by $\varrho_2 = \varrho_2(U, V)$. Now the hypothesis to be tested is $H_0 : \varrho_1 = \varrho_2$.

Calculate from sample I the estimate $r_1 \approx \varrho_1$ and from sample II the estimate $r_2 \approx \varrho_2$. Random variables

$$Z_1 = \frac{1}{2} \ln \frac{1+r_1}{1-r_1} \quad \text{and} \quad Z_2 = \frac{1}{2} \ln \frac{1+r_2}{1-r_2}$$

can be regarded as they come — by a good approximation — from normal distribution with variances

$$D^2(Z_1) = \frac{1}{n_1-3} \quad \text{and} \quad D^2(Z_2) = \frac{1}{n_2-3},$$

respectively. If hypothesis H_0 is true the expected values, $E(Z_1)$ and $E(Z_2)$, are equal and, consequently, random variable

$$W = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

has distribution $N(0, 1)$. So if the value of W falls into interval $(-2, +2)$ hypothesis H_0 may be accepted at a level 95 per cent.

The procedure given above may be used, e.g., to decide whether the coefficient of correlation between flood peaks and flood durations is the same or not at two different gauges.

The comparison of more than two empirical correlation coefficients is also possible in the following way: denote the empirical correlation coefficients calculated from samples consisting of $n_1, n_2, \ldots, n_k$ elements by $r_1, r_2, \ldots, r_k$ and the correspond-

ing theoretical values by $\varrho_1, \varrho_2, \ldots, \varrho_k$. To check null hypothesis $H_0: \varrho_1 = \varrho_2 = \ldots = \varrho_k$ calculate the statistic

$$\chi^2 = \sum_{i=1}^{k} (n_i - 3)(Z_i - \bar{Z})^2$$

which follows χ^2 distribution with parameter $(k-1)$. Here

$$Z_i = \frac{1}{2} \ln \frac{1+r_i}{1-r_1} \quad \text{and} \quad \bar{Z} = \frac{\sum_{i=1}^{k} (n_i - 3) Z_i}{\sum_{i=1}^{k} (n_i - 3)}.$$

If $\chi^2 < \chi^2(\varepsilon)$ that is when the calculated χ^2 is less than the critical value contained in Table T.5 for $(1-\varepsilon)$ then H_0 is accepted at this level. (Note that the value of k, as compared to the number of sample elements, should be small since in the formula (7.12) of the expected value the term of correction, $\frac{\varrho}{2(n-1)}$, has been neglected.)

7.1.3. THE MEDIAL CORRELATION

Pairs of random variables occurring in the practice of hydrology show in many cases monotonic tendencies. Rising stages involve the increase of flow values, a heavy rainfall in the catchment results in rising stages in the channel, etc. Departures from these tendencies are induced mostly by the behavior of additional variables.

Random variables X and Y are said to be positive quadrant dependent if the inequality

$$H(x, y) = P(X < x, Y < y) \geqq P(X < x) P(Y < y) = \\ = F(x) G(y) \tag{7.2}$$

is satisfied for any quadrants $\{X < x, Y < y\}$. The concept of positive or negative quadrant dependence (see Eq. 7.2) as introduced by Lehmann [B. 20] reflects the positive or negative association between the variables. Numerous so-called nonparametric measures have been introduced to quantify the monotonic association. In this section the so-called medial correlation is presented; it was introduced by Mosteller [B. 24] while its statistical analysis was performed by Blomquist [B. 6].

Let the joint distribution function of random variables X and Y be denoted by $H(x, y)$. Let $F(x)$ be the distribution function of X and $G(y)$ the same for Y. Furthermore, denote by $\tilde{x}_{1/2}$ the median of distribution of variable X and by $\tilde{y}_{1/2}$ the same for Y that is let

$$F(\tilde{x}_{1/2}) = G(\tilde{y}_{1/2}) = \frac{1}{2}.$$

Using the medians divide the sample space into four quadrants in the following way

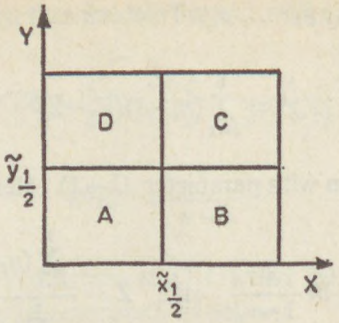


Figure 66

(see Figure 66.) (For sake of simplicity the X and Y values are considered positive.) Now calculate the probability

$$\tilde{\varrho}_{1/2} = P(A+C)-P(B+D) = P(A+C)-[1-P(A+C)] = 2P(A+C)-1, \tag{7.13}$$

the estimator of which is the Mosteller—Blomquist statistic. Due to the property of the median points

$$P(A)+P(B) = 1/2$$

$$P(C)+P(B) = 1/2$$

that is $P(A)=P(C)$ and because $P(A+C)=P(A)+P(C)$, one can obtain that

$$\tilde{\varrho}_{1/2} = P(A+C)-P(B+D) = 4P(A)-1. \tag{7.14}$$

As $P(A)=H(\tilde{x}_{1/2}, \tilde{y}_{1/2})$ so

$$\tilde{\varrho}_{1/2} = 4H(\tilde{x}_{1/2}, \tilde{y}_{1/2})-1. \tag{7.15}$$

Measure $\tilde{\varrho}_{1/2}$ is called medial correlation, it is applicable very well to get a quick and simple information on the relation between two random variables. When the joint distribution function, $H(x, y)$, is not known then $\tilde{\varrho}_{1/2}$ is to be estimated from the sample. Let $\tilde{x}_{1/2}$ and $\tilde{y}_{1/2}$ be the sample medians which determine the event $\hat{A}=\{X<\tilde{x}_{1/2}, Y<\tilde{y}_{1/2}\}$. Using the k/n relative frequency of event $\hat{A}$, value of $\tilde{\varrho}_{1/2}$ is estimated by

$$\hat{\varrho}_{1/2} = 4\frac{k}{n}-1. \tag{7.16}$$

This estimation requires rather little calculation as what is to be made is just to count the points in set A and to calculate relative frequency k/n.

Since relative frequency k/n is an asymptotically unbiased estimate of probability $P(A)=H(\tilde{x}_{1/2}, \tilde{y}_{1/2})$ therefore

$$E\left(\frac{k}{n}\right) = H(\tilde{x}_{1/2}, \tilde{y}_{1/2})$$

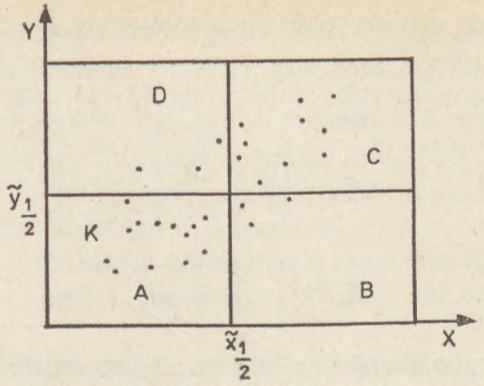


Figure 67

so that

$$E(\hat{\varrho}_{1/2}) = E\left(4\frac{k}{n}-1\right) = 4E\left(\frac{k}{n}\right)-1 \approx$$

$$\approx 4H(\tilde{x}_{1/2}, \tilde{y}_{1/2})-1 = \tilde{\varrho}_{1/2}.$$

Furthermore,

$$D\left(\frac{k}{n}\right) \approx \sqrt{\frac{H(\tilde{x}_{1/2}, \tilde{y}_{1/2})[1-H(\tilde{x}_{1/2}, \tilde{y}_{1/2})]}{n}}.$$

If X and Y are independent then $H(\tilde{x}_{1/2}, \tilde{y}_{1/2}) = F(\tilde{x}_{1/2})G(\tilde{y}_{1/2}) = 1/4$ whence $E(\varrho_{1/2}) = 0$. Since in this case

$$D\left(\frac{k}{n}\right) = \sqrt{\frac{\frac{1}{4}\cdot\frac{3}{4}}{n}} = \frac{\sqrt{3}}{4\sqrt{n}}$$

if follows that $D^2(\hat{\varrho}_{1/2}) = 4^2 \cdot \frac{3}{4^2 n}$ that is

$$D(\hat{\varrho}_{1/2}) = \sqrt{\frac{3}{n}}.$$

In case of independent random variables value of $\hat{\varrho}_{1/2}$ satisfies the approximate relation

$$P\left(-2\sqrt{\frac{3}{n}} \leqq \hat{\varrho}_{1/2} < 2\sqrt{\frac{3}{n}}\right) \approx 0.95.$$

Measure $\tilde{\varrho}_{1/2}$ has several advantageous properties. When X and Y are independent then $\tilde{\varrho}_{1/2} = 0$. Furthermore, when there is a monotonic functional relation between X and Y, that is if $Y = \varphi(X)$, then $\tilde{\varrho}_{1/2} = 1$ if $\varphi(.)$ is monotonic increasing and $\tilde{\varrho}_{1/2} = -1$ if $\varphi(.)$ is monotonic decreasing. (The reverse of these statements is, in general, not true.) Relation

(7.17) $$-1 \leqq \tilde{\varrho}_{1/2} < 1$$

is always satisfied since, as it can be seen in the following, $\tilde{\varrho}_{1/2}$ is a special correlation coefficient:

Let ξ_x and η_y indicator variables of the events $\{X<x\}$ and $\{Y<y\}$, respectively, that is

$$\xi_x = \begin{cases} 1 & \text{if } X < x, \\ 0 & \text{if } X \geqq x \end{cases}$$

$$\eta_y = \begin{cases} 1 & \text{if } Y < y \\ 0 & \text{if } Y \geqq y. \end{cases}$$

Calculate the coefficient of correlation between random variables ξ_x and η_y. Applying formula (7.1) it can be written that

$$\tilde{\varrho}(\xi_x, \eta_y) = \frac{E(\xi_x, \eta_y) - E(\xi_x)E(\eta_y)}{D(\xi_x)D(\eta_y)} = \tag{7.18}$$

$$= \frac{H(x, y) - F(x)G(y)}{\sqrt{F(x)[1-F(x)]G(y)[1-G(y)]}}.$$

If $x=\tilde{x}_{1/2}$ and $y=\tilde{y}_{1/2}$ then

$$\varrho(\xi\tilde{x}_{1/2}, \eta\tilde{y}_{1/2}) = \frac{H(\tilde{x}_{1/2}, \tilde{y}_{1/2}) - F(\tilde{x}_{1/2})G(\tilde{y}_{1/2})}{\sqrt{F(\tilde{x}_{1/2})[1-F(\tilde{x}_{1/2})]G(\tilde{y}_{1/2})[1-G(\tilde{y}_{1/2})]}} = \tag{7.19}$$

$$= \frac{H(\tilde{x}_{1/2}, \tilde{y}_{1/2}) - \frac{1}{4}}{\frac{1}{4}} = 4H(\tilde{x}_{1/2}, \tilde{y}_{1/2}) - 1 = \tilde{\varrho}_{1/2}.$$

Quantity $\tilde{\varrho}(\xi_x, \eta_y)$ in formula (7.18) is called indicator correlation (see [B. 30]). So the medial correlation, $\tilde{\varrho}_{1/2}$, is a special case of indicator correlation.

From Eq. (7.18) it can be deducted that if X and Y are independent that is when $H(x, y)=F(x)G(y)$ then $\tilde{\varrho}(\xi_x, \eta_y)=0$, consequently, following from (7.19), $\tilde{\varrho}_{1/2}=0$. It can also be seen that, in case of a positive quadrant dependence: $H(x, y)\geqq$ $\geqq F(x)G(y)$, $\tilde{\varrho}\geqq 0$.

To realize that a monotonic (increasing or decreasing) functional relation, $Y=$ $=\varphi(X)$, between random variables X and Y involves that $|\tilde{\varrho}|=1$ now such an important notion is introduced which will be relied on several times later on.

Definition. In case of $H(x, y)>F(x)G(y)$ the set of coupled points $(\tilde{x}_\alpha, \tilde{y}_\alpha)$ while in case of $H(x, y)<F(x)G(y)$ the set of coupled points $(\tilde{x}_\alpha, \tilde{y}_{1-\alpha})$ is called quantile curve.

In the definition $\tilde{x}_\alpha$ and $\tilde{y}_\alpha$ are the α-quantiles which satisfy equation $F(\tilde{x}_\alpha)=$ $=G(\tilde{y}_\alpha)=\alpha$ while $G(\tilde{y}_{1-\alpha})=1-\alpha$, $\alpha\in[0, 1]$.

In the following it is supposed that both $F(x)$ and $G(y)$, the marginal distribution functions, are strictly monotonic increasing functions.

If between the random variables X and Y there is a monotonic increasing or decreasing functional relationship, $Y=\varphi(X)$, then holds the following

Lemma 1. Let there between the random variables X and Y be a monotonically increasing (or decreasing) functional relation, $Y=\varphi(X)$, then $\tilde{y}_\alpha=\varphi(\tilde{x}_\alpha)$ (or $\tilde{y}_{1-\alpha}=\varphi(\tilde{x}_{1-\alpha})$) where $\tilde{x}_\alpha$ and $\tilde{y}_\alpha$ (or $\tilde{x}_{1-\alpha}$ and $\tilde{y}_{1-\alpha}$) are the α- and $(1-\alpha)$- quantiles of the distributions of variables X and Y, respectively.

Due to Definition 1 this lemma asserts that in these cases the regression curve of the random variables X and Y, the function $y=\varphi(x)$, and the quantile curves are identical.

The lemma can be proved simply. Let $Y=\varphi(X)$ be monotonically increasing, then

$$\alpha = P(Y < \tilde{y}_\alpha) = P(\varphi(X) < \tilde{y}_\alpha) =$$

$$= P(X < \varphi^{-1}(\tilde{y}_\alpha)).$$

Since from the definition of α-quantile $P(X<\tilde{x}_\alpha)=\alpha$ it follows that $\varphi^{-1}(\tilde{y}_\alpha)=\tilde{x}_\alpha$ that is

$$\tilde{y}_\alpha = \varphi(\tilde{x}_\alpha) \quad \text{for all} \quad \alpha\in[0, 1]. \tag{7.20}$$

The decreasing case is similar.
If $\alpha=1/2$ then in both cases

$$\tilde{y}_{1/2} = \varphi(\tilde{x}_{1/2}) \tag{7.22}$$

which means that the quantile curve passes through the point defined by the medians.

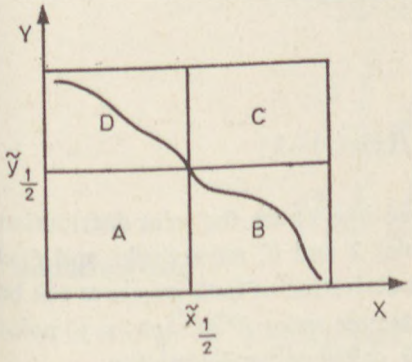


Figure 68

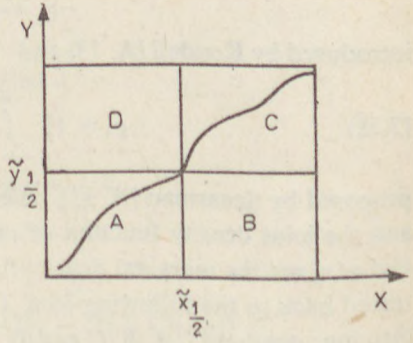


Figure 68/a

On this basis as indicated by the figure, for a monotonic increasing functional relation

$$\tilde{\varrho}_{1/2} = 4P(A)-1 = 4\cdot\frac{1}{2}-1 = 1 \quad (\text{as } P(A) = P(C)) \tag{7.23}$$

while for a monotonic decreasing functional relation

$$\tilde{\varrho}_{1/2} = 4P(A) - 1 = 4 \cdot 0 - 1 = -1. \tag{7.24}$$

If the joint distribution of the random variables X and Y is a two-dimensional normal then the following relation holds between the medial correlation, $\tilde{\varrho}_{1/2}$ and the correlation coefficient, ϱ:

$$\varrho = \sin \frac{\pi}{2} \tilde{\varrho}_{1/2}. \tag{7.25}$$

which can be obtained by direct calculation.

Note that the value of indicator correlation $\tilde{\varrho}$ in Eq. (7.18) can be calculated easily not only for the median point $(\tilde{x}_{1/2}, \tilde{y}_{1/2})$ but also for any point $(\tilde{x}_\alpha, \tilde{y}_\alpha)$ of the quantile curve. Denoting by $\tilde{x}_\alpha$ and $\tilde{y}_\alpha$ the α-quantiles of the marginal distributions then according to Eq. (7.18) relation

$$\tilde{\varrho}_\alpha = \frac{H(\tilde{x}_\alpha, \tilde{y}_\alpha) - \alpha^2}{\alpha - \alpha^2} \tag{7.26}$$

will be obtained where $\alpha \in [0, 1]$. Formula (7.26) incorporates formula (7.19) as a special case.

7.1.4. KENDALL'S τ_1 AND SPEARMAN'S ϱ_s

Among the measures representing the closeness of connection between random variables are commonly used the formulae

$$\tau = 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x, y) h(x, y)\, dx\, dy - 1 \tag{7.27}$$

introduced by Kendall [A. 13] and

$$\varrho_s = 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x, y) f(x) g(y) - 3 \tag{7.28}$$

proposed by Spearman [B. 35], where $H(x, y)$ and $h(y, x)$ are the joint distribution and the joint density function of random variables X and Y, respectively, and $f(x)$ and $g(y)$ are the marginal density functions. The derivation of both measures can be traced back to the following idea. Dividing the sample space R^2 at each (x, y) point into four quadrants, A, B, C and D, see Figure 69, calculate the probability

$$\begin{aligned} \tilde{\varrho}(x, y) &= P(A+C) - P(B+D) = P(A+C) - \\ &\quad -[1 - P(A+C)] = 2P(A+C) - 1 \end{aligned} \tag{7.29}$$

which is, obviously, a function of variables x and y. The quantity $\tilde{\varrho}(x, y)$ gives essentially the measure by which, at point (x, y), the tendency of monotonic increase between the random variables X and Y is stronger than the tendency of monotonic

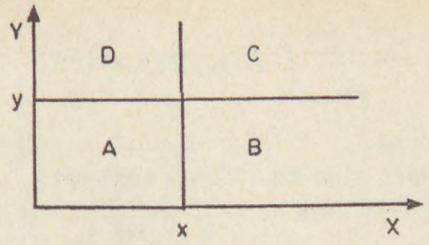


Figure 69

decrease. Averaging the quantity $\tilde{\varrho}(x, y)$ and using the joint density function $h(x, y)$ we obtain the expression

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mu(x, y)h(x, y)\,dx\,dy. \tag{7.30}$$

Now we turn to prove that this is equal to τ. Since for point (x, y) relations

$$P(A) = H(x, y) \quad \text{and} \quad P(C) = 1-F(x)-G(y)+H(x, y)$$

hold and with these

$$P(A+C) = P(A)+P(C) = 2H(x, y)-F(x)-G(y)+1$$

and

$$\tilde{\varrho}(x, y) = 2P(A+C)-1 = 4H(x, y)-2F(x)-2G(y)+1$$

it follows that

$$\tau = 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x, y)h(x, y)dx\,dy-$$

$$-2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x)h(x, y)dx\,dy-$$

$$-2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(y)h(x, y)dx\,dy-1.$$

Considering that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x)h(x, y)dx\,dy = \int_{-\infty}^{\infty} F(x)\left[\int_{-\infty}^{\infty} h(x, y)dy\right]dx =$$

$$= \int_{-\infty}^{\infty} F(x)f(x)dx = \frac{F^2(x)}{2}\Bigg|_{-\infty}^{\infty} = \frac{1}{2}$$

and that, similarly,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x)h(x, y)dx\,dy = \frac{1}{2}$$

what is obtained is

$$\tau = 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x, y)h(x, y)dx\,dy - 1, \tag{7.31}$$

which is Kendall's measure.

However, if quantity $\tilde{\varrho}(x, y)$ in Eq. (7.29) is averaged by using the product of the densities an expression of the form

$$\begin{aligned} \frac{1}{3}\varrho_s &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{\varrho}(x, y)f(x)g(y)dx\,dy = \\ &= 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x, y)f(x)g(y)dx\,dy - \\ &-2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x)f(x)g(y)dx\,dy = \\ &= 2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(y)f(x)g(y)dx\,dy - 1 = \\ &= 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x, y)f(x)g(y)dx\,dy - 1 \end{aligned} \tag{7.32}$$

is obtained which equals to one third of Spearman's measure.

It can be seen that the values of both measures, τ contained in Eq. (7.27) and ϱ_s included in Eq. (7.28) are zero if X and Y are independent. On the other hand, the values of both measures are $+1$ if a monotonic increasing functional relation holds between X and Y and -1 if Y is a monotonic decreasing function of variable X. In addition, the value of both τ and ϱ_s is invariant under monotonic (increasing or decreasing) transformation of both variables.

For the determination of the closeness of a stochastic relation the use of formulae (7.27) and (7.28) can take place only when the joint distribution function $H(x, y)$ of the variables is known. Since in practice this is usually not the case the value of both τ and ϱ_s is estimated from a sample. Whereas the estimation of the medial correlation $\tilde{\varrho}_{1/2}$ can be obtained by an extremely simple calculation the estimation of τ and ϱ_s is rather tedious. Below the estimation of τ is presented while for the estimation of ϱ_s we refer to the literature (see [B. 16]).

Let a sample of size n be

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n). \tag{I}$$

An estimation of τ can be obtained for each point (X_i, Y_i) in the following manner. The points (X_i, Y_i) and (X_j, Y_j) will be called concordant if $(X_i - X_j)(Y_i - Y_j) > 0$. In the opposite case they are discordant. Now the estimator of τ can be obtained from

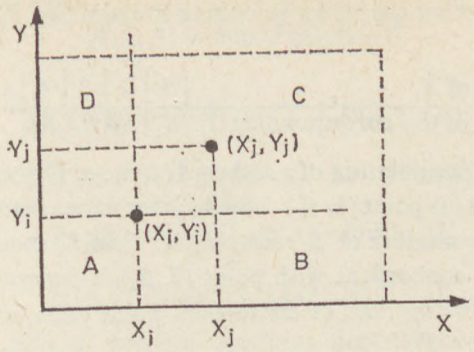


Figure 70

the following expression:

$$\frac{\tau+1}{2} = \frac{\text{number of concordant points}}{\binom{n}{2}} \tag{7.33}$$

The number of the concordant points can be counted in the following way (see Figure 71.).

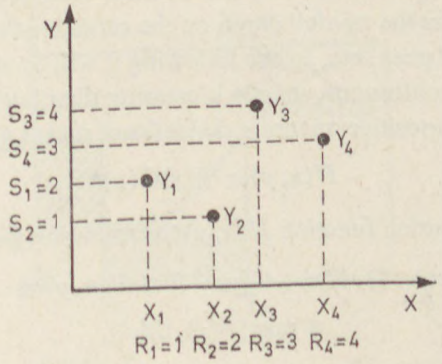


Figure 71

Accept the X co-ordinates of points in sample (I) in their natural order and attach to them ranks corresponding to their indices. So the rank of the lowest X_i will be 1, that of the next lowest one will be 2 and so on. By projecting the Y co-ordinates of (X_i, Y_i) points to the Y-axis an ordered sample will be obtained where for Y_i its rank, S_i, that is a number corresponding to its location in the order of magnitudes is substituted, as it is seen in Fig. 71. Point (X_i, Y_i) is substituted by a rank pair (R_i, S_i) in a manner that the point pairs conserve their concordance or discordance.

Write under one another the ranks belonging to pairs (X_i, Y_i) in a table as shown below:

Rank of X_i	1	2	3	4	...	n
Rank of the corresponding Y_i	S_1	S_2	S_3	S_4		S_n

Ranks S_i constitute a permutation of numbers $1, 2 \ldots, n$. Determine now the number of points concordant with point $(1, S_1)$. This number varies between 0 (if $S_1 = n$) and $n-1$ (if $S_1 = 1$). (The number of S_i ranks higher than S_1 should be counted.) Then the number of points concordant with point $(2, S_2)$ is determined (excluding point $(1, S_1)$), etc. Finally, the number of concordant point pairs obtained in this way is divided by $\binom{n}{2} = \frac{n(n-1)}{2}$, the number of all pairs, and what is obtained is the same from which the estimation of τ can be calculated.

We remark that the right hand side of (7.33) is the average of the estimations of probabilities $P(A+C)$ taken for each point (x_i, y_i).

The attention of readers interested in further details is drawn to work [B. 16].

7.1.5. EXAMINATION OF THE POSITIVE QUADRANT DEPENDENCE

In the practice of hydrology, so in the hydrology of floods as well, the case of positive quadrant dependence occurs frequently. This indicates a monotonic increasing tendency between the random variables X and Y. Such cases are, e.g., the development of a stage Y if X denotes the rainfall depth on the catchment or if Y is the flood duration and X is the flood peak, etc. In the following it will be supposed that both variables, X and Y, have continuous, strictly increasing distribution functions, $F(x)$ and $G(y)$, respectively. The positive quadrant dependence means, that

$$H(x, y) \geqq F(x)G(y)^*$$

For any joint distribution function $H(x, y)$ the following inequality holds:

$$\max [0, F(x)+G(y)-1] \leqq H(x, y) \leqq \leqq \min [F(x), G(y)], \tag{7.35}$$

which will be shown below. Quantities at the left and righ hand side of inequality (7.35), respectively, are called Frechét bounds.

To realize the validity of this inequality consider events $A = \{X < x\}$, $B = \{Y < y\}$ and $AB = \{X < x, Y < y\}$. Obviously

$$P(AB) \leqq P(A)$$

and

$$P(AB) \leqq P(B)$$

that is

$$P(AB) \leqq \min [P(A), P(B)]$$

* See: [B.14]

whence

$$H(x, y) = \min [F(x), G(y)].$$

Furthermore,

$$P(A+B) = P(A)+P(B)-P(AB) \leqq 1$$

that is

$$P(AB) \geqq P(A)+P(B)-1, \quad \text{as} \quad P(AB) \geqq 0$$

and

$$P(AB) \geqq \max [0, P(A)+P(B)-1]$$

so that

$$H(x, y) \geqq \max [0, F(x)+G(y)-1].$$

Concludingly, in case of positive quadrant dependence that is when relation $H(x, y) \geqq F(x)G(y)$ is satisfied inequality

$$F(x)G(y) \leqq H(x, y) \leqq \min [F(x), G(y)] \tag{7.36}$$

will hold.

Equality $H(x, y)=F(x)G(y)$ holds if and only if the random variables X and Y are independent.

Equality $H(x, y)=\min [F(x), G(y)]$ holds if and only if between random variables X and Y there is a monotonic increasing functional relation: $Y=\varphi(X)$. In this case, as it is apparent form the figure below (Figure 72), the whole mass of probabilities is located on the quantile curve, $Y=\varphi(x)=G^{-1}[F(x)]$:

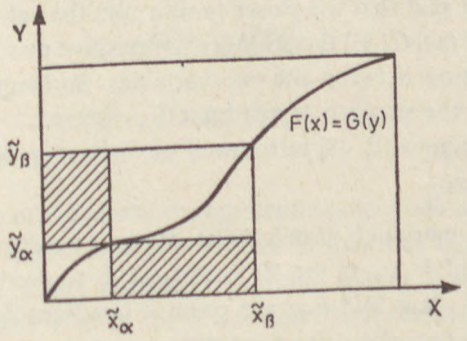


Figure 72

If $H(x, y)=\min [F(x), G(y)]$ then, with $\beta>\alpha$,

$$H(\tilde{x}_\alpha, \tilde{y}_\alpha) = F(\tilde{x}_\alpha) = \alpha$$

$$H(\tilde{x}_\beta, \tilde{y}_\alpha) = G(\tilde{y}_\alpha) = \alpha$$

that is

$$H(\tilde{x}_\beta, \tilde{y}_\alpha) - H(\tilde{x}_\alpha, \tilde{y}_\alpha) = 0$$

for all $x \in [0, 1]$ and $\beta>\alpha$, which implies that the probability measure of area above (and similarly below) the quantile curve is zero.

$$H(\tilde{x}_\alpha, \tilde{y}_\beta) = F(\tilde{x}_\alpha) = \alpha.$$

This implies that the value of the density function is zero except at the points of the quantile curve.

The marginal distributions of the two-dimensional distribution function

$$H(x, y) = \min [F(x), G(y)] \text{ are}$$

$$H(x, +\infty) = \min [F(x), G(+\infty)] = F(x),$$

$$H(+\infty, y) = \min [F(+\infty), G(y)] = G(y).$$

By virtue of relationship (7.36) $\min [F(x), G(y)]$ is the uniformly greatest bivariate distribution surface whose marginal distributions are $F(x)$ and $G(y)$, respectively. On the other hand, if $H(x, y) = \min [F(x), G(y)]$ then, in accordance with the foregoing, there is a monotonic increasing functional relation between the random variables X and Y, in the form $Y = \varphi(X)$. This represents, obviously, the closest relation between them that is this is the case where the positive quadrant dependence is the strongest.

When X and Y are independent the marginal distributions of the bivariate distribution function $H(x, y) = F(x)G(y)$ are also $F(x)$ and $G(y)$.

The positive quadrant dependence is the lowest when X and Y are independent.

In respect of all such X and Y random variables whose joint distribution function, $H(x, y)$ satisfies inequality (7.36) and whose marginal distributions are $F(x)$ and $G(y)$, respectively, it may be said that the closer (uniformly) the value of $H(x, y)$ to distribution function $\min [F(x), G(y)]$ the stronger the positive quadrant dependence that is the closest the relation between the two variables. So inequality (7.36) defines a certain graduation for the strength of quadrant dependence.

In his paper Yanamigoto [B. 48] introduced the following definition for the grade of quadrant dependence:

Definition 2. If the marginal distributions of the two-dimensional distribution functions $H_1(x, y)$ and $H_2(x, y)$ are $F(x)$ and $G(y)$, respectively, and $H_1(x, y) \leqq H_2(x, y) \leqq F(x)G(y)$, it is said that the positive quadrant dependence of the random variables is stronger when the joint distribution is $H_1(x, y)$ than when it is $H_2(x, y)$.

In general the determination of the one-dimensional distributions of the random variables X and Y is easier (e.g., through test of fit) but it is much more difficult for their joint distribution function, $H(x, y)$. Two such measures will be presented.

Starting from inequality (7.36), if a positive quadrant dependence is the case, the inequality

$$0 \leqq H(x, y) - F(x)G(y) \leqq \min [F(x), G(y)] - F(x)G(y) \tag{7.37}$$

is always satisfied. Consequently, for all (x, y) points this may be written in the form

$$H(x, y) - F(x)G(y) = \lambda(x, y)\{\min [F(x), G(y)] - [F(x)G(y)]\} \tag{7.38}$$

where λ is a continuous function with bounds $0 \leqq \lambda(x, y) \leqq 1$. Let the function

$$\lambda(x, y) = \frac{H(x, y) - F(x)G(y)}{\min [F(x), G(y)] - F(x)G(y)} \tag{7.39}$$

the function of connection between X and Y with distribution $F(x)$ and $G(y)$, respectively. The function $\lambda(x, y)$ at a given point (x, y) is the ratio of the difference between the values $H(x, y)$ and $F(x)G(y)$ and of the maximum difference possible at this point. Note that at the median point $(\tilde{x}_{1/2}, \tilde{y}_{1/2})$ the value of function $\lambda(x, y)$ is just the same as that of the medial correlation $\tilde{\varrho}_{1/2}$ since

$$\lambda(\tilde{x}_{1/2}, \tilde{y}_{1/2}) = \frac{H(\tilde{x}_{1/2}, \tilde{y}_{1/2}) - \frac{1}{4}}{\frac{1}{2} - \frac{1}{4}} =$$

$$= 4H(\tilde{x}_{1/2}, \tilde{y}_{1/2}) - 1 = \tilde{\varrho}_{1/2}$$

The following average value of $\lambda(x, y)$:

$$\lambda^* = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{H(x, y) - F(x)G(y)}{\min [F(x), G(y)] - F(x)G(y)} \cdot \tag{7.40}$$

$$\cdot f(x)g(y)dx\,dy =$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \lambda(x, y) f(x)g(y)dx\,dy$$

will be considered one of the measures representing the grade of the positive quadrant dependence. Considering formula (7.37) it can be seen that $0 \leqq \lambda^* \leqq 1$ and that $\lambda^* = 0$ if and only if X and Y are independent while $\lambda^* = 1$ if and only if $H(x, y) = \min [F(x), G(y)]$ that is when there is a functional relation between the random variables X and Y, which is, of course, monotonically increasing.

Let now $H_1(x, y) \leqq H_2(x, y) \leqq F(x)G(y)$ and

$$\lambda_1^* = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{H_1(x, y) - F(x)G(y)}{\min [F(x), G(y)] - F(x)G(y)} f(x)g(y)dx\,dy,$$

$$\lambda_2^* = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{H_2(x, y) - F(x)G(y)}{\min [F(x), G(y)] - F(x)G(y)} f(x)g(y)dx\,dy.$$

Hence

$$\lambda_1^* - \lambda_2^* = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{H_1(x, y) - H_2(x, y)}{\min [F(x), G(y)] - F(x)G(y)} f(x)g(y)dx\,dy \geqq 0 \tag{7.41}$$

that is, using λ^* as a measure, if $H_1 > H_2$, H_1 is more quadrant-dependent than H_2. Another measure can be given by the following ratio:

$$\lambda^{**} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [H(x,y) - F(x)G(y)]dx\,dy}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{\min[F(x), G(x)] - F(x)G(y)\}dx\,dy}. \tag{7.42}$$

It is easy to see that, in case of positive quadrant dependence, $0 \leqq \lambda^{**} \leqq 1$, furthermore that $\lambda^{**} = 0$, if and only if X and Y are independent while $\lambda^{**} = 1$ if and only if there is a monotonic increasing functional relation between X and Y. Furthermore, if $H_1 \geqq H_2 \geqq F(x)G(y)$ then, using λ^{**} as a measure, H_1 has a larger quadrant-dependence than H_2.

Remark 1: Measure λ^{**} is the quotient of two volumes, viz. the ratio of the volume between distribution surfaces $H(x, y)$ and $F(x)G(y)$ and the volume between surfaces $\min[F(x), G(y)]$ and $F(x)G(y)$.

Remark 2: Measure λ^{**} can be called covariance quotient. It is known (see, e.g., Lehmann [B. 20]) that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [H(x,y) - F(x)G(y)]dx\,dy = \tag{7.43}$$

$$= E(XY) - E(X)E(Y) = \operatorname{cov}(X, Y).$$

On this basis

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{\min[F(x)G(y)] - F(x)G(y)\}dx\,dy = \tag{7.44}$$

$$\operatorname{cov}_+(X, Y)$$

where $\operatorname{cov}_+(X, Y)$ is the covariance between random variables X and Y if their joint distribution function is $\min[F(x), G(y)]$. So

$$\lambda^{**} = \frac{\operatorname{cov}(X,Y)}{\operatorname{cov}_+(X,Y)} = \frac{\dfrac{\operatorname{cov}(X,Y)}{D(X)D(Y)}}{\dfrac{\operatorname{cov}_+(X,Y)}{D(X)D(Y)}} = \frac{\varrho}{\varrho_+} \tag{7.45}$$

which means that λ^{**} is at the same time a correlation quotient.

Remark 3: In the knowledge of marginal distributions $F(x)$ or $G(y)$, respectively, measure λ^{**} can be estimated very simply from a statistical sample. First the correlation coefficient is estimated by using the usual method (see formula 7.6) and then, in the knowledge of the marginal distributions, it is already possible to calculate the

maximum correlation coefficient ϱ_+ since, after calculating function $\varphi(x)=G^{-1}[F(x)]$,

$$\operatorname{cov}_+(X,Y)=E[X,\varphi(X)]-E(X)E(Y)=$$

$$=\int_{-\infty}^{\infty} x\varphi(x)f(x)dx-\int_{-\infty}^{\infty} xf(x)dx\int_{-\infty}^{\infty} yg(y)dy$$

and what is still to do is to divide by the product of standard deviations which can also be calculated as the marginal distributions are known.

Remark 4: If the quantile curve relating to random variables X and Y is a straight line that is when $y=G^{-1}[F(x)]=ax+b$ then $\varrho_+=1$ so that in this case $\lambda^{**}=\varrho$. In this way, on the basis of Remark 1, for distributions of this type a new geometric interpretation has been obtained to the correlation coefficient.

If the quantile curve $y=G^{-1}[F(x)]$ is not a linear function then $\varrho_+<1$, consequently, in this case $\lambda^{**}>\varrho$. Now λ^{**} is a better measure for the closeness of relation between variables X and Y since through ϱ rather the linearity of their relation is measured.

Special cases:

a) Let $H(x,y)$ be a two-dimensional normal distribution with parameters $E(X)=m_1$, $D(X)=\sigma_1$, $E(Y)=m_2$ and $D(Y)=\sigma_2$ and with correlation coefficient ϱ. With these the marginal distributions are

$$F(x)=\frac{1}{\sigma_1\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(t-m_1)^2}{2\sigma_1^2}}dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-m_1}{\sigma_1}} e^{-\frac{u^2}{2}}du=\Phi\left(\frac{x-m_1}{\sigma_1}\right),$$

$$G(y)=\frac{1}{\sigma_2\sqrt{2\pi}}\int_{-\infty}^{y} e^{-\frac{(t-m_2)^2}{2\sigma_2^2}}dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{y-m_2}{\sigma_2}} e^{-\frac{v^2}{2}}dv=\Phi\left(\frac{y-m_2}{\sigma_2}\right).$$

The equation of the quantile curve is $G(y)=F(x)$ that is $\tilde{y}=G^{-1}[F(x)]$. In our example

$$\Phi\left(\frac{y-m_2}{\sigma_2}\right)=\Phi\left(\frac{x-m_1}{\sigma_1}\right)$$

whence

$$\tilde{y}(x)=\frac{\sigma_2}{\sigma_1}x+m_2-\frac{\sigma_2}{\sigma_1}m_1 \tag{7.46}$$

which is the equation of a straight line so that by virtue of Remark 4 $\lambda^{**}=\varrho$.

b) Let

$$H(x,y)=\lambda\min[F(x),G(y)]+(1-\lambda)F(x)G(y) \tag{7.47}$$

where $0\leqq\lambda\leqq 1$, $\lambda=\text{const}$.

Hence

$$\lambda = \frac{H(x, y) - F(x)G(y)}{\min[F(x), G(y)] - F(x)G(y)} = \tag{7.48}$$

$$= \lambda(x, y) = \text{const.}$$

Consequently,

$$\lambda^* = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \lambda(x, y) h(x, y) dx\, dy = \tag{7.49}$$

$$= \lambda \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) dx\, dy = \lambda.$$

Furthermore,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [H(x, y) - F(x)G(y)]\, dx\, dy =$$

$$= \lambda \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{\min[F(x), G(y)] - F(x)G(y)\}\, dx\, dy$$

that is

$$\lambda = \lambda^{**} = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [H(x, y) - F(x)G(y)]\, dx\, dy}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{\min[F(x), G(y)] - F(x)G(y)\}\, dx\, dy} \tag{7.50}$$

So for a distribution as given by Eq. (7.47) $\lambda = \lambda^* = \lambda^{**} = \frac{\varrho}{\varrho_+}$.

7.1.6. TESTING DEPENDENCE BY MEANS OF QUANTILE VALUES

In general, the purpose of measuring stochastic relation between random variables is to decide to what extent these variables may be considered independent or, in other terms, how far their stochastic relation may be regarded close, to what extent they may be associated with a monotonic increasing or a monotonic decreasing tendency.

Conduct now an examination to the question if what a quick procedure can be developed to have a preliminary information on a stochastic relation, on the basis of numerical values obtained for the indicator correlation $\tilde{\varrho}_x$.

If only the k_1, k_2, k_3 frequencies of sample points located in the quadrants defined by points $(\tilde{x}_{1/4}, \tilde{y}_{1/4})$, $(\tilde{x}_{1/2}, \tilde{y}_{1/2})$, $(\tilde{x}_{3/4}, \tilde{y}_{3/4})$ are figured out (see Figure 73), their ratio itself is already a rather informative indication on the stochastic relation.

In case of independence it holds that

$$H(\tilde{x}_{1/4}, \tilde{y}_{1/4}) = \frac{1}{16}\,;\; H(\tilde{x}_{1/2}, \tilde{y}_{1/2}) = \frac{1}{4} = \frac{4}{16};\; H(\tilde{x}_{3/4}, \tilde{y}_{3/4}) = \frac{9}{16}$$

that is their ratio is

$$k_1 : k_2 : k_3 = 1 : 4 : 9 \tag{7.80}$$

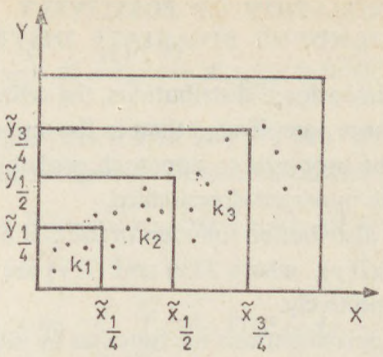


Figure 73

while for a functional relation it is

$$k_1 : k_2 : k_3 = 1 : 2 : 3. \tag{7.81}$$

So if the ratio of empirical values is close to either of the two series above a corresponding decision may be taken. When the result is not sharp the division can be utilized in a χ^2-test. When a large sample is treated the number of quantiles applied may be greater.

In the course of statistical analyses such a test whether the random variables $X_1, X_2, \ldots, X_n$ are independent or not is needed frequently. In an event where a (pairwise) dependence may be suspected this question can be answered through an approximation where the indicator correlation is utilized in the following manner:

Produce the pairs (X_1, X_2), (X_2, X_3), $\ldots$, (X_{n-2}, X_{n-1}), (X_{n-1}, X_n), plot them as points on the plane and apply the procedure above. (As a solution to this problem the Wald—Wolfowitz test can be found in literature, see [B. 42], which is based on serial correlation and whose execution is extremely tedious.) E.g., when what should be known is whether the $X_1, X_2, \ldots, X_{30}$ exceedances observed in the Tisza river at Szeged in the sequence of second quarters (see Table T.1) may be considered a statistical sample which represents an X random variable whose distribution is a given $F(x)$; that is, whether there is a dependence between the subsequent data or not, then, as described in the foregoing, a test of independence may be carried out before testing the fit.

The location pattern of points (X_1, X_2), (X_2, X_3), $\ldots$, (X_{29}, X_{30}) is shown in Fig. 72. In this case the ratio of the number of points located in the quadrants $(\tilde{x}_{1/4}, \tilde{y}_{1/4})$, $(\tilde{x}_{1/2}, \tilde{y}_{1/2})$, $(\tilde{x}_{3/4}, \tilde{y}_{3/4})$ is

$$k_1 : k_2 : k_3 = 2 : 9 : 16 = 1 : 4.5 : 8$$

which, apparently, differs only slightly from the ratio 1:4:9 that is it reflects conditions characteristic to independence. (Considering that the sample tested is a relatively small one the ratio obtained seems to be convincing to accept independence.)

7.1.7. SIMPLE APPROXIMATION OF POSITIVELY QUADRANT-DEPENDENT BIVARIATE DISTRIBUTIONS

When dealing with two-dimensional distributions, the calculation of probabilities for different events is much more complicated than in the one-dimensional case.

A simple method will be proposed to approach probabilities for two-dimensional events in case of positively quadrant-dependence.

Let $H(x, y)$ be the joint distribution function of the pair of random variables (X, Y), for which $H(x, y) \geqq F(x)G(y)$, where $F(x)$ and $G(y)$ are the marginal distribution functions of X and Y respectively.

Let us investigate the approximations the function by means of the values of

$$H_\lambda(x, y) = \lambda \min(F, G) + (1-\lambda)FG \tag{7.51}$$

in the sense of quadratic mean deviation. Choosing λ such, that

$$\varphi(\lambda) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (H-H)^2 f g \, dx \, dy = \min \tag{7.52}$$

As

$$H_\lambda - H = (H_\lambda - FG) - (H - FG) =$$
$$= \lambda[\min(F, G) - FG] - (H - FG)$$

we obtain

$$\varphi(\lambda) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (H_\lambda - H)^2 f g \, dx \, dy = \lambda^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\min(F, G) - FG)^2 f g \, dx \, dy - \tag{7.53}$$

$$-2\lambda \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [\min(F, G) - FG][H - FG] f g \, dx \, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (H - FG)^2 f g \, dx \, dy \geqq 0.$$

Introducing the quantities:

$$\mu = 90 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (H - FG)^2 f g \, dx \, dy \text{ (Hoeffding [B. 11]) and} \tag{7.54}$$

$$\nu = 90 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [\min(F, G) - FG][H - FG] f g \, dx \, dy \tag{7.55}$$

we have

$$\varphi(\lambda) = \frac{\lambda^2}{90} - 2\frac{\lambda \nu}{90} + \frac{\mu}{90}. \tag{7.56}$$

(Obviously $\nu \geqq \mu$ and a simple calculation shows, that $\mu = \nu = 1$ if $H = \min(F, G)$.)

The value of $\varphi(\lambda)$ will be minimum if

$$\varphi'(\lambda) = \frac{2\lambda - 2\nu}{90} = 0 \quad \text{i.e.} \quad \lambda = \nu \tag{7.57}$$

Then for (7.56) we get:

$$\varphi(\nu) = \frac{\mu - \nu^2}{90} \geqq 0, \quad \text{i.e.} \quad \nu \geqq \mu \geqq \nu^2 \tag{7.58}$$

Hence:

$$\varphi(v)=\frac{\mu-v^2}{90}\leqq\frac{v(1-v)}{90}\leqq\frac{1}{360}=0.0027<3.10^{-3}. \tag{7.59}$$

The mean-square deviation between H and H_λ can of course be much less. For example let us consider the Morgenstern-distribution (see: [B. 23]).

$$H=FG+\alpha F(1-F)G(1-G)\quad(0\leqq\alpha\leqq 1) \tag{7.60}$$

For this distribution:

$$\mu=90\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(H-FG)^2fg\,dx\,dy=$$

$$=90\alpha^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(F-F^2)^2(G-G^2)^2fg\,dx\,dy=\frac{\alpha^2}{10}$$

$$v=90\alpha\iint\limits_{F\leqq G}F^2(1-F)G(1-G)^2fg\,dx\,dy+$$

$$+90\alpha\iint\limits_{F>G}F(1-F)^2G^2(1-G)fg\,dx\,dy\approx\frac{3\alpha}{10}.$$

In this case

$$\varphi(v)=\frac{\mu-v^2}{90}=\frac{\alpha^2}{9000}=0.0001\alpha<\frac{1}{9}10^{-3}.$$

The result is similar for the distribution

$$H=\min(F,G)-\alpha\min(F,G)(1-F)(1-G)\quad(0\leqq\alpha\leqq 1) \tag{7.61}$$

as well.

In case of $F(x)=1-e^{-\beta x}$, $G(y)=1-e^{-\gamma y}$, this bivariate distribution is the joint distribution function of the exceedance X and the duration Y of a flood-peak above of a sufficiently high level c for the River Tisza (See: [B. 31]).

CHAPTER 8

Regression analysis

8.1. METHODS TO CALCULATE REGRESSION

8.1.1. THE LEAST SQUARES METHOD. THE REGRESSION CURVE

The least squares method is a very important aid of probability theory and mathematical statistics. The method is used to determine and interpret theoretical quantities in probability theory and to estimate functions and constants from measurements in mathematical statistics. It is the analysis of results through mathematical statistics that may provide a support to justify the application of this method; however, cases where this is not the most suitable method will also be seen while in other cases the efficiency of this method will be pointed out. Below the method will be presented first as a means of determining the so-called regression curves for two random variables, followed by a brief discussion on the case of multivariate functional relations.

Consider the pair of random variables X and Y. On the basis of theoretical considerations or empirical data it is known about them that although neither of them is determined accurately by the other, nevertheless, the possible trends in the values of one of them has an influence in a specified way on the values of the other. The value of Y (which can be measured in a more difficult way or only later) is attempted to be approached by a certain function $g(x)$ of variable X. Examine now the question that what a function $g(x)$ of X will approach Y "the best" in the sense that

$$E\{[Y-g(x)]^2\} = \text{minimum}. \tag{8.1}$$

By the aid of the joint density function of the pair of random variables (X, Y) this condition may be written in the form

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[y-g(x)]^2 h(x, y)\,dx\,dy = \text{minimum}. \tag{8.2}$$

This condition is satisfied by

$$g(x) = E(Y|X = x) \tag{8.3}$$

which is a function representing the conditional expected value of variable Y under the condition that $X=x$. This can be proved as follows.

As the expected value of the conditional expected value of a random variable is equal to the unconditional expected value (see Section 2.1.7/b) it follows that

$$E\{[Y-g(x)]^2\} = E\{E[Y-g(x)]^2|X = x\}. \tag{8.4}$$

Furthermore, according to Steiner's theorem (see formula 2.4.2)

$$E\{[Y-g(x)]^2|X=x\} = \text{minimum} \quad \text{if} \quad g(x) = E(Y|X=x) \tag{8.5}$$

for all fixed values $X=x$. This involves that

$$E\{[Y-g(x)]^2\} = \text{minimum} \quad \text{if} \quad g(x) = E(Y|X=x). \tag{8.6}$$

Function $E(Y|X=x)=\bar{y}(x)$ is called the regression curve of Y related to X. Under all conditions where $X=x$ the Y values fluctuate around the expected value $E(Y|X=x)=\bar{y}(x)$ and so if X value is measured and Y value is calculated by using formula $\bar{y}(x)=+E(Y|X=x)$ then what is expected is that through many measurements the errors (deviations) will be balanced. Using the joint density function the formula of a regression curve can be written as follows:

$$\bar{y}(x) = \frac{\int_{-\infty}^{\infty} yh(x, y)\,dy}{\int_{-\infty}^{\infty} h(x, y)\,dy}. \tag{8.7}$$

Unfortunately, in most cases the joint density function of the two random variables is not known and therefore in the derivation of the formula depicting the regression curve difficulties are encountered.

The regression curve of X related to Y, function $E(X|Y=y)=\bar{x}(y)$, may be interpreted in a completely analogous way. Note here that if X and Y are independent then $\bar{y}(x)=E(Y|X=x)=E(Y)=c_1$ (const.) and $\bar{x}(y)=c_2$ (const.).

When the joint density function $h(x, y)$ is not known or when the density function is too sophisticated then, by the aid of the sample (cluster of points on a plane): (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), which represents the pair of random variables (X, Y), the regression curve is substituted possibly by a simple curve, a straight line or a polynom.

8.1.2. REGRESSION IN CASE OF BIVARIATE NORMAL DISTRIBUTION

In cases where the joint distribution of random variables X and Y is two-dimensional normal that is when their joint density function is

$$h(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} e^{-\frac{1}{2(1-\varrho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\varrho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]} \tag{8.8}$$

one can write the equation of the regression curve that is the formula of the function expressing the conditional expected value, $E(Y|X=x)=y(x)$, by utilizing formula

(8.7), as

$$E(Y|X=x)=\bar{y}(x)=\frac{\int_{-\infty}^{\infty} vh(x,v)\,dv}{\int_{-\infty}^{\infty} h(x,v)\,dv}=m_2+\varrho\frac{\sigma_2}{\sigma_1}x-\varrho\frac{\sigma_2}{\sigma_1}m_1 \tag{8.9}$$

where $m_1=E(X)$; $m_2=E(Y)$, $\sigma_1=D(X)$, $\sigma_2=D(Y)$ and $\varrho=\varrho(X,Y)$ is the correlation coefficient. So in case of bivariate normal distribution the regression curve is a straight line:

$$\bar{y}(x)=m_2+\varrho\frac{\sigma_2}{\sigma_1}(x-m_1) \tag{8.10}$$

that is

$$\frac{\bar{y}-m_2}{\sigma_2}=\varrho\frac{x-m_1}{\sigma_1}. \tag{8.11}$$

Relationship (8.11) means that if X and Y values are standardized that is if variables

$$Y^*=\frac{Y-m_2}{\sigma_2} \quad \text{and} \quad X^*=\frac{X-m_1}{\sigma_1}$$

are introduced then the regression curve of variable Y^* related to X^* is a straight line of the form

$$y=\varrho x. \tag{8.12}$$

This line passes through the origo and its direction is determined by $\varrho=\varrho(X,Y)$, the coefficient of correlation between the two variables. Note that a curve of regression between two variables may be a straight line even when their joint distribution is not bivariate normal. If the regression between X and Y is linear the equation of the regression line, $y=ax+b$, will be given.

Remark:

We have seen in formula (7.47) that the quantile curve for two-dimensional normal distribution is

$$\tilde{y}(x)=\frac{\sigma_2}{\sigma_1}(x-m_1)+m_2=G^{-1}\{F(x)\}.$$

From (8.10) follows that

$$\bar{y}(x)-m_2=\varrho\frac{\sigma_2}{\sigma_1}(x-m_1)=\varrho\tilde{y}(x)=\varrho G^{-1}[F(x)].$$

It means that in the normal case:

$$\bar{y}(x)=\varrho G^{-1}[F(x)]+(1-\varrho)E(Y) \tag{8.13}$$

which is a quite simple relation between the quantile curve and the linear regression line.

The fact is pointed out here that if the random variables X and Y have the joint distribution $H(x, y)$ with marginals $F(x)$ and $G(y)$ and the marginals are the same type of distribution i.e. $F(x)$ and $G(y)$ differ from each other only in the value of parameters, then the quantile curve is linear and (8.13) holds.

8.1.3. APPLICATION OF THE QUANTILE CURVE TO RAPID DETERMINATION OF THE RELATION BETWEEN THE MAGNITUDE OF EXCEEDANCE AND FLOODING DURATION

The corresponding values of X exceedances and Y flood durations in case of flood waves observed in the Tisza river at Szeged are given in Table T.1. This is the case utilized to demonstrate how to express the relation of random variables X and Y by means of the quantile curve. Plotting the corresponding values as a cluster of points on a plane the pattern of dispersion shown in Fig. 74 is obtained. To have a measure for the closeness of relation between the random variables X and Y the indicator correlation $\tilde{\varrho}_\alpha$ is calculated with $\alpha=1/2$. Using the empirical medians, $\tilde{x}_{1/2}$ and $\tilde{y}_{1/2}$, what is obtained is

$$\tilde{\varrho}_{1/2} = 4\frac{k}{n} - 1 = 4 \cdot \frac{14}{30} - 1 = 0.86$$

which value indicates a close relation so that a search for the shape of a function is worth while. On the basis of those set forth in Section 8.1.10 if there is a monotonic functional relation $Y=\varphi(X)$ between the random variables X and Y the function describing this relation is the quantile curve of both variables: $\varphi(x)=G^{-1}[F(x)]$. The quantile curve is a curve passing through the joint quantile points $(\tilde{x}_{\alpha 1}, \tilde{y}_{\alpha 1})$, $(\tilde{x}_{\alpha 2}, \tilde{y}_{\alpha 2})$, ..., $(\tilde{x}_{\alpha n}, \tilde{y}_{\alpha n})$. Produce now, by using the table, the ordered samples

$$X_1^* < X_2^* < \ldots < X_{31}^*$$

and

$$Y_1^* < Y_2^* < \ldots < Y_{31}^*$$

and, with respect to the fact that each element of the ordered sample is a certain quantile of the distribution, plot the set consisting of points (X_1^*, Y_1^*), (X_2^*, Y_2^*),..., ..., (X_{31}^*, Y_{31}^*) and connect these points by straight sections. In this way a polygon will be formed which provides the approximate shape of the quantile curve (see Fig. 74). Apparently, the relation between variables X and Y is of linear character. Since, as it was said in Section 6.3.3, the magnitude of exceedances, the random variable X, is distributed exponentially, it follows that if the duration of flooding, Y, is a linear function of X then Y is also distributed exponentially.

The quantile curve of exponentially distributed two random variables is given as

$$\varphi(x) = G^{-1}[F(x)] = \frac{E(Y)}{E(X)} \cdot x.$$

With the data of Table T.1

$$E(Y) \approx \overline{Y} = 23.58 \quad \text{and} \quad E(X) \approx \overline{X} = 100.35.$$

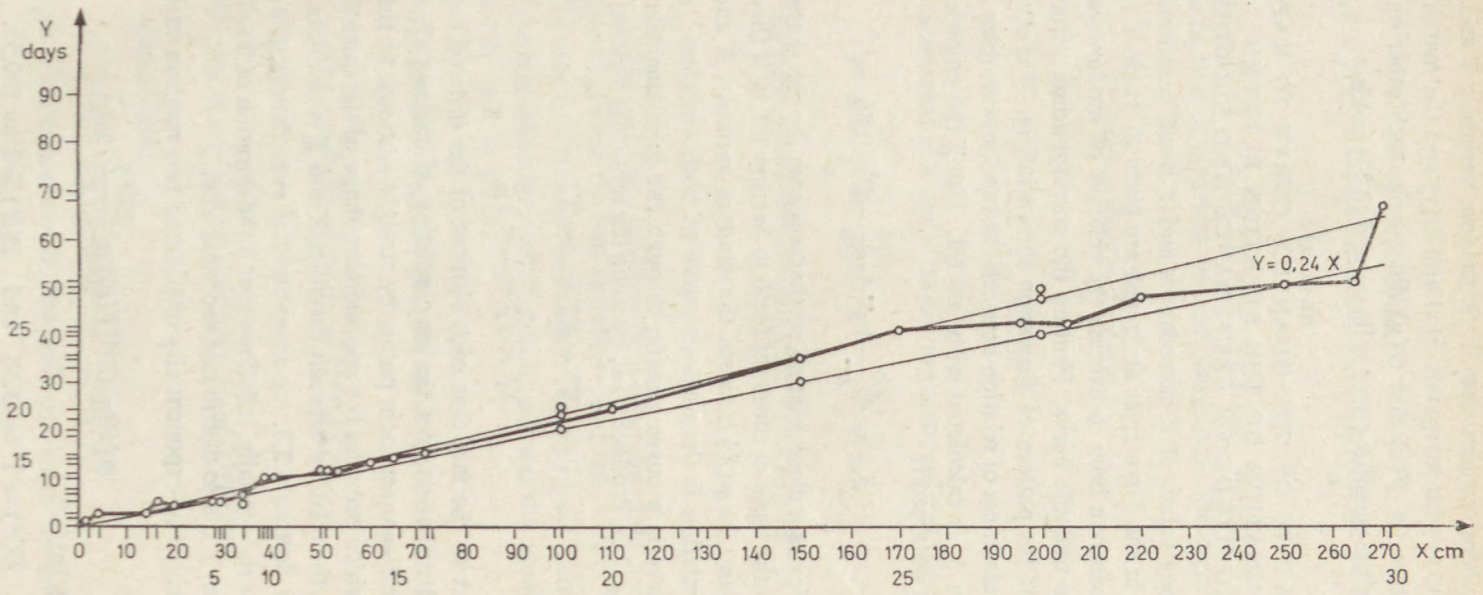

Y
days
90
80
70
60
50
25 40
30
20 20
15
10
0
Y = 0,24 X
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 X cm
5 10 15 20 25 30

Figure 74

So the equation giving the quantile curve is

$$\varphi(x) = G^{-1}[F(x)] = \frac{23.58}{100.35}x \approx \frac{24}{100}x = 0.24x.$$

By means of formula (8.13) we have

$$\bar{y}(x) \approx \varrho G^{-1}[F(x)] + (1-\varrho)E(Y) = 0.86 \cdot 0.24x + 0.04 \cdot 23.58 = 20.64 + 0.94.$$

In this case we have the following rapid estimation for the duration of flood:

$$E(Y|X) \approx \frac{X\,\text{cm}}{5} + 1\,\text{days},$$

which is valid only for the River Tisza at Szeged in the second quarter of the year.

8.1.4. ESTIMATION OF LINEAR REGRESSION FROM A STATISTICAL SAMPLE

When the distribution of a random vector variable (X, Y) is not known the equation of the regression curve $E(Y|X=x)=y(x)$ cannot be derived. Suppose that a two-dimensional statistical sample (x_1, y_1), (x_2, y_2) ..., (x_n, y_n) is available for a vector variable (X, Y). After plotting the sample as a cluster of points on a plane the shape of this cluster will frequently suggest such an approximation where the relation between the variables X and Y is represented by a linear function. In such cases it is supposed that the regression of variable Y related to X has the form $Y=aX+b$. By virtue of the least squares principle the quantification of a and b may take place under the condition that $E(Y-aX-b)^2=$minimum. The determination of constants a and b requires the knowledge of the first and second moments of the random variables in the parenthesis and this necessitates the knowledge of the joint density function of the pair of variables (X, Y). Therefore the straight line $y=ax+b$ is derived directly from the sample, based on the condition that

$$f(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \text{minimum}.$$

Now the values (x_i, y_i) are given numbers while the parameters a and b are not known and what is to be determined is the extreme value of the bivariate function of these variables. The solution is obtained by equating the derivates to zero:

$$\frac{\partial f(a, b)}{\partial a} = -2 \sum_{i=1}^{n} (y_i - ax_i - b)x_i = 0 \tag{8.14}$$

$$\frac{\partial f(a, b)}{\partial b} = -2 \sum_{i=1}^{n} (y_i - ax_i - b) = 0. \tag{8.15}$$

From Eqs. (8.14) and (8.15) the following system of linear equations is obtained for the unknown a and b:

$$(\sum x_i^2)a + (\sum x_i)b = \sum x_i y_i$$
$$(\sum x_i)a + nb = \sum y_i.$$

By the application of Cramer's rule:

$$a = \frac{\begin{vmatrix} \sum x_i y_i & \sum x_i \\ \sum y_i & n \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum y_i & n \end{vmatrix}} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} =$$

$$= \frac{\frac{\sum x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\sum x_i^2}{n} - \bar{x}^2},$$

(8.16)

$$b = \frac{\begin{vmatrix} \sum x_i^2 & \sum x_i y_i \\ \sum x_i & \sum y_i \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum y_i & n \end{vmatrix}} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} =$$

$$= \frac{\frac{\sum x_i^2}{n} \bar{y} - \bar{x} \frac{\sum x_i y_i}{n}}{\frac{\sum x_i^2}{n} - \bar{x}^2}.$$

In this way practically the problem has been solved but the formulae obtained will be somewhat transformed as follows:

$$a = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\sum x_i^2}{n} - \bar{x}^2} = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}}{\frac{\sum (x_i - \bar{x})^2}{n}} =$$

$$= \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}}{\sqrt{\frac{\sum (x_i - x)^2}{n} \frac{\sum (y_i - \bar{y})^2}{n}}} \frac{\sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}}{\sqrt{\frac{\sum (x_i - \bar{x})}{n}}} = r \frac{S_y}{S_x}$$

(8.17)

$$b = \frac{\frac{\sum x_i^2}{n} \bar{y} - x^2 y + \bar{x}^2 \bar{y} - \bar{x} \frac{\sum x_i y_i}{n}}{\frac{\sum x_i^2}{n} - \bar{x}} =$$

$$= \bar{y} - \frac{\frac{\sum x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\sum x_i^2}{n} - \bar{x}^2} \bar{x} = \bar{y} - a\bar{x}.$$

Let be recalled here that the estimation of correlation coefficient ϱ is the empirical correlation coefficient

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{8.18}$$

and that to estimate the expected value of variables X and Y, the respective sample means $\bar{x}$ and $\bar{y}$ can be used while the estimations of variances $D(X)$ and $D(Y)$ can be given by

$$S_x = \sqrt{\frac{\sum (x_i - x)^2}{n}}$$

and

$$S_y = \sqrt{\frac{\sum (y_i - y)^2}{n}},$$

respectively.

So the result is in accordance with the theoretical results contained in (8.10) and at the same time the statistical estimations of the theoretical parameters contained in (8.10) have also been produced.

Consequently, the formula which can be used to calculate the empirical regression line of Y related to X from a statistical sample is

$$y = r \frac{S_y}{S_x} x + \bar{y} - r \frac{S_y}{S_x} \bar{x}. \tag{8.19}$$

8.1.5. REGRESSION SURFACE AND PLANE

Usually the quantities playing part in hydrological research depend on several factors. It is necessary, therefore, to clarify how a given critical quantity depends on other quantities and out of these which the given quantity depends strongly or weakly on. In many cases such a question should also be answered whether a strong relation between two quantities is causal indeed or the semblance of a close relation is caused by other quantities. Below r random variables will be considered and out of them one will be tested as a function of the remaining variables. All the variables involved are supposed to be random variables so that the fluctuation of this vector variable may be attributed to chance.

Denote the set of random variables concerned by a random vector $(X_1, X_2, \ldots, X_r)$ and the joint density function of these variables by $f(x_1, x_2, \ldots, x_r)$. If a variable, X_1, is to be approached by some function of the remaining variables, $g(x_2, x_3, \ldots, x_r)$, then — as it was seen — the least squares method will lead to the conditional expected value:

$$E(X_1|x_1, x_2, \ldots, x_r) = g(x_2, x_3, \ldots, x_r) =$$

$$= E(X_1|X_2 = x_x, X_3 = x_3, \ldots, X_r = x_r) =$$

$$= \frac{\int_{-\infty}^{\infty} x_1 f(x_1, x_2, \ldots, x_r)\, dx_1}{\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_r)\, dx_1}.$$

So X_1 — in the sense of averages mentioned several times before — may be substituted by the conditional mean of the vector variable $(X_2, X_3, \ldots, X_r)$. The above expression of $g(x_2, x_3, \ldots, x_r)$ is called the regression surface of X_1 related to $(X_2, X_3, \ldots, X_r)$. In the special case where the joint distribution of variables $(X_1, X_2, \ldots, X_r)$ is n-dimensional normal the regression surface is a plane that is

$$E(X_1 | X_2 = x_2, \ldots, X_r = x_r) =$$

$$= a_{12}X_2 + a_{13}X_3 + \ldots + a_{1r}X_r + a_1.$$

Instead of variable X_1, of course, any one of the variables may be expressed as a function of the remaining ones.

Just as the regression curve, which belogs to two variables, can be substituted by a straight line, when dealing with more than two random variables frequently an approximation through a linear function to the chosen X_1 variable is considered satisfactory as well: such a regression "plane" (in fact an $(r-1)$-dimensional, so-called, hyper plane) is determined to which — in accordance with the principle of least squares — the deviations of the X_1 values are, in average, the least.

For sake of simplicity suppose that the expected value of each variable is zero that is

$$E(X_i) = 0 \quad (i = 1, 2, \ldots, r).$$

This can be attained by subtracting the respective expected value from each random variable.

Now what is searched is an $(r-1)$-variate linear function,

$$a_{12}X_2 + a_{13}X_3 + \ldots + a_{1r}X_r,$$

with which

$$E(X_1 - a_{12}X_2 - a_{13}X_3 \ldots a_{1r}X_r)^2 = \text{minimum}. \tag{8.20}$$

Unlike in the case of two variables here the coefficients are expressed by the aid of the (unknown) moments and then the moments are estimated from the sample.

With the assumption that the joint distribution of the variables concerned is a continuous r-dimensional distribution, Eq. (8.20) can be written in the form of an integral, too:

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} (x_1 - a_{12}x_2 \ldots a_{1r}x_r)^2 f(x_1, x_2, \ldots, x_r)\, dx_1 \ldots dx_r = \text{minimum} \tag{8.21}$$

where $f(x_1, x_2, \ldots, x_r)$ is the joint density function.

To denote the mixed moments (covariance) the notation used is

$$b_{ij} = b_{ji} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i x_j f(x_1, \ldots, x_r)\, dx_1 \ldots dx_r. \tag{8.22}$$

In a special case:

$$b_{ii} = E(X_i^2) = D^2(X_i) \quad (i = 1, 2, \ldots, r). \tag{8.23}$$

Through derivation the following linear equation system is obtained for the unknown coefficients, a_{ij}:

$$b_{i2}a_{12} + b_{i3}a_{13} + \ldots + b_{ir}a_{1r} = b_{i1} \quad (i = 2, 3, \ldots, r). \tag{8.24}$$

Supposing that the covariance matrix

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1r} \\ b_{21} & b_{22} & \ldots & b_{2r} \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ b_{r1} & b_{r2} & \ldots & b_{rr} \end{bmatrix} \tag{8.25}$$

is not singular (that is $|B| \neq 0$), expression

$$a_{ik} = -\frac{B_{1k}}{B_{11}} \quad (k = 1, 2, \ldots, r) \tag{8.26}$$

is obtained as solution where B_{1k} and B_{11} denote the algebraic subdeterminants belonging to the elements b_{1k} and b_{11}, respectively.

The correlation coefficient of variables X_i and X_j can also be calculated by using the elements of matrix $\mathbf{B}$:

$$\varrho(X_i, X_j) = \frac{E(X_i, X_j)}{D(X_i)D(X_j)} = \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}} \tag{8.27}$$

$$(i, j = 1, 2, \ldots, r).$$

Produce now the difference

$$Y_1 = X_1 - a_{12}X_2 - a_{13}X_3 - \ldots - a_{1r}X_r =$$

$$= \frac{1}{B_{11}} \sum_{j=1}^{r} B_{1j}X_j,$$

the so-called residue, through subtracting the "best" linear approximation of X_1 from X_1. It is easy to realize that Y_1 is uncorrelated to variables X_2, X_3, ..., X_r and, in addition, when an r-dimensional normal distribution is the case, it is independent of them but is positively correlated to X_1 since

$$E(Y_1, X_i) = \frac{1}{B_{11}} \sum_{j=1}^{r} b_{1j}b_{ij} = \tag{8.28}$$

$$= \begin{cases} \dfrac{|B|}{B_{11}} & \text{if} \quad i = 1 \\ 0 & \text{if} \quad i \neq 1. \end{cases}$$

(Here $|B|$ denotes the determinant of the covariant matrix **B**.) Since according to our assumption the expected values of variables $X_1, X_2, \ldots, X_r$ are equal to zero the consequence is that $E(Y_1)=0$ and so

$$D^2(Y_1) = E(Y_1^2) = E(Y_1 \cdot X_1) = \frac{|B|}{B_{11}} 0.$$

The fact that $|B|$ and B_{11} are non-negative is a consequence of the symmetry of matrix **B**.

The coefficient of correlation between Y_1 and X_1 is

$$(Y_1, X_1) = \frac{E(Y_1 \cdot X_1)}{D(Y_1)D(X_1)} = \sqrt{\frac{|B|}{b_{11}B_{11}}}.$$

If the statistical sample used to represent an r-dimensional random variable $(X_1, X_2, \ldots, X_r)$ is

$$(X_{1j}, X_{2j}, \ldots, X_{rj}) \quad (j = 1, 2, \ldots, r)$$

the following estimate may be used in the formulae above:

$$b_{ik} \approx \hat{b}_{ik} = \frac{1}{r-1} \sum_{j=1}^{r} (x_{ij} - \bar{x}_{i.})(x_{kj} - \bar{x}_{k.}).$$

Applying this the equation of the empirical regression plane is

$$X_1 = \frac{1}{\hat{B}_{11}} [\hat{B}_{12}(x_2 - \bar{x}_{2.}) + \hat{B}_{13}(x - \bar{x}_{3.}) + \ldots + \hat{B}_{1r}(x_r - \bar{x}_{r.})] \tag{8.29}$$

where $\hat{B}_{ij}$ is the estimate of subdeterminant B_{ij} if the b_{ij} elements of matrix **B** are substituted by the $\hat{b}_{ij}$ numbers.

8.1.6. MULTIVARIATE LINEAR FUNCTIONAL RELATIONSHIP. GAUSSIAN NORMAL EQUATIONS

Suppose now that a variable y is a linear function of variables $x_1, x_2, \ldots, x_s$:

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_s x_s$$

where the coefficients $a_1, a_2, \ldots, a_s$ are to be determined on the basis of measurements.

Suppose that measurements had taken place at points $x_{1i}, x_{2i}, \ldots, x_{si}$ $(i = 1, 2, \ldots, n)$ and that from an experiment which had been performed on the i-th system of value a quantity subjected to random fluctuation, Y_i, was obtained for y that is

$$Y_i = a_1 x_{1i} + a_2 x_{2i} + \ldots + a_s x_{si} + \varepsilon_i \quad (i = 1, 2, \ldots, n) \tag{8.30}$$

where the errors ε_i are supposed to have zero as expected value, to be uncorrelated and to have the same variance:

$$E(\varepsilon_i) = 0, \quad D^2(\varepsilon_i) = \sigma^2;$$

$$E(\varepsilon_i \varepsilon_j) = 0 \quad \text{if} \quad i \neq j \quad (i, j = 1, 2, \ldots, n).$$

In accordance with the least squares method as a starting point for the estimation of the coefficients the following condition has to be used:

(8.31)
$$F(a_1, a_2, \ldots, a_s) =$$

$$= \sum_{i=1}^{n} (Y_i - a_1 X_{1i} - a_2 X_{2i} - \ldots - a_s X_{si})^2 = \text{minimum}.$$

Hence

$$\frac{\partial F}{\partial a_k} = -\sum_{i=1}^{n} (Y_i - a_1 X_{1i} - a_2 X_{2i} - \ldots - a_s X_{si}) X_{ki} = 0 \quad (k = 1, 2, \ldots, s)$$

that is

$$\sum_{i=1}^{n} \sum_{j=1}^{s} a_j X_{ji} X_{ki} = \Sigma Y_i X_{ki}$$

and so

(8.32)
$$\sum_{j=1}^{s} a_j \sum_{k=1}^{n} X_{ji} X_{ki} = \sum_{i=1}^{n} Y_i X_{ki} \quad (k = 1, 2, \ldots, s).$$

A simpler form of this equation can be obtained if the notation

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \quad X_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1n} \end{bmatrix}; \quad X_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2n} \end{bmatrix}; \quad \ldots, X_s = \begin{bmatrix} X_{si} \\ X_{s2} \\ \vdots \\ X_{sn} \end{bmatrix}$$

is introduced.

Eq. (8.32) above can also be written in the form of a scalar product:

$$\sum_{j=1}^{s} a_j (X_j, X_k) = (Y_i, X_k) \quad (k = 1, 2, \ldots, s).$$

The whole equation system in a detailed form is

(8.33)
$$\begin{aligned} &a_1(X_1, X_1) + a_2(X_2, X_1) + \ldots + \\ &+ a_s(X_s, X_1) = (Y, X_1) \\ &a_1(X_1, X_2) + a_2(X_2, X_2) + \ldots + \\ &+ a_s(X_s, X_2) = (Y, X_2) \\ &\vdots \\ &a_1(X_1, X_3) + a_2(X_2, X_3) + \ldots + \\ &+ a_s(X_s, X_3) = (Y, X_3). \end{aligned}$$

Equations (8.33) are called Gaussian or normal equations.

The solution to equation system (8.33) provides the estimates of coefficients α_1, α_2, ..., α_s:

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_s \end{bmatrix} = \begin{bmatrix} \\ \mathbf{D}^{-1} \\ \\ \end{bmatrix} \begin{bmatrix} (Y, X_1) \\ (Y, X_2) \\ \vdots \\ (Y, X_s) \end{bmatrix} \tag{8.34}$$

where matrix **D**, the so-called Gram matrix, is evidently a symmetrical one:

$$\mathbf{D} = \mathbf{D}^* = \begin{bmatrix} (X_1, X_1) & (X_1, X_2) & \dots & (X_1, X_s) \\ (X_2, X_1) & (X_2, X_2) & \dots & (X_2, X_s) \\ \vdots & & & \\ (X_s, X_1) & (X_s, X_2) & \dots & (X_s, X_s) \end{bmatrix}. \tag{8.35}$$

This is essentially a geometric problem: out of s linearly independent vectors X_1, X_2, ..., X_s located in the n-dimensional space which one is closest to vector Y.

Figure 75 below illustrates the solution to the case where $s=2$:

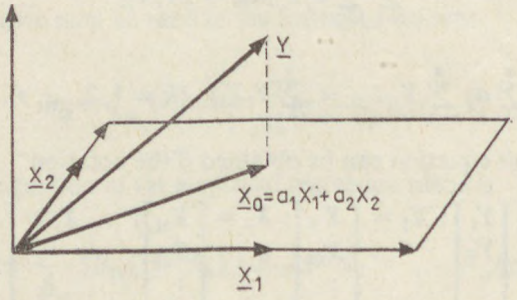


Figure 75

It is seen that vector X_0, the solution, is the orthogonal projection of Y on the plane (subspace) spanned by vectors X_1 and X_2. Since in this case vector $Y-X_0$ is perpendicular to the subspace consisting of all linear combinations of vectors X_1 and X_2 it follows that it is perpendicular to all vectors in this sub-space, including vectors X_1 and X_2, so that

$$(Y-X_0, X_1) = 0$$
$$(Y-X_0, X_2) = 0.$$

Hence

$$(Y, X_1) = (X_0, X_1) = (a_1X_1+a_2X_2, X_1) =$$
$$= a_1(X_1, X_1)+a_2(X_2, X_1)$$
$$(Y, X_2) = (X_0, X_2) = (a_1X_1+a_2X_2, X_2) =$$
$$= a_1(X_1, X_2)+a_2(X_2, X_2).$$

The situation is analogous in case of any finite s-dimensional sub-space.

The solution to the normal equation system in unambiguous when the $|\mathbf{D}|$ determinant of matrix $\mathbf{D}$ differs from zero that is when vectors $X_1, X_2, \ldots, X_s$ are linearly independent. In this case it can be proved that the α_i components of the solution vector $(\alpha_1, \alpha_2, \ldots, \alpha_s)$ are unbiased estimates of coefficients $a_i (i=1, 2, \ldots, s)$ and, what is more, it is these estimates that have the least variance among all the unbiased linear estimates. (Theorem of Gauss.)

The variance and correlation coefficient of estimates α_i can be calculated in a relatively simple manner. Denote by D_{jk} the $(s-1)$ order subdeterminant belonging to the element (X_j, X_k) of determinant $|D|$ and introduce the notation $Q_{jk}=D_{jk}/|D|$. It can be shown that

$$E[(\alpha_j - a_j)(\alpha_k - a_k)] = Q_{jk}\sigma^2 \quad (j, k = 1, 2, \ldots, s).$$

If $j=k$ then the variance of random variable α_j is obtained:

$$D^2(\alpha_j) = Q_{jj}\sigma^2 \quad (j = 1, 2, \ldots, s).$$

The correlation coefficient is

$$\varrho(\alpha_j, \alpha_k) = \frac{Q_{jk}}{\sqrt{Q_{jj}Q_{kk}}} (j, k = 1, 2, \ldots, s).$$

If the σ^2 variance of errors ε_i is not known, a situation occurring in most of the cases, an estimation can be found to σ^2 through formula

$$S^2 = \frac{\sum_{i=1}^{n} (Y_i - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_s x_{si})^2}{n-s}$$

which is an unbiased estimator of σ^2:

$$E(S^2) = \sigma^2.$$

With an assumption that the distribution of each error ε_i is normal the joint distribution of $(\alpha_1, \alpha_2, \ldots, \alpha_s)$ will be an s-dimensional normal distribution and the distribution of S^2 will be a χ^2 distribution with parameter $(n-s)$ and with variance

$$D^2(S^2) = \frac{2\sigma^4}{n-s}.$$

So in case of normal distribution S^2 is a strongly consistent estimate to σ^2.

Note that in the above application of the principle of least squares such an assumption was made that the values of variables X_i were known accurately. In practice sometimes the situation is different from this. Nevertheless, if the values of variables X_i can be measured much more accurately than those of Y the above statistical statement may be considered valid, in approximation.

8.1.7. POLYNOMIAL REGRESSION

A case where the planar representation of a sample (X_i, Y_i) $(i=1, 2, \ldots, n)$ shows that no good approximation would be obtained by linear regression can be discussed in a quite similar way. Now variable Y is approached by a polynom of variable X:

$$Y \sim a_0+a_1X+a_2X^2+\ldots+a_sX^s.$$

Here the number of coefficients to be determined is $(s+1)$. Consider the case where the locations of measurements are $x_1, x_2, \ldots, x_n$ and the result of measurement at a point x_i is the random variable Y_i:

$$Y_i = a_0+a_1x_i+a_2x_i^2+\ldots+a_sx_i^s+\varepsilon_i.$$

Suppose that variables ε_i are uncorrelated random variables with equal variances:

$$E(\varepsilon_i) = 0, \quad E(\varepsilon_i, \varepsilon_j) = 0 \quad \text{if} \quad i \neq j,$$

$$D^2(\varepsilon_i) = \sigma^2 \quad (i, j = 1, 2, \ldots, n).$$

The least squares method leads now to the following system of equations which corresponds to the normal equations:

$$\sum_{j=1}^{s} a_j(X_j, X_k) = (Y, X_k)$$

where

$$X_j = \begin{bmatrix} x_1^j \\ x_2^j \\ \vdots \\ x_s^j \end{bmatrix}; \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{bmatrix}.$$

From a statistical point of view the estimates $\alpha_0, \alpha_1, \ldots, \alpha_s$ obtained for the unknown coefficients by solving the equation system behave in the same manner as in the previous case.

Formally the estimation of coefficients may be performed in the same way when both X and Y are random variables; however, in this case the statistical behavior of estimates may no longer be examined in such a simple way. It can also be seen that the calculation of polinomial regression can be traced back to the case of multivariate linear regression.

8.1.8. PARTIAL CORRELATION

Consider the random variables $X_1, X_2, \ldots, X_r$ and calculate the coefficient of correlation between variables X_i and X_k:

$$\varrho(X_i, X_k) = \varrho_{ik} =$$

$$= \frac{E\{[X_i-X(X_i)][X_k-E(X_k)]\}}{D(X_i)D(X_k)}$$

which is called customarily the coefficient of total correlation. This coefficient, as it was explained in Section 2.1.7, might take any value between -1 and $+1$ and if its absolute value was close to 1 then it was said there was a close correlation between X_i X_k. A close correlation may appear also as a consequence of a situation where the development of both X_i and X_k values is influenced by their dependence on the other random variables. In fact, there is no causal relation between X_i and X_k so that if the impact of the other variables were eliminated no close correlation could be found between them. Here the question may arise how to eliminate from a relation between two random variables the impact of other variables.

Out of n random variables select two variables, say X_1 and X_2 (the numbering of variables is, of course, arbitrary). Express both X_1 and X_2 through the best approximate linear combinations composed of variables $X_3, X_4, \ldots, X_n$:

$$X_1 = \sum_{j=3}^{n} a_{ij} X_j; \; X_2 = \sum_{k=3}^{n} a_{2j} X_j.$$

Now construct the variables

$$Y_1 = X_1 - \sum_{j=3}^{n} a_{1j} X_j \; Y_2 = X_2 - \sum_{k=3}^{n} a_{2j} X_j,$$

the so-called residues, for which it can be shown that they are uncorrelated to variables $X_3, \ldots, X_n$, moreover, in case of joint normal distribution, they are independent of them but there are positive correlations between Y_1 and X_1 and between Y_2 and X_2.

The coefficient of correlation between residues Y_1 and Y_2 is called partial correlation between X_1 and X_2, related to variables $(X_3, \ldots, X_n)$. The parial correlation is considered to reflect the relation between X_1 and X_2 in its cleaned form, without the impact of variables $X_3, \ldots, X_n$.

The partial correlation can be calculated by using the elements of matrix **B** as given in (8.25):

$$\varrho(Y_1, Y_2) = \frac{E(Y_1, Y_2)}{D(Y_1)D(Y_2)} = -\frac{B_{12}}{\sqrt{B_{11}B_{12}}} \tag{8.36}$$

where B_{ij} denotes the algebraic subdeterminant belonging to the element b_{ij} of matrix **B**.

If $n=3$, that is when in the course of analysing the relation between X_1 and X_2 what is to be eliminated is the impact of a single third variable X_3, the partial correlation can be calculated as well by utilizing the coefficient of the pairwise total correlations. Introducing the notation

$$Y_1 = X_1 - a_{13}X_3; \quad Y_2 = X_2 - a_{23}X_3;$$

$$\varrho_{12} = \varrho(X_1, X_2); \quad \varrho_{13} = \varrho(X_1, X_3);$$

$$\varrho_{23} = \varrho(X_2, X_3);$$

formula

$$\varrho(Y_1, Y_2) = \frac{\varrho_{12} - \varrho_{13}\varrho_{23}}{\sqrt{(1-\varrho_{13}^2)(1-\varrho_{23}^2)}} \tag{8.37}$$

is obtained. When the joint distributions are not known then, utilizing the statistical sample and using the methods desribed in the previous sections the corresponding empirical correlation coefficients are drawn into calculation.

8.1.9. MULTIPLE CORRELATION

In certain cases there may be a requirement that the stochastic relation of a certain random variable X_1 should be measured by a set of variables $X_2, X_3, \ldots, X_n$, instead of a single one. Resting on the foregoing it seems to be expedient that this relation is expressed by the relation between variable X_1 and variable

$$X_1^* = a_{12}X_2 + a_{13}X_3 + \ldots + a_{1n}X_n \tag{8.38}$$

(the "best" linear approximation of X_1) that is a correlation coefficient

$$\varrho(X_1, X_1^*) = \frac{E(X_1, X_1^*)}{D(X_1)D(X_1^*)} \tag{8.39}$$

is formed which is called the coefficient of multiple correlation between X_1 and $(X_2, \ldots, X_n)$. The coefficient of multiple correlation can be expressed as well by using the covariance matrix B as it is given in Eq. (8.25):

$$\varrho(X_1, X_1^*) = \sqrt{1 - \frac{|B|}{b_{11}B_{11}}}\,. \tag{8.40}$$

APPENDIX

COMBINATORIAL TOOLS

Combinatory deals with the counting problems related to finite sets. Let A denote the set of the first n integers

$$A = \{1, 2, \ldots, n\}$$

i.e. a set of n different elements.

a) *Permutation*

As a first problem we discuss the following question: what is the number of different arrays of the elements, $1, 2, \ldots, n$ while each number may occur only once? Any number can be put in the first place, any of the remaining $n-1$ can be put in the second, any of the remaining $n-2$ in the third, etc. Consequently, the number of all the possible arrays is $n.(n-1)\ldots 3.2.1.$

Each ordering of numbers $1, 2, \ldots, n$ is called a permutation.

Let the set of all possible permutations of $1, 2, \ldots, n$ be denoted by $\mathscr{P}_n$ that is called the space of permutations of order n. The number of elements of $\mathscr{P}_n$ is denoted by P_n. Then

$$P_n = n \cdot (n-1) \ldots 3 \cdot 2 \cdot 1 = n!$$

i.e. n factorial. Thus, the value of $n!$ is to be calculated by multiplying the natural numbers from 1 to n.

The sequence $n!$ is a fairly fast increasing function of n.

Table 1.1 contains the results of a number of permutations.

If the value of n is large, $n!$ becomes extremely large. It usually happens that the exact value of $n!$ is of not too much importance, in many cases we are rather interested in its order of magnitude. Thus, the need for approximating $n!$ may arise. A very good estimation for $n!$ can be obtained by use of the Stirling-formula

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n},$$

where $e = 2.7182\,81\ldots$ is the Napierian number for natural logarithm. For example, if $n = 10$ then the Stirling-formula gives

$$\left(\frac{10}{e}\right)^{10} \sqrt{20\pi} = 3\,598\,600.$$

Table 1.1

Table of Factorials

n	$n!$	n	$n!$
1	1	26	40 329 146 10^{19}
2	2	27	10 888 869 10^{21}
3	6	28	30 488 834 10^{22}
4	24	29	88 417 620 10^{23}
5	120	30	26 525 286 10^{25}
6	720	31	82 228 387 10^{26}
7	5 040	32	26 313 084 10^{28}
8	40 320	33	86 833 176 10^{29}
9	362 880	34	29 525 280 10^{31}
10	3 628 800	35	10 333 148 10^{33}
11	39 916 800	36	37 199 333 10^{34}
12	47 900 160	37	13 763 753 10^{36}
13	62 270 208	38	52 302 262 10^{37}
14	87 178 291	39	20 397 882 10^{39}
15	13 076 774	40	81 591 528 10^{40}
16	20 922 790	41	33 452 527 10^{42}
17	35 568 743	42	14 050 061 10^{44}
18	64 023 737	43	60 415 263 10^{45}
19	12 164 510	44	26 582 716 10^{47}
20	24 329 020	45	11 962 222 10^{49}
21	51 090 942	46	55 026 222 10^{50}
22	11 240 007	47	25 862 324 10^{52}
23	25 852 017	48	12 413 916 10^{54}
24	62 044 840	49	60 828 186 10^{55}
25	15 511 210	50	30 414 093 10^{57}

(0! = 1, is a purposeful convention)

Comparison to the precise value of 10! = 3628800 yields a relative error of 0.8 per cent. The larger the number of n the better the approximation, i.e. the smaller the relative error.

b) *Permutations with repetition*

Assume that certain elements $k_1, k_2, \ldots, k$ of a set $A = \{a_1, a_2, \ldots a_n\}$ are identical, and

$$k_1 + k_2 + \ldots + k_r = n.$$

If all possible sequence of these elements are considered than the number of different arrays is:

$$P_n^{(\text{rep})} = \frac{n!}{k_1! k_2! \ldots k_r}.$$

As an example, consider the letters of the word MATHEMATICA. The number of all different sequences is:

$$\frac{11!}{2!\,3!\,2!\,1!\,1!} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 6 \cdot 2} = 1\,663\,200.$$

c) *Variations*

Let k elements be chosen from set $A=\{1, 2, \ldots, n\}$ one after the other and be written in the order of selection. The sequence of elements is called a variation. What is the number of different k-element variations from the elements of set A? Any of the n elements can be selected for the first place, any of the $n-1$ elements for the second, any of the $n-2$ for the third, etc. For the k-th place any of the remaining $(n-k+1)$ elements may be selected. Therefore, the number of k-element variations of set A is

$$V_k^n = n(n-1)\ldots(n-k+1) = \frac{n!}{(n-k)!}.$$

Consider now the following allocation problem: Given k elements, $a_1, a_2, \ldots, a_k$. Let them be located into $n \geqq k$ cells labelled by $1, 2, \ldots, n$ in such a way that any cell may contain one element only. What is the number of the possible allocations?

Element a_1 can be set in any of the n cells. Element a_2 then can be put in any of the remaining $(n-1)$ cells. The number of all possible allocations is:

$$n(n-1)\ldots(n-k+1).$$

d) *Variations with repetition*

Consider again the previous allocation problem but without the constraint to allow at most one element in one cell. (It is allowed to allocate all elements in the same cell.) Therefore, any cell may contain any subset of $a_1, a_2, \ldots, a_k$. What is the number of possible allocations in this case? Any of the n cells may be selected for element a_1 any for element a_2, etc. To allocate elements a_1 and a_2 we have altogether n^2 possibilities, similarly, we have n^3 possibilities to allocate elements a_1, a_2 and a_3, etc. The number of possibilities to allocate elements $a_1, a_2, \ldots, a_k$ is then:

$$V_k^n(\text{rep}) = n^k.$$

e) *Combinations*

Consider again the allocation problem where elements $a_1, a_2, \ldots, a_k$ are allocated into n cells in such a way that there is only one element in one cell. Assume that elements $a_1, a_2, \ldots, a_k$ are identical, i.e. $a_1=a_2=\ldots=a_k$.

Say, k undistinguishable objects are allocated in n cells. How many possibilities do we have?

The allocation of k undistinguishable objects in n cells is equivalent to the selection of k different cells from n different cells, irrespectively of the order of selection. These

allocations are called combinations. The number of all distinguishable combinations is now:

$$C_k^n = \frac{n(n-1)\ldots(n-k+1)}{1\cdot 2\cdot \ldots \cdot k} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

Expression $C_k^n = \binom{n}{k}$ is for the number of all subsets containing k elements of set $A=\{1, 2, \ldots, n\}$. Consequently, out of $1, 2, \ldots, n$ numbers altogether $\binom{n}{k}$ options are available to select k different numbers.

As an application the *binomial law* may be mentioned used to perform the n-th power of an expression consisting of two members:

$$(p+q)^n = (p \overset{1}{+} q)(p \overset{2}{+} q)\ldots(p \overset{n}{+} q).$$

Multiplication is performed in such a way as to select one member from each factor, these are then multiplied in turn and the products are added up. If q is selected from each factor then q^n is obtained. There is only one way to get this expression. If member p has been selected out of k factors and q from the remaining $n-k$ then one may get members like $p^k q^{n-k}$. These can be obtained in the same number as k terms may be selected from n terms, i.e. in $\binom{n}{k}$ different ways. Therefore

$$(1.1) \qquad (p+q)^n = \binom{n}{0} p^0 q^n + \binom{n}{1} p q^{n-1} + \ldots +$$

$$+ \binom{n}{k} p^k q^{n-k} + \ldots + \binom{n}{n} p^n q^0$$

The computation of $\binom{n}{k}$ for large values of n and k is extremely cumbersome due to the large values of higher factorials. The problem, therefore, is how to estimate $\binom{n}{k}$ for large n and k values? In case if n is even having the form $2m$ and $k=m$ the use of Stirlings formula is a good estimate for the middle (largest) element:

$$\binom{2m}{m} = \frac{(2m)!}{(m!)^2} \frac{\left(\frac{2m}{e}\right)^{2m} \sqrt{2\pi \cdot 2m}}{\left(\frac{m}{e}\right)^{2m} \cdot 2\pi m} = \frac{2^{2m}}{\sqrt{\pi \cdot m}}.$$

This estimate will be needed many times in the sequel.

The symmetry property

$$\binom{n}{k} = \binom{n}{n-k}$$

can be obtained easily by substituting the factorials:

$$\frac{n!}{k!(n-k)!} = \frac{n!}{(n-k)!\,[n-(n-k)]!}$$

It should be noted that as a special case of Eq. (1.1) one has

$$(1.2) \qquad (1+1)^n = \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n} = 2^n.$$

On the basis of Eq. (1.2) one can answer the question of how many subsets the finite set $A=\{1, 2, \ldots, n\}$ will have. Set A has

$\binom{n}{1}$ one-element subsets

$\binom{n}{2}$ two-elements subsets

$\binom{n}{3}$ three-element subsets

$\binom{n}{k}$ k-element subsets, ...

If empty set $\emptyset$ and the full set $A=\{1, 2, \ldots, n\}$ are also considered as subsets of A then the total number of subsets is:

$$\binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n} = 2^n$$

f) *Combinations with repetition*

The notion of combination with repetition may be explained by the following allocation problem. Given n cells, labelled by the number 1 to n we allocate k identical (indistinguishable) objects in them. How many combinations are there possible? (Here, more than one object can be put into one cell, moreover all objects may be put into one cell.) For example, for $n=4$ and $k=5$ one possible allocation is:

As n cells are represented by $n-1$ lines, these $n-1$ lines are to be allocated amongst k points each representing some identical object. Every allocation is a configuration of $n-1+k$ sings, from which $n-$ and k are identical, respectively. The number of all distinguishable cases is, obviously, by permutation with repetition:

$$\frac{(n+k-1)!}{k!(n-1)!} = \binom{n+k-1}{k} = \binom{n+k-1}{n-k}.$$

Table T.4

The normal distribution function

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.5000	0.35	0.6368	0.70	0.7580	1.05	0.8531
0.01	0.5040	0.36	0.6406	0.71	0.7611	1.06	0.8554
0.02	0.5080	0.37	0.6443	0.72	0.7642	1.07	0.8577
0.03	0.5120	0.38	0.6480	0.73	0.7673	1.08	0.8599
0.04	0.5160	0.39	0.6517	0.74	0.7703	1.09	0.8621
0.05	0.5199	0.40	0.6554	0.75	0.7734	1.10	0.8643
0.06	0.5239	0.41	0.6591	0.76	0.7764	1.11	0.8665
0.07	0.5279	0.42	0.6628	0.77	0.7794	1.12	0.8686
0.08	0.5319	0.43	0.6664	0.78	0.7823	1.13	0.8708
0.09	0.5359	0.44	0.6700	0.79	0.7853	1.14	0.8729
0.10	0.5398	0.45	0.6736	0.80	0.7881	1.15	0.8749
0.11	0.5438	0.46	0.6772	0.81	0.7910	1.16	0.8770
0.12	0.5478	0.47	0.6808	0.82	0.7939	1.17	0.8790
0.13	0.5517	0.48	0.6844	0.83	0.7967	1.18	0.8810
0.14	0.5557	0.49	0.6879	0.84	0.7995	1.19	0.8830
0.15	0.5596	0.50	0.6915	0.85	0.8023	1.20	0.8849
0.16	0.5636	0.51	0.6950	0.86	0.8051	1.21	0.8869
0.17	0.5675	0.52	0.6985	0.87	0.8078	1.22	0.8888
0.18	0.5714	0.53	0.7019	0.88	0.8106	1.23	0.8907
0.19	0.5753	0.54	0.7054	0.89	0.8133	1.24	0.8925
0.20	0.5793	0.55	0.7088	0.90	0.8159	1.25	0.8944
0.21	0.5832	0.56	0.7123	0.91	0.8186	1.26	0.8962
0.22	0.5871	0.57	0.7157	0.92	0.8212	1.27	0.8980
0.23	0.5910	0.58	0.7190	0.93	0.8238	1.28	0.8997
0.24	0.5948	0.59	0.7224	0.94	0.8264	1.29	0.9015
0.25	0.5987	0.60	0.7237	0.95	0.8289	1.30	0.9032
0.26	0.6026	0.61	0.7291	0.96	0.8315	1.31	0.9049
0.27	0.6064	0.62	0.7324	0.97	0.8340	1.32	0.9066
0.28	0.6193	0.63	0.7357	0.98	0.8365	1.33	0.9082
0.29	0.6141	0.64	0.7380	0.99	0.8389	1.34	0.9099
0.30	0.6179	0.65	0.7422	1.00	0.8413	1.35	0.9115
0.31	0.6217	0.66	0.7454	1.01	0.8438	1.36	0.9131
0.32	0.6256	0.67	0.7486	1.02	0.8461	1.37	0.9147
0.33	0.6293	0.68	0.7517	1.03	0.8485	1.38	0.9162
0.34	0.6331	0.69	0.7549	1.04	0.8508	1.39	0.9177

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1.40	0.9192	1.75	0.9599	2.20	0.9861	2.90	0.9981
1.41	0.9207	1.76	0.9608	2.22	0.9868	2.92	0.9982
1.42	0.9222	1.77	0.9616	2.24	0.9875	2.94	0.9984
1.43	0.9236	1.78	0.9625	2.26	0.9881	2.96	0.9985
1.44	0.9251	1.79	0.9633	2.28	0.9887	2.98	0.9986
1.45	0.9265	1.80	0.9641	2.30	0.9893	3.00	0.9989
1.46	0.9279	1.81	0.9649	2.32	0.9898	3.20	0.9993
1.47	0.9292	1.82	0.9656	2.34	0.0904	3.40	0.9996
1.48	0.9306	1.83	0.9664	2.36	0.9909	3.60	0.9998
1.49	0.9319	1.84	0.8671	2.38	0.9913	3.80	0.9999
1.50	0.9332	1.85	0.9678	2.40	0.9918		
1.51	0.9345	1.86	0.9686	2.42	0.9922		
1.52	0.9357	1.87	0.9693	2.44	0.9927		
1.53	0.9370	1.88	0.9699	2.46	0.9931		
1.54	0.9382	1.89	0.9706	2.48	0.9934		
1.55	0.9394	1.90	0.9713	2.50	0.9938		
1.56	0.9406	1.91	0.9719	2.52	0.9941		
1.57	0.9418	1.92	0.9726	2.54	0.9945		
1.58	0.9429	1.93	0.9732	2.56	0.9948		
1.59	0.9441	1.94	0.9738	2.58	0.9951		
1.60	0.9452	1.95	0.9744	2.60	0.9953		
1.61	0.9463	1.96	0.9750	2.62	0.9956		
1.62	0.9474	1.97	0.9756	2.64	0.9959		
1.63	0.9484	1.98	0.9761	2.66	0.9961		
1.64	0.9495	1.99	0.9767	2.68	0.9963		
1.65	0.9505	2.00	0.9772	2.70	0.9965		
1.66	0.9515	2.02	0.9783	2.72	0.9967		
1.67	0.9525	2.04	0.9793	2.74	0.9969		
1.68	0.9535	2.06	0.9803	2.76	0.9971		
1.69	0.9545	2.08	0.9812	2.78	0.9973		
1.70	0.9554	2.10	0.9821	2.80	0.9974		
1.71	0.9564	2.12	0.9830	2.82	0.9976		
1.72	0.9572	2.14	0.9838	2.84	0.9977		
1.73	0.9582	2.16	0.9846	2.86	0.9978		
1.74	0.9591	2.18	0.9854	2.88	0.9980		

Table T.5

The χ^2 distribution

n \ p	0.99	0.98	0.95	0.90	0.80	0.70	0.50
1	0.000	0.000	0.003	0.016	0.064	0.148	0.455
2	0.020	0.040	0.103	0.211	0.446	0.713	1.386
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357
5	0.554	0.752	1.145	1.010	2.343	3.000	4.351
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338
20	8.260	9.237	10.851	12.443	14.578	17.266	19.337
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336

0.30	0.20	0.10	0.05	0.02	0.01	0.001	p / n
1.074	1.642	2.706	3.841	5.412	6.635	10.827	1
2.408	3.219	4.605	5.991	7.824	9.210	13.815	2
3.665	4.642	6.251	7.815	9.837	11.345	16.268	3
4.878	5.989	7.779	9.488	11.668	13.277	18.465	4
6.064	7.289	9.236	11.070	13.388	15.086	20.517	5
7.231	8.558	10.645	12.592	15.033	16.812	22.457	6
8.383	9.803	12.017	14.067	16.622	18.475	24.322	7
9.524	11.030	13.362	15.507	18.168	20.090	26.125	8
10.656	12.242	14.684	16.919	19.679	21.666	27.877	9
11.781	13.442	15.987	18.307	21.161	23.209	29.588	10
12.899	14.631	17.275	19.675	22.618	24.725	31.264	11
14.011	15.812	18.549	21.026	24.054	26.217	32.909	12
15.119	16.985	19.812	22.362	25.472	27.688	34.528	13
16.222	18.151	21.064	23.685	26.873	29.141	36.123	14
17.322	19.311	22.307	24.996	28.259	30.578	37.697	15
18.418	20.465	23.542	26.296	29.633	32.000	39.252	16
19.511	21.615	24.769	27.587	30.995	33.409	40.790	17
20.601	22.760	25.989	28.869	32.346	34.805	42.312	18
21.689	23.900	27.204	30.144	33.687	36.191	42.820	19
22.775	25.038	28.412	31.410	35.020	37.566	45.315	20
23.858	26.171	29.615	32.671	36.343	38.932	46.797	21
24.939	27.301	30.813	33.924	37.659	40.289	48.268	22
26.018	28.429	32.007	35.172	38.968	41.638	49.728	23
27.096	29.553	33.196	36.415	40.270	42.980	51.179	24
28.172	30.675	34.382	37.652	41.566	44.314	52.620	25
29.246	31.795	35.563	38.885	42.856	45.642	54.052	26
30.319	32.912	36.741	40.113	44.140	46.963	55.476	27
31.391	34.027	37.916	41.337	45.419	48.278	56.793	28
32.461	35.139	39.087	42.557	46.693	49.588	58.302	29
33.530	36.250	40.256	43.773	47.962	50.892	59.703	30

Table T.6

The $K(z)$ function

z	$K(z)$	z	$K(z)$	z	$K(z)$
0.28	0.000001	0.71	0.305471	1.14	0.851394
0.29	0.000004	0.72	0.322265	1.15	0.858038
0.30	0.000009	0.73	0.339113	1.16	0.864142
0.31	0.000021	0.74	0.355981	1.17	0.870612
0.32	0.000046	0.75	0.372833	1.18	0.876548
0.33	0.000091	0.76	0.389640	1.19	0.882258
0.34	0.000171	0.77	0.406372	1.20	0.887750
0.35	0.000303	0.78	0.423002	1.21	0.893030
0.36	0.000511	0.79	0.439505	1.22	0.898104
0.37	0.000826	0.80	0.455857	1.23	0.902972
0.38	0.001285	0.81	0.472041	1.24	0.907648
0.39	0.001929	0.82	0.488030	1.25	0.912132
0.40	0.002808	0.83	0.503808	1.26	0.916432
0.41	0.003972	0.84	0.519366	1.27	0.920556
0.42	0.005476	0.85	0.534682	1.28	0.924505
0.43	0.007377	0.86	0.549744	1.29	0.928288
0.44	0.009730	0.87	0.564546	1.30	0.931908
0.45	0.012590	0.88	0.579070	1.31	0.935370
0.46	0.016005	0.89	0.593316	1.32	0.938682
0.47	0.020022	0.90	0.607270	1.33	0.941848
0.48	0.024683	0.91	0.620928	1.34	0.944872
0.49	0.030017	0.92	0.634286	1.35	0.947756
0.50	0.036055	0.93	0.647338	1.36	0.959512
0.51	0.042814	0.94	0.660082	1.37	0.953142
0.52	0.050306	0.95	0.672516	1.38	0.955650
0.53	0.058534	0.96	0.684636	1.39	0.958040
0.54	0.067497	0.97	0.696444	1.40	0.960348
0.55	0.077183	0.98	0.707940	1.41	0.962486
0.56	0.087577	0.99	0.719126	1.42	0.964552
0.57	0.098656	1.00	0.730000	1.43	0.966516
0.58	0.110395	1.01	0.740566	1.44	0.968382
0.59	0.122760	1.02	0.750826	1.45	0.970158
0.60	0.135718	1.03	0.760780	1.46	0.971846
0.61	0.149223	1.04	0.770434	1.47	0.973448
0.62	0.163225	1.05	0.779794	1.48	0.974970
0.63	0.177753	1.06	0.788860	1.49	0.976412
0.64	0.192677	1.07	0.797636	1.50	0.977782
0.65	0.207987	1.08	0.806128	1.51	0.979080
0.66	0.223637	1.09	0.814342	1.52	0.980310
0.67	0.239582	1.10	0.822282	1.53	0.981476
0.68	0.255780	1.11	0.829950	1.54	0.982578
0.69	0.272189	1.12	0.837356	1.55	0.983622
0.70	0.288765	1.13	0.844502	1.56	0.984610

z	$K(z)$	z	$K(z)$	z	$K(z)$
1.57	0.985544	1.93	0.998837	2.29	0.999944
1.58	0.986426	1.94	0.998924	2.30	0.999949
1.59	0.987260	1.95	0.999004	2.31	0.999954
1.60	0.988048	1.96	0.999079	2.32	0.999958
1.61	0.988791	1.97	0.999149	2.33	0.999962
1.62	0.989492	1.98	0.999123	2.34	0.999965
1.63	0.990154	1.99	0.999273	2.35	0.999968
1.64	0.990777	2.00	0.999329	2.36	0.999970
1.65	0.991364	2.01	0.999380	2.37	0.999973
1.66	0.991917	2.02	0.999428	2.38	0.999976
1.67	0.992438	2.03	0.999474	2.39	0.999978
1.68	0.992928	2.04	0.999516	2.40	0.999980
1.69	0.993389	2.05	0.999552	2.41	0.999982
1.70	0.993828	2.06	0.999588	2.42	0.999984
1.71	0.994230	2.07	0.999620	2.43	0.999986
1.72	0.994612	2.08	0.999650	2.44	0.999987
1.73	0.994972	2.09	0.999680	2.45	0.999988
1.74	0.995309	2.10	0.999705	2.46	0.999989
1.75	0.995625	2.11	0.999723	2.47	0.999990
1.76	0.995922	2.12	0.999750	2.48	0.999991
1.77	0.996200	2.13	0.999770	2.49	0.999992
1.78	0.996460	2.14	0.999790	2.50	0.9099925
1.79	0.996704	2.15	0.999806	2.55	0.9999956
1.80	0.996932	2.16	0.999822	2.60	0.9999974
1.81	0.997146	2.17	0.999838	2.65	0.9999984
1.82	0.997346	2.18	0.999852	2.70	0.9999990
1.83	0.997533	2.19	0.999864	2.75	0.9999994
1.84	0.997707	2.20	0.999874	2.80	0.9999997
1.85	0.997870	2.21	0.999886	2.85	0.99999982
1.86	0.998023	2.22	0.999896	2.90	0.99999990
1.87	0.998145	2.23	0.999904	2.95	0.99999994
1.88	0.998297	2.24	0.999912	3.00	0.99999997
1.89	0.998421	2.25	0.999920		
1.90	0.998536	2.26	0.999926		
1.91	0.998644	2.27	0.999934		
1.92	0.998744	2.28	0.999940		

Table T.7

Student distribution

n \ p	0.90	0.80	0.70	0.60	0.50	0.40	0.30
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076
15	0.128	0.258	0.393	0.536	0.691	0.866	1,074
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064
21	0.127	0.257	0.391	0.532	0.686	0.859	1.663
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058
27	0.127	0.256	0.309	0.531	0.684	0.855	1.057
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055
30	0.127	0.156	0.389	0.530	0.683	0.854	1.055
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050
60	0.126	0.254	0.387	0.527	0.679	0.848	1.046
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036

0.20	0.10	0.05	0.02	0.01	0.002	p / n
3.078	6.314	12.706	31.821	63.057	636.619	1
1.886	2.920	4.303	6.965	9.925	31.598	2
1.638	2.353	3.182	4.541	5.841	12.941	3
1.533	2.132	2.776	3.747	4.604	8.610	4
1.476	2.015	2.571	3.365	4.032	6.859	5
1.440	1.943	2.447	3.143	3.707	5.959	6
1.415	1.895	2.365	2.998	3.499	5.405	7
1.397	1.860	2.306	2.896	3.355	5.041	8
1.383	1.833	2.262	2.821	3.250	4.781	9
1.372	1.812	2.228	2.764	3.169	4.587	10
1.363	1.796	2.201	2.718	3.106	4.437	11
1.356	1.782	2.179	2.681	3.055	4.318	12
1.350	1.771	2.160	2.650	3.012	4.221	13
1.345	1.761	2.145	2.624	2.977	4.140	14
1.341	1.753	2.131	2.602	2.947	4.073	15
1.337	1.746	2.120	2.583	2.921	4.015	16
1.333	1.740	2.110	2.567	2.898	3.965	17
1.330	1.734	2.101	2.552	2.878	3.922	18
1.328	1.729	2.093	2.539	2.861	3.883	19
1.325	1.725	2.086	2.528	2.845	3.850	20
1.323	1.721	2.080	2.518	2.831	3.819	21
1.321	1.717	2.074	2.508	2.819	3.792	22
1.319	1.714	2.069	2.500	2.807	3.767	23
1.318	1.711	2.064	2.492	2.797	3.745	24
1.316	1,708	2.060	2.485	2.787	3.725	25
1.315	1.706	2.056	2.479	2.779	3.707	26
1.314	1.703	2.052	2.473	2.771	3.690	27
1.313	1.701	2.048	2.467	2.763	3.674	28
1.311	1.699	2.045	2.462	2.756	3.659	29
1.310	1.697	2.042	2.457	2.750	3.646	30
1.303	1.684	2.021	2.423	2.704	3:551	40
1.296	1.671	2.000	2.390	2.660	3.460	60
1.289	1.658	1.980	2.358	2.617	3.373	120
1.282	1.645	1.960	2.326	2.576	3.291	∞

Table T.8

Table for the Poisson distribution

k \ λ	1	2	3	4
0	0.36783	0.13534	0.04978	0.01831
1	0.36788	0.27067	0.14936	0.07326
2	0.18394	0.27067	0.22404	0.14653
3	0.06131	0.18045	0.22404	0.19537
4	0.01532	0.09022	0.16803	0.19537
5	0.00306	0.03609	0.10082	0.15629
6	0.00051	0.01203	0.05040	0.10420
7	0.00007	0.00343	0.02160	0.05954
8		0.00085	0.00810	0.02977
9		0.00019	0.00270	0.01322
10		0.00003	0.00081	0.00529
11			0.00022	0.00192
12			0.00005	0.00064
13			0.00001	0.00019
14				0.00005
15				0.00001
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				

k \ λ	5	10	15	16
0	0.00673	0.00004	0.00000	0.00000
1	0.03369	0.00045	0.00000	0.00000
2	0.08422	0.00227	0.00003	0.00001
3	0.14037	0.00756	0.00017	0.00007
4	0.17547	0.01891	0.00064	0.00030
5	0.17547	0.03783	0.00193	0.00098
6	0.14622	0.06305	0.00483	0.00262
7	0.10444	0.09007	0.01037	0.00599
8	0.06527	0.11260	0.01944	0.01198
9	0.03626	0.15511	0.03240	0.02131
10	0.01813	0.12511	0.04861	0.03409
11	0.00824	0.11374	0.06628	0.04959
12	0.00343	0.09478	0.08285	0.06612
13	0.00132	0.07290	0.09560	0.08138
14	0.00047	0.05207	0.10244	0.09301
15	0.00015	0.03471	0.10244	0.09921
16	0.00004	0.02169	0.09603	0.09921
17	0.00001	0.01276	0.08473	0.09338
18		0.00709	0.07061	0.08300
19		0.00373	0.05574	0.06989
20		0.00183	0.04181	0.05592
21		0.00088	0.02986	0.04260
22		0.00040	0.02036	0.03098
23		0.00017	0.01328	0.02155
24		0.00007	0.00830	0.01437
25		0.00002	0.00498	0.00919
26		0.00001	0.00287	0.00566
27			0.00159	0.00335
28			0.00085	0.00191
29			0.00044	0.00105
30			0.00022	0.00056
31			0.00010	0.00029
32			0.00005	0.00014
33			0.00002	0.00007
34			0.00001	0.00003

LITERATURE

A) Books

A.1 Chow, Ven Te: *Handbook of Applied Hydrology*. Mc Graw Hill, New York. 1964.

A.2 Cramer, H.: *Mathematical Methods of Statistics*. 8th printing, Princeton Univ. Press, 1958.

A.3 Cramer, H. and Leadbetter, M. R.: *Stationary and Related Stochastic Processes*. Wiley and Sons, New York. 1967.

A.4 Doob, J. L.: *Stochastic Processes*. Wiley and Sons, New York. 1953.

A.5 Ezekiel, M. and Rox, K. A.: *Methods of Correlation and Regression Analysis*. Wiley and Sons, New York. 1959.

A.6 Feller, W.: *An Introduction to Probability Theory and its Applications*. Wiley and Sons, New York. 1957.

A.7 Ferguson, T. S.: *Mathematical Statistics, A Decision Theoretic Approach*. Academic Press New York—London. 1967.

A.8 Fisz, M.: *Probability Theory and Mathematical Statistics*. 3rd ed. Wiley and Sons, New York. 1963.

A.9 Gnedenko, B. V.: *The Theory of Probability*. 4th ed. Chelsea, New York, 1968.

A.10 Hajek, J.: *Nonparametric Statistics*. Holden Day Inc, San Francisco. 1969.

A.11 Hall, M. J.: *Urban Hydrology*. Elsevier S. P, London—New York. 1984.

A.12 Kartvelisvili, N. A.: *Stochasticheskaya Gidrologiya*. Gidrometeoizdat, Leningrad. 1975.

A.13 Kendall, M.: *Rank Correlation Methods*. Griffin London—New York. 1955.

A.14 Kendall, M.G—Stuart, A.: *The Advanced Theory of Statistics, 1—3*. Griffin, London. 1966.

A.15 Kolmogorov, A. N.: *Foundations of the Theory of Probability*. 2nd ed. Chelsea New York, 1956.

A.16 Lehmann, E. L.: *Testing Statistical Hypothesis*. Wiley and Sons, New York. 1959.

A.17 Lehmann, E. L.: *Nonparametrics*. Holden Day Inc., San Francisco, 1975.

A.18 Loéve, M.: *Probability Theory*. D. Van Nostrand Company. Inc., Princeton. 1963.

A.19 Parzen, E.: *Modern Probability Theory and its Applications*. Wiley and Sons, New York. 1960.

A.20 Parzen, E.: *Stochastic Processes*. Holden Day Inc., San Francisco. 1962.

A.21 Reimann, J.—V. Nagy I.: *Hydrological Statistics* (in Hungarian). Tankönyvkiadó, Budapest. 1984.

A.22 Reimann, J.—Tóth, J.: *Probability Calculus and Mathematical Statistics*. (In Hungarian.) Tankönyvkiadó, Budapest. 1985.

A.23 Rényi, A.: *Probability Theory*. Akadémiai Kiadó, Budapest. 1970.

A.24 Shaw, E. M.: *Hydrology in Practice*. Van Nostrand Reinhold Co. Ltd., Berkshire, England. 1983.

A.25 Vincze, I.: *Mathematical Statistics with Applications in Industry* (in Hungarian). Műszaki Kiadó. 1968.

A.26 Wald, A.: *Statistical Decision Functions*. Wiley and Sons, New York. 1961.

A.27 Yevjevich, V.: *Probability and Statistics in Hydrology. Water Resources Publications*, Fort Collins, Colorado, USA. 1972.

A.28 Yevjevich, V.: *Stochastic Processes in Hydrology. Water Resources Publications*, Fort Collins, Colorado, USA. 1972.

B.1 Arnold, H. J.: "Small Sample Power for the One-Sample Wilcoxon-Test for Non-normal Shift Alternatives". *Ann. Math. Statist.* **27.** 1767—1778. 1965.

B.2 Bardzik, A.:"Hydrological Models of Flood-wave transformation in River-channels used for outflow prediction." *Proc. of the International Conference on "Hydrological Processes in the cathment"*. Cracow. 1986.

B.3 Bhattacharya, P. K.: "Estimation of probability density function and its derivates." *Sankhya* **29.** 373—382. 1967.

B.4 Billingsley, P.: "Statistical Methods in Markov Chains". *Ann. Math. Statist.* **32.** 12—40, 1961.

B.5 Blomqvist, N.: "On a Measure of Dependence between Two Random Variables". *Ann. Math. Statist.* **21.** 593—600. 1950.

B.6 Diaconis, P—Efron, B.: "Testing for independence in a two-way table. New interpretations of the chi-square statistics." *The Annals of Statistics.* 1985. Vol. 13. No. 3. 845—874.

B.7 Gnedenko, B. V. and Korolyuk, V. C.: "On the Maximal Deviation between Two Empirical Distributions". *Dokl. Akad. Nauk*, SSSR **80.** 525—528, 1951.

B.8 Gupta, V. K.—Duckstein, L.—Peebles, R. W.: "On the joint distribution of the largest flood and its time of occurrence." *Water Resour. Res.*, 12/2. 295—304, 1976.

B.9 Gumbel, E. J.: "Bivariate Exponential Distributions." *Amer. Stat. Association Journal*, December. 1960.

B.10 Haan, C. T—Johnson, H. P—Brakensiek, D. L.: "Hydrologic Modelling of Small watersheds". *ASAE Monograph.* **5.** 1982.

B.11 Hoeffding, W.: "A nonparametric test of independence." *Ann Math. Statist*, **19.** 546—557.

B.11 Karr, A.: "Two extreme value processes arising in hydrology." *J. Appl. Probab.* **13.** 190—194, 1976.

B.12 Kavvas, M. L.: "Stochastic Trigger Model for Flood Peaks".
1. Development of the Model. *Water Resour. Res.* 18(2). 383—398, 1982.
2. Application of the Model to the Flood Peaks of Goksu—Karahacili. *Water Resour. Res.* 18(2). 399—411, 1982.

B.13 Kiefer, J. and Wolfowitz, J.: "Optimum Designs in Regression Problems". *Ann Math. Statist.* **30.** 271—294. 1959.

B.14 Konijn, H. S.: "On the Power of Certain Tests for Independence in Bivariate Populations". *Ann. Math. Statists.* **27.** 300—323, 1956.

B.15 Konecny, F. and Nachtnebel, H. P.: "Extreme value Process and the Evaluation of Risk in Flood Analysis." *Arbeitsbericht*, Institut f. Wasserwirtschaft, Universität für Bodenkultur, Wien. 1983.

B.16 Kruskal, H. W.: "Ordinal Measures of Association." *Amer. Stat. Association Jornal*, 1958.

B.17 Kuczera, G.: "Robust Flood Frequency Models". *Water Resour. Res.* **18.** 2. 315—324, 1982.

B.18 Lancaster, H. O.: "Ordinal Measures of Association" *J. Ann. Statist. Assoc.* **53.**

B.19 Lehmann, E. L.: "The power of rank tests". *Ann. Math. Statist.* **24.** 23—42, 1953.

B.20 Lehmann, E. L.: "Some Concepts of Dependence". *Ann. Math. Statist.* **37.** 1137—1153. 1966.

B.21 Linfoot, E. H.: "An informational measure of Correlation". *Information and Control.* **1.** 85—89. 1957.

B.22 Lloyd, E. H.: "What is and what is not, a Markov Chain?" *Journal of Hydrology*, **22.** 1—28. 1974.

B.23 Morgenstern, D.: "Einfache Beispiele zweidimensionaler Verteilungen." *Mitt. Math. Statist.* **8.** 234—235. (1956)

B.24 Mosteller, F.: 'On some useful "inefficient" statistics'. Unpublished thesis. Princeton Univ. 1946.

B.25 Nadaraja, E. A.: "On nonparametric estimates of density functions and regression" *Teor. Verojatnost i Primenen* **10.** 199—203, 1965.

B.26 Parzen, E.: "On estimation of probability density function and mode". *Ann. Math. Statist.* **27.** 832—837. 1956.

B.27 Rao, U. V. R.—Savage, I. R. and Sobel, M.: "Contributions to the Theory of Rank Order Statistics; The Two-sample Censored Case". *Ann. Math. Statist.* **31.** 415—426. 1960.

B.28 Reimann, J. and Vincze, I.: "On the Comparison of Two Samples with Slightly Different Sizes". *Publ. Math. Inst. Hungar. Acad. Sci.* Vol. V. Ser. A. Fasc 3. 1960.

B.29 Reimann, J.: "Unsymmetrical Random Walk on the Plane and in the Space with Absorbing Barriers". *Acta Math. Sci. Hung.* Vol. XV. Fasc. 3—4. 1964.

B.30 Reimann, J.: "The Statistical Treatment of Flood Peaks". *UNESCO Technical Reports on Scientific and Practical Results of Selected IHD Projects.* Paris. 1974.

B.31 Reimann, J.: "Investigation of positively quadrant dependent bivariate distributions." *Periodica Polytechnica.* Vol. 32. Nos 1—2.

B.32 Révész, P.: "On Empirical Density Function". *Periodica Math. Hung.* Vol. 2. (1—4) 85—110. 1972.

B.33 Rosenblatt, M.: "Remarks on some non-parametric. estimates of density function". *Ann. Math. Statist.* **27.** 832—837. 1956.

B.34 Sarkadi, K.: "On Testing for Normality". *Publ. Math. Inst. Hungar. Acad. Sci.* Vol V. Ser A. Fasc. 3. 1960.

B.35 Spearman, C.: "The proof and measurement of association between two things". *Amer. Journal of Psychology*, **15.** (1904)

B.36 Taesombut, V. and Yevjevic, V.: "Use of Partial Flood Series of Estimating. Distribution of Maximum Annual Flood Peak". *Hydrology Papers*, Colorado State Univ. Fort Collins, Colorado, Okt. 1978.

B.37 Takács, L.: "On holding-time problems" (in Hungarian). *MTA III. Oszt. Közleményei*, VII. 3—4. 1957.

B.38 Todorovic, P.: "Stochastic Models of Floods." *Water Resour. Res.*, **14.** (2) 345—356, 1978.

B.39 Todorovic, P. and Woolhiser, D. A.: "On the time when the extreme flood occurs". *Water Resour. Res.* **7.** (5) 1144—1150. 1971.

B.40 Tusnády, G.: "On Testing Density Functions". *Periodica Math. Hung.* **5.** 161—169. 1974.

B.41 Wald, A.: "The fitting of straight lines if both variables are subject to error". *Ann. Math. Statist.* **11.** No. 3. 1940.

B.42 Wald, A. and Wolfowitz, J.: "On a Test Wether Two Samples Are from the Same Population". *Ann. Math. Statist.* **11.** 147—162. 1940.

B.43 Waylen, P. and Ming-Ko Woo: "Prediction of Annual Floods Generated by Mixed Processes". *Water Resour. Res.* 18 (4). 1283—1286, 1982.

B.44 Vincze, I.: "On some joint distributions and joint limiting distributions in the theory of order statistics". *Publ. Math. Inst. Hungar. Acad. Sci.* **4.** 29—47. 1959.

B.45 Vincze, J.: "On the Cramèr—Frèchet—Rao inequality in the nonregular case". *Contributions to Statistics.* Academia. Prague. 1979.

B.46 Vincze, J.: "Remark to the derivation of the Cramèr—Frèchet—Rao inequality in the regular case". *Lecture notes in mathematical statistics.* Springer-Verlag, Wien, 1986.

B.47 Wolfowitz, J.: "Asymptotic Distribution of Runs Up and Down". *Ann. Math. Statist.* **15.** 163—172. 1944.

B.48 Yanagimoto, T.: "On Measures of Association and Related Problems". *Ann. Inst. Statist. Math.* 1969.

B.49 Zelenhasic, E.: "Theoretical Probability Distributions for Flood Peaks". *Hydrology Papers.* **42.** Colorado State Univ. Fort Collins. 1970.

C) Tables

C.1 Bolshev, L. N.—Smirnov, N. W.: *Statistical Tables* (in Russian) Nauka. Moscow. 1969.

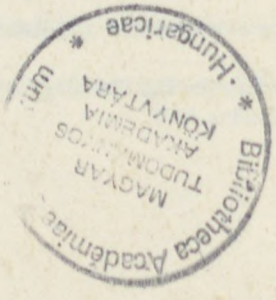
C.2 Finney, D. J.—Latscha, R.—Benneth, B. M.—Hsu, P.: *Tables for Testing Significance in a 2×2 Contingency Table.* Cambridge. Univ. Press. 1963.

C.3 Hald, A.: *Statistical Tables and Formulas,* Wiley and Sons. New York. 1960.

C.4 Owen, D. B.: *Tables for computing bivariate normal probabilities.* Ann. Math. Statist. 27. 1956.

C.5 Owen, D. B.: *Handbook of Statistical Tables.* Addison—Wesley. Reading Mass. 1962.

C.6 Pearson, E. S.—Hartley, H. O.: *Biometrical Tables for Statistician.* Cambridge University Press. 1962.



The statistical analysis of flood-waves will be efficient only if the hydrologist possesses a rather broad statistical knowledge and the methods of statistics are combined according to the nature of the problem at hand. This book tries to provide for this task as far as the limits of its reasonable extent will allow.

The basic objective of this book is to introduce the reader to the probabilistic and statistical model building techniques related to flood-problems (or other hydrologic problems).

In order to understand techniques presented in this book nothing beyond the knowledge of elementary calculus and combinatorial tools is assumed.

And now a few words about the structure of the book. The first few chapters summarize the bases of probability theory, the elements of Markov-chains and Markov processes, which are illustrated through a number of examples (chapters 1—3). Chapters 4—6 cover the basic methods of mathematical statistical analysis and decision theory, illustrated with hydrological problems. The methods are of general character and are applicable in many branches of engineering practice. Chapters 7 and 8 are devoted to the investigation of connections between random variables. This part of the book is the so called generalized correlation and regression theory which is a very useful tool in the engineering research and planning.

The book was written in the hope that the statistical techniques contained therein will help the hydrologist, the hydrologist student, or other civil engineers to get the most possible information from the results of observations for practical purposes.

ISBN 963 05 4832 1

J. Reimann

MATHEMATICAL STATISTICS