Probability Theory, Random Processes and Mathematical Statistics

Mathematics and Its Applications

Managing Editor:

M. HAZEWINKEL
Centre for Mathematics and Computer Science, Amsterdam, The Netherlands

Volume 344

Probability Theory, Random Processes and Mathematical Statistics

by

Yu. A. Rozanov
Steklov Mathematical Institute,
Moscow, Russia

SPRINGER SCIENCE+BUSINESS, MEDIA, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-94-010-4201-7 ISBN 978-94-011-0449-4 (eBook)
DOI 10.1007/978-94-011-0449-4

This is a completely revised and updated translation of the
original Russian work of the same title,
Moscow, Nauka, © 1980 (second edition).
Translated by the author.

Printed on acid-free paper

Originally published by Kluwer Academic Publishers in 1995
Softcover reprint of the hardcover 2nd edition 1995

CONTENTS

Preface

Probability Theory, Theory of Random Processes and Mathematical Statistics are important areas of modern mathematics and its applications. They develop rigorous models for a proper treatment for various 'random' phenomena which we encounter in the real world. They provide us with numerous tools for an analysis, prediction and, ultimately, control of random phenomena. Statistics itself helps with choice of a proper mathematical model (e.g., by estimation of unknown parameters) on the basis of statistical data collected by observations.
This volume is intended to be a concise textbook for a graduate level course, with carefully selected topics representing the most important areas of modern Probability, Random Processes and Statistics.

The first part (Ch. 1–3) can serve as a self-contained, elementary introduction to Probability, Random Processes and Statistics. It contains a number of relatively simple and typical examples of random phenomena which allow a natural introduction of general structures and methods. Only knowledge of elements of real/complex analysis, linear algebra and ordinary differential equations is required here.

The second part (Ch. 4–6) provides a foundation of Stochastic Analysis, gives information on basic models of random processes and tools to study them. Here a familiarity with elements of functional analysis is necessary. Our intention to make this course fast-moving made it necessary to present important material in a form of examples.

Yu. Rozanov

Annotation

The book consists of two parts which differ one from another in their contents and the style of exposition. The first one discusses many relatively simple problems which lead to different models of probability and random processes, as well as basic methods of mathematical statistics, including typical applications. The second part presents elements of general analysis of random processes.

CHAPTER 1

Introductory Probability Theory

1. The Notion of Probability

1.1. EQUIPROBABLE OUTCOMES

Imagine a usual coin tossing with two possible outcomes ω = 'head' or 'tail' each of them having the probability $1/2$. In another example of a dice tossing with six possible equiprobable outcomes $\omega = 1, \ldots, 6$; what is the probability of the event $\{\omega \text{ is even}\}$? The answer is of course $1/2$ (why?).

In a similar way, one can imagine a lot Ω of N different outcomes each having the same probability $\mathbf{P}(\omega) = 1/N$; what is the probability $\mathbf{P}(A)$ of the event $\{\omega \in A\}$ for a given subset $A \subseteq \Omega$? The natural answer is

$$\mathbf{P}(A) = \frac{N(A)}{N}, \tag{1.1}$$

where $N(A)$ is the number of elements in the set $A \subseteq \Omega$. Here, all possible events can be represented by corresponding *sets* $A \subseteq \Omega$, with the empty one $A = \varnothing$ as the *impossible event* of the probability $\mathbf{P}(A) = 0$, and $A = \Omega$ as the *certain event* of the probability $P(A) = 1$. Moreover, one can operate with events as we do with sets, when $A^c = \Omega \backslash A$ corresponds to the *complementary event* to $A \subseteq \Omega$, $A_1 \cup A_2$ corresponds to the union (sum) of events $A_1, A_2 \subseteq \Omega$ etc. In particular,

$$(A_1 \cup A_2)^c = A_1^c \cap A_2^c,$$

where on the right-hand side is the intersection of the complementary events A_1^c, A_2^c to A_1, A_2. (Sometimes the intersection $A_1 \cap A_2$ of any $A_1, A_2 \subseteq \Omega$ is also called the *product* of the events A_1, A_2 and denoted by $A_1 \cdot A_2$.)

According to (1.1), for any *disjoint* events $A_1, \ldots, A_n \subseteq \Omega$ the probability of their union

$$A = \bigcup_{k=1}^{n} A_k$$

is

$$\mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \mathbf{P}(A_k), \tag{1.2}$$

which follows of course from the relation

$$N\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} N(A_k).$$

In spite of the simplicity of the above scheme, sometimes it might be difficult to find out which outcomes ω can be considered *equiprobable* (i.e. having the same probability $P(\omega) = 1/N$). For example, in a simultaneous tossing of two coins one deals with the following outcomes: 'two heads', 'two tails', 'head and tail.' Are these $N = 3$ outcomes equiprobable? Or should we consider the following $N = 4$ outcomes:

$$\Omega = \{\text{'head-head'}, \text{'head-tail'}, \text{'tail-head'}, \text{'tail-tail'}\},$$

each of them having the probability $\mathbf{P}(\omega) = 1/4$?

1.2. EXAMPLES

Random sampling. Suppose we *randomly* choose l objects from a lot containing n objects of which m are 'defective'. The sample contains a random number $\xi = 0, 1, \ldots, \min(l, m)$ of defective objects. Any of

$$N = \binom{n}{l} = \frac{n!}{l!(n-l)!}$$

possible samples is equiprobable. What is the probability to choose a sample with $\xi = k$ defective objects? Let $A = \{\xi = k\}$ be the corresponding event. Then

$$N\{\xi = k\} = \binom{m}{k} \cdot \binom{n-m}{l-k}$$

and according to (1.1)

$$\mathbf{P}\{\xi = k\} = \frac{\binom{m}{k}\binom{n-m}{l-k}}{\binom{n}{l}}, \quad k = 0, 1, \ldots, \min(l, m). \tag{1.3}$$

The system of probabilities (1.3) is called the *hypergeometric distribution.*

Random allocations. Consider random placement of n 'particles' into r 'cells' yielding all possible distributions $(n_1, \ldots, n_r)$ of particles (n_1 is the number of particles in the first cell, n_2 in the second cell, etc.). The total number N of all allocations is the number of all $n_1, \ldots, n_r$ such that $n_1 + \cdots + n_r = n$ which is the coefficient $(1/n!)\, f^{(n)}(0)$ in the well-known Taylor series

$$f(x) \equiv \left(\sum_{n_1=0}^{\infty} x^{n_1} \right) \cdots \left(\sum_{n_r=0}^{\infty} x^{n_r} \right) = (1-x)^{-r} = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(0) x^n, \quad |x| < 1.$$

Therefore

$$N = \frac{1}{n!} \left[\frac{d^n}{dx^n} (1-x)^{-r} \right]_{x=0} = \frac{1}{n!} r(r+1) \cdots (r+n-1) = \binom{n+r-1}{n}.$$

Assuming that all allocations are equiprobable, we find the corresponding probability

$$\mathbf{P}(n_1, \ldots, n_r) = \binom{n+r-1}{n}^{-1}. \tag{1.4}$$

Consider now a different situation when the random placement $(n_1, \ldots, n_r)$ is subject to the condition $n_k = 0$ or 1, $k = 1, \ldots, n$ (this implies, of course, $n \leqslant r$). The total number N of all such allocations is obviously $N = \binom{r}{n}$ so, assuming any $(n_1, \ldots, n_r)$ equiprobable, we obtain for the corresponding probability

$$\mathbf{P}(n_1, \ldots, n_r) = \binom{r}{n}^{-1}. \tag{1.5}$$

Note that we did not distinguish between particles at the derivation of the probability distributions (1.4), (1.5) of random allocation.*

Suppose that we are dealing now with n *different* particles which are placed randomly into r cells in such a way that any allocation $(i_1, \ldots, i_n)$ is equiprobable, where $1 \leqslant i_k \leqslant r$ is the cell number of the k-th particle, $k = 1, \ldots, n$. It is clear that the total number of all such allocations is $N = r^r$ so that the corresponding probability is

$$\mathbf{P}(i_1, \ldots, i_r) = r^{-n}.$$

* This assumption is satisfied by certain 'elementary particles' considered in Quantum Physics; see, e.g., W. Feller: An Introduction to Probability Theory and Its Applications, vol. I, Wiley, New York etc. 1968.

The event $A = (n_1, \ldots, n_r)$, considered in (1.4), (1.5) in the case of indistinguishable particles, occurs for

$$N(n_1, \ldots, n_r) = \frac{n!}{n_1! \cdots n_r!}$$

different allocations $(i_1, \ldots, i_n)$ satisfying r conditions

$$\sum_{i_k=1} 1 = n_1, \quad \ldots, \quad \sum_{i_k=r} 1 = n_r.$$

According to (1.1), the probability of this event is

$$\mathbf{P}(n_1, \ldots, n_r) = \frac{n!}{n_1! \cdots n_r!} r^{-n}. \tag{1.6}$$

1.3. CONDITIONAL PROBABILITY

Let us return to the general scheme with a lot Ω of N equiprobable outcomes $\omega \subseteq \Omega$ and the probability of any event $A \subseteq \Omega$ determined by (1.1). Suppose we know that some event B does occur; what is the probability of A in the new situation? The corresponding probability is called the conditional probability of A given B, and is usually denoted by $P(A \mid B)$. It is clear that if B occurs, the outcomes $\omega \subseteq A$ are necessarily in B and among them only $\omega \in AB$ are in favour of A. Hence, assuming that all $\omega \in B$ are *equiprobable*, we get that

$$\mathbf{P}(A \mid B) = \frac{N(AB)}{N(B)} = \frac{N(AB)}{N} \Big/ \frac{N(B)}{N},$$

where $N(AB)$, $N(B)$ is the number of outcomes in AB, B, respectively. Hence the *conditional probability* of A given B can be defined by

$$\mathbf{P}(A \mid B) = \frac{P(AB)}{P(B)}. \tag{1.7}$$

For example, one can easily observe that in the random allocation scheme (1.4), the conditional probability of $A = (n_1, \ldots, n_r)$ given $B = \{n_k = 0 \text{ or } 1,\ k = 1, \ldots, r\}$ equals to

$$\mathbf{P}(A \mid B) = \binom{r+n-1}{n}^{-1},$$

which is the same as the probability (1.5).

Consider again the general equiprobable scheme (1.1), and suppose that $B_1, \ldots, B_n$ are *disjoint* (i.e. mutually exclusive) events whose union

$$\bigcup_k B_k = \Omega$$

is the certain event Ω. Then, for any event $A \subseteq \Omega$, the *total probability formula*

$$\mathbf{P}(A) = \sum_{k=1}^{n} \mathbf{P}(A \mid B_k) P(B_k) \tag{1.8}$$

holds.

According to (1.1), the above formula follows from the relation

$$N(A) = \sum_{k=1}^{n} N(AB_k).$$

Obviously (1.8) holds as well for any *disjoint* $B_1, \ldots, B_n$ and any A such that

$$A \subseteq \bigcup_{k=1}^{n} B_k.$$

EXAMPLE (*The best choice problem*). Imagine a fastidious bride who is to select the best among n candidates upon seeing them successively, under the condition that a rejected one is lost forever. Of course, it would be unwise to marry the first candidate if the number n is large. On the other hand, if she refuses too many of them, she might lose the best. How can she make the best choice? It is assumed that the bride is capable of ranking the candidates by assessing the 'quality' of every of them by a real number ξ_k, i.e., ξ_k is the 'quality' of the k-th successive candidate, $k = 1, \ldots, n$.

A sensible strategy for the bride is the following. First she decides on some m ($1 \leqslant m \leqslant n$) and then chooses the first among the last $n-m$ candidates who is better than the previous m ones, i.e. her choice is the smallest $\tau = m+1, \ldots, n$, such that $\xi_\tau > \max(\xi_1, \ldots, \xi_m)$. (Of course, there is a chance to lose all the candidates if none of the last $n-m$ ones is better than the first m.) The strategy depends on m, and the corresponding probability to choose the best possible candidate

$$p_m = \mathbf{P}\{\xi_\tau = \max(\xi_1, \ldots, \xi_n)\}$$

can be maximized by an appropriate choice of $m = m_n$, i.e.

$$p_{m_n} = \max_{1 \leqslant m \leqslant n} p_m.$$

Let us find the probability p_m in the case of *equiprobable* orderings $\xi_{i_1} < \cdots < \xi_{i_n}$. It is clear that a successive ξ_{l+1} can be anywhere between the previous $\xi_1, \ldots, \xi_l$ as well as either larger or smaller than all of them, which results in $l+1$ different orderings, so that the total number of all possible outcomes (orderings) $\xi_{i_1} < \cdots < \xi_{i_n}$ is

$$N = 1 \cdots l(l+1) \cdots n = n!.$$

Consider the probability $\mathbf{P}(B_k)$ of $B_k = \{\tau = k\}$, $k = m+1, \ldots, n$. Of course, under the event B_k, the first m points can be ordered arbitrarily, with the number of successive locations of ξ_{l+1} with respect to the previous $\xi_1, \ldots, \xi_l$, $l = 1, \ldots, m-1$, explained above. The successive ξ_{m+l}, $l = 1, \ldots, k-m-1$, can be correspondingly put each into $m+l-1$ intervals only as $\xi_{m+l} < \max(\xi_1, \ldots, \xi_m)$, while for $\xi_k = \max(\xi_1, \ldots, \xi_k)$ there is only one possibility of choice of the interval. Finally, the successive ξ_{k+l}, $l = 1, \ldots, n-k$, can be in any of the $k+l$ positions with respect to $\xi_1, \ldots, \xi_{k+l-1}$. Hence, the total number of outcomes favourable to B_k is

$$N(B_k) = 1 \cdots m \cdot m \cdots (k-2)(k+1) \cdots n = n! \frac{m}{(k-1)k};$$

so that

$$\mathbf{P}(B_k) = \frac{N(B_k)}{N} = \frac{m}{(k-1)k}, \quad k = m+1, \ldots, n.$$

Let A be the event that one has made the best choice, then

$$p_m = \mathbf{P}(A) = \sum_{k=m+1}^{n} \mathbf{P}(A \mid B_k)\mathbf{P}(B_k),$$

where

$$\mathbf{P}(A \mid B_k) = \frac{N(AB_k)}{N(B_k)} = \frac{k}{n},$$

as the event AB_k differs from B_k in the way that the point ξ_k is the very right so that there are only $k + l - 1$ choices for location of ξ_{k+l}, $l = 1, \ldots, n - k$, and consequently

$$N(AB_k) = 1 \cdots m \cdot m \cdots (k-2)k \cdots (n-1) = n! \frac{m}{(k-1)n}.$$

Thus,

$$p_m = \frac{m}{n} \sum_{k=m}^{n-1} \frac{1}{k}.$$

It is easy to see that

$$\frac{m}{n} \sum_{k=m+1}^{n} \frac{1}{k} = \frac{m}{n} \sum_{k=m+1}^{n} \frac{\frac{1}{n}}{\frac{k}{n}} \leqslant \frac{m}{n} \int_{\frac{m}{n}}^{1} \frac{\mathrm{d}x}{x} = -\frac{m}{n} \log \frac{m}{n}$$

$$\leqslant \frac{m}{n} \sum_{k=m}^{n-1} \frac{\frac{1}{n}}{\frac{k}{n}} = \frac{m}{n} \sum_{k=m}^{n-1} \frac{1}{k}$$

and for large n ($n \to \infty$), since $1/\mathrm{e}$ is the maximum point of $-x \log x$, $0 < x < 1$, we obtain for the optimal quantities

$$m_n \sim \frac{n}{\mathrm{e}}, \qquad p_{m_n} \sim \frac{1}{\mathrm{e}} \quad (\mathrm{e} = 2.718\ldots).$$

1.4. INDEPENDENT EVENTS

Given two events A_1, A_2, it is natural to think of A_1 as being *independent* of A_2 if the occurrence of A_2 has no effect on the occurrence of A_1, i.e., if the corresponding conditional probability satisfies

$$\mathbf{P}(A_1 \mid A_2) = \mathbf{P}(A_1).$$

In view of (1.7), the above independence of A_1 from A_2 is equivalent to

$$\mathbf{P}(A_1 A_2) = \mathbf{P}(A_1) \cdot \mathbf{P}(A_2). \tag{1.9}$$

EXAMPLE. Let A_1 be the event that a card randomly picked from a deck is a spade, and A_2 the event that the card is a queen. Are these two events independent? The question is not easily answered by the intuition alone. Using the formal definition (1.9) of independence, in the case of a full deck (52 cards) with 13 spades and 4 queens we conclude that A_1 and A_2 are *independent*, as the probability of the event $A_1 \cdot A_2$ to pick up the queen of spades is

$$\mathbf{P}(A_1 A_2) = \frac{1}{52} = \frac{1}{4} \cdot \frac{1}{13} = \mathbf{P}(A_1) \cdot \mathbf{P}(A_2).$$

However the situation is quite different in the case of a deck which contains some blank cards in addition; in particular, for a large number $n \to \infty$ of blank cards, we obviously have $\mathbf{P}(A_1) \to 0$, while $\mathbf{P}(A_1 \mid A_2) \equiv 1/4$, independently of n.

Independent trials. Consider two experiments, e.g., throwing a coin and a dice. Usually we consider them as *independent*, and this intuitive feeling is very much consistent with the general formal definition (1.9). Indeed, consider the general case of two independent experiments with the corresponding outcomes $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$ of the total numbers N_1 and N_2, respectively, assuming that any joint outcome $\omega = (\omega_1, \omega_2) \in \Omega$ is *equiprobable*, where Ω is the direct product

$$\Omega = \Omega_1 \times \Omega_2,$$

and the total number of outcomes $\omega = (\omega_1, \omega_2) \in \Omega$ is $N = N_1 \cdot N_2$. Then, any two events $A_1 \subseteq R_1$ and $A_2 \subseteq \Omega_2$ are independent in the sense of (1.9). More precisely, the event $A_1 \cdot A_2$ corresponds to the direct product $A_1 \times A_2$ having $N(A_1 \times A_2) = N_1(A_1) \cdot N_2(A_2)$ of outcomes $\omega = (\omega_1, \omega_2)$, where $N_1(A_1)$, $N_2(A_2)$ is the number of outcomes in A_1, A_2, respectively. Therefore

$$\mathbf{P}(A_1 A_2) = \frac{N(A_1 A_2)}{N} = \frac{N_1(A_1) N_2(A_2)}{N_1 N_2} = \mathbf{P}(A_1)\mathbf{P}(A_2).$$

Of course, one can encounter a similar situation with several independent experiments (trials) $\Omega_1, \ldots, \Omega_n$, when the whole thing can be described by the direct product

$$\Omega = \Omega_1 \times \cdots \times \Omega_n \tag{1.10}$$

with *equiprobable* outcomes $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ in the corresponding trials, $k = 1, \ldots, n$. Here,

$$\mathbf{P}(\omega) = \mathbf{P}(\omega_1) \cdots \mathbf{P}(\omega_n), \tag{1.11}$$

and for any events $A_k \subseteq \Omega_k$ $(k = 1, \ldots, n)$ in *different* trials one has

$$\mathbf{P}(A_{i_1} \cdots A_{i_m}) = \mathbf{P}(A_{i_1}) \cdots \mathbf{P}(A_{i_m}), \tag{1.12}$$

for any disjoint indices $i_1, \ldots, i_m = 1, \ldots, n$ (check it!).

Relations (1.11) and (1.12) reflect mutual independence of the experiments $\Omega_1, \ldots, \Omega_n$; moreover, they are accepted in Probability Theory as a formal definition of *independent trials* (*events*). The opposite case of highly dependent trials can be illustrated by formally taking all Ω_k to be the same, in which case the outcomes $\omega_1 = \omega_2 = \cdots = \omega_n$ are the most dependent.

EXAMPLE. Consider random allocation of n different particles, where the k-th particle is placed into any of r cells, and any outcome $\omega_k = i_k$ ($i_k = 1, 2, \ldots, r$ is the cell number) is equiprobable. Assume that the corresponding trials (allocations) Ω_k, $k = 1, \ldots, n$, are independent, i.e., the particles behave independently of each other; then their distribution law is given by (1.6).

1.5. PROBABILITY AND FREQUENCY

Consider a sequence Ω_k, $k = 1, \ldots, n$, of *independent* trials which are of a similar nature, and an event $A = A_k$ associated with Ω_k (e.g., A is the occurrence of 'head' in coin tossing). Consider the *frequency* $n(A)/n$ of the event A, where $n(A)$ is the number of trials in which A occurred. For large n $(n \to \infty)$ one can observe the remarkable phenomenon of the near coincidence

$$\frac{n(A)}{n} \approx \mathbf{P}(A), \tag{1.13}$$

Table I. Number of occurencies of 'heads' in a series of 100 experiments of 100 coin tossings

Trial numbers	Number of heads										Total
0 – 1,000	54	46	53	55	46	54	41	48	51	53	501
– 2,000	48	46	40	53	49	49	48	54	53	45	485
– 3,000	43	52	58	51	51	50	52	50	53	49	509
– 4,000	58	60	54	55	50	48	47	57	52	55	536
– 5,000	48	51	51	49	44	52	50	46	53	41	485
– 6,000	49	50	45	52	52	48	47	47	47	51	488
– 7,000	45	47	41	51	49	59	50	55	53	50	500
– 8,000	53	52	46	52	44	51	48	51	46	54	497
– 9,000	45	47	46	52	47	48	59	57	45	48	494
– 10,000	47	41	51	48	59	51	52	55	39	41	484

which reflects the famous *Law of Large Numbers* of Probability Theory (it will appear in a general form in Sect. 8 of this chapter). As an illustration of (1.13), we present Table I containing the number of occurrences of 'heads' in a series of 100 experiments each corresponding to a sequence of 100 coin tossings.*

2. Some Probability Models

2.1. TRIALS WITH COUNTABLE OUTCOMES

Consider, for example, a sequence of coin tossing up to the first moment n when 'head' appears ($n = 1, 2, \ldots$). Here, every outcome is a sequence of $(n-1)$ 'tails' and 'head' at the end, which can be denoted as $\omega = n$ ($n = 1, 2, \ldots$). Clearly, the probability of $\{\omega = n\}$ is

$$\mathbf{P}\{\omega = n\} = \left(\frac{1}{2}\right)^{n-1} \cdot \frac{1}{2} = 2^{-n}; \quad n = 1, 2, \ldots .$$

The last formula formally holds for the outcome $\{\omega = \infty\}$, too, as its probability is zero. For example, the probability of $\omega = n$ being even is

$$\mathbf{P}\{\omega = 2k;\ k = 1, 2, \ldots\} = \sum_{k=1}^{\infty} 2^{-2k} = \frac{1}{4}\left(1 - \frac{1}{4}\right)^{-1} = \frac{1}{3}.$$

In the general trial with a *countable* number of outcomes $\omega \in \Omega$ with prescribed probabilities $\mathbf{P}(\omega) \geqslant 0$,

$$\sum_{\omega \in \Omega} \mathbf{P}(\omega) = 1, \tag{2.1}$$

the probability of any event $A \subseteq \Omega$ is defined by

$$\mathbf{P}(A) = \sum_{\omega \in A} \mathbf{P}(\omega). \tag{2.2}$$

Let several trials $\Omega_1, \ldots, \Omega_n$ be given, which can jointly be described by outcomes $\omega = (\omega_1, \ldots, \omega_n)$, $\omega_k \in \Omega_k$ $(k = 1, \ldots, n)$, as the direct product

$$\Omega = \Omega_1 \times \cdots \times \Omega_n. \tag{2.3}$$

* See the reference on p. 3.

Assume that the trials Ω_k, $k = 1, \ldots, n$, are *independent*. Then the probability of every outcome $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ is defined as

$$\mathbf{P}(\omega) = \mathbf{P}(\omega_1) \ldots \mathbf{P}(\omega_n), \tag{2.4}$$

summing up in total to

$$\sum_{\omega \in \Omega} \mathbf{P}(\omega) = \left[\sum_{\omega_1 \in \Omega_1} \mathbf{P}(\omega_1) \right] \cdots \left[\sum_{\omega_n \in \Omega_n} \mathbf{P}(\omega_n) \right] = 1.$$

Here, any events $A_k \subseteq \Omega_k$, $k = 1, \ldots, n$, belonging to different trials, are *independent* in the sense of (1.11). This can easily be verified by formally representing A_k in the product trial $\Omega = \Omega_1 \times \cdots \times \Omega_n$ by the corresponding direct product of $A_k \subseteq \Omega_k$ and the rest of Ω_j, $j \neq k$. (For example, A_1 can be represented by $A_1 \times \Omega_2 \times \cdots \times \Omega_n$.) Thus, the product $A_1 \ldots A_n$ can be represented by the direct product $A_1 \times \cdots \times A_n \subset \Omega$, and the general argument of (2.1)–(2.4) applies.

2.2. BERNOULLI TRIALS

Consider an event $A = A_k \subseteq \Omega_k$, where Ω_k, $k = 1, \ldots, n$, are *independent* trials of a similar nature (for example, A is the occurrence of 'head' in a series of n coin tossing). Put $\omega_k = 1$ if the event $A = A_k$ occurs, $\omega_k = 0$ otherwise. Then, as for as we are interested in the event A only, we can take $\Omega_k = \{1, 0\}$ as the two-point set, $k = 1, \ldots, n$. The direct product

$$\Omega = \{1, 0\}^n,$$

consisting of all $\{1, 0\}$-sequences $\omega = (\omega_1, \ldots, \omega_n)$, represents all possible outcomes $\omega \in \Omega$ of interest. Let the probability of A be the same for each $k = 1, \ldots, n$:

$$\mathbf{P}(A) = p;$$

what is the probability that A occurs m times in the trial series? According to the general model (2.3), (2.4), we can define it as

$$\mathbf{P}(m) = \binom{n}{m} p^m q^{n-m}, \quad q = 1 - p, \quad m = 0, 1, \ldots, n, \tag{2.5}$$

where $\binom{n}{m}$ is the number of all outcomes $\omega = (\omega_1, \ldots, \omega_n)$ with

$$\sum_{k=1}^{n} \omega_k = m,$$

as any such outcome has the same probability $\mathbf{P}(\omega) = p^m(1-p)^{n-m}$.

Formula (2.5) gives the *Bernoulli* (or *binomial*) *probability distribution*, with

$$\sum_m \mathbf{P}(m) = \sum_{m=0}^{n} \binom{n}{m} p^m q^{n-m} = (p+q)^n = 1.$$

The sum

$$\sum_m m\mathbf{P}(m) = \sum_{m=1}^{n} m \frac{n!}{m!(n-m)!} p^m q^{n-m} = np \sum_{m=0}^{n-1} \binom{n-1}{m} p^m q^{n-1-m} = np$$

is called the *mean value* of the probability distribution $\mathbf{P}(m)$, $m = 0, 1, \ldots$.

Let us introduce the *Poisson probability distribution*

$$\mathbf{P}(m) = \frac{a^m}{m!} \mathrm{e}^{-a}, \quad m = 0, 1, \ldots, \tag{2.6}$$

$$\sum_m \mathbf{P}(m) = 1,$$

where $a > 0$ is a parameter which coincides with the corresponding *mean value*

$$\sum_m m\mathbf{P}(m) = \sum_{m=1}^{\infty} m \frac{a^m}{m!} \mathrm{e}^{-a} = a\mathrm{e}^{-a} \sum_{m=0}^{\infty} \frac{a^m}{m!} = a.$$

2.3. LIMIT POISSON DISTRIBUTION

Consider a large series of the Bernoulli trials with a small probability $p = \mathbf{P}(A)$ of the occurrence of the event A. We ask how the corresponding probabilities will behave when $n \to \infty$, $p \to 0$ and the mean value

$$np = a$$

remains constant.

To answer this question, introduce the *generating function* $f(z)$ of a probability distribution $\mathbf{P}(m)$, $m = 0, 1, \ldots$, given by the power series

$$f(z) = \sum_m \mathbf{P}(m) z^m, \quad |z| \leqslant 1,$$

of the complex variable $z = re^{iu}$, $r \leqslant 1$, $-\pi \leqslant u \leqslant \pi$, $i = \sqrt{-1}$. The coefficients of the power series are given by the well-known formula

$$\begin{aligned} \mathbf{P}(m) &= \frac{1}{m!} f^{(m)}(0) = \frac{1}{2\pi} \int_{|z|=1} \frac{f(z)}{z^m} \, dz \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-imu} f(e^{iu}) \, du, \quad m = 0, 1, \ldots . \end{aligned}$$

For the Bernoulli distribution

$$\mathbf{P}_n(m) = \binom{n}{m} p^m q^{n-m}, \quad q = 1 - p, \quad m = 0, \ldots, n,$$

with the corresponding n, see (2.5), the generating function is

$$\begin{aligned} f_n(z) = \sum_m \mathbf{P}_n(m) z^m &= \sum_{m=0}^{n} \binom{n}{m} (zp)^m (1-p)^{n-m} \\ &= \left[1 - p(1-z)\right]^n = \left[1 - \frac{a(1-z)}{n}\right]^n, \quad a = np. \end{aligned}$$

Obviously for $|z| \leqslant 1$ as $n \to \infty$ we have

$$\left| \ln f_n(z) + a(1-z) \right| = n \left| \ln \left[1 - \frac{a(1-z)}{n}\right] + \frac{a(1-z)}{n} \right| \leqslant \frac{C}{n} \longrightarrow 0,$$

which shows that

$$f_n(z) \longrightarrow e^{-a(1-z)} = f(z)$$

uniformly in $|z| \leqslant 1$.

The limit function

$$f(z) = \mathrm{e}^{-a}\mathrm{e}^{az} = \sum_{m=0}^{\infty} \left(\frac{a^m}{m!} \mathrm{e}^{-a} \right) z^m, \quad |z| \leqslant 1,$$

turns out to be the generating function of the Poisson distribution with the mean value a. As a result, we have the convergence of the corresponding coefficients

$$\begin{aligned} \mathbf{P}_n(m) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{-imu} f_n(\mathrm{e}^{iu})\, \mathrm{d}u \\ &\longrightarrow \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{-imu} f(\mathrm{e}^{iu})\, \mathrm{d}u = \mathbf{P}(m), \quad m = 0, 1, \ldots, \end{aligned}$$

which gives the following *Poisson approximation of the Bernoulli distribution*:

$$\mathbf{P}_n(m) = \binom{n}{m} p^m q^{n-m} \sim \mathbf{P}(m) = \frac{a^m}{m!} \mathrm{e}^{-a}, \quad m = 0, 1, \ldots . \tag{2.7}$$

EXAMPLE (*The raisin roll problem*). Suppose N raisin rolls of equal size are baked from a batch of dough into which n raisins have been carefully mixed before. Then, clearly the number of raisins will vary from roll to roll, although the average number of raisins per roll is just $a = n/N$. What is the probability that a given roll contains at least one raisin?

It is natural to assume that the volume of the raisins is much smaller than that occupied by the dough, and the raisins move around freely and virtually independently during the mixing, hence whether or not a given raisin ends up in chosen roll does not depend on what happens to other raisins. Clearly, after careful mixing, raisins will be approximately uniformly distributed throughout the dough, i.e., each raisin has the probability

$$p = \frac{1}{N}$$

of ending up in a given roll. Then, we can interpret the problem in terms of a series of n Bernoulli trials, where 'success' in the k-th trial means that the k-th raisin ends up in a chosen roll. Suppose, both the number N of rolls and the number n of raisins are large, so that, in particular, $p = 1/N$ is small. Then the number of 'successes' in the n trials, or the number of raisins in a given roll, is approximately Poisson distributed, i.e., the probability $\mathbf{P}(m)$ of finding m raisins in the roll is

$$\mathbf{P}(m) \approx \frac{a^m}{m!} \mathrm{e}^{-a},$$

where

$$a = np = \frac{n}{N}.$$

Therefore, the probability $\mathbf{P}$ of finding at least one raisin is

$$\mathbf{P} = 1 - \mathbf{P}(0) \approx 1 - \mathrm{e}^{-a}.$$

EXAMPLE (*Radioactive decay*). It is experimentally observed that radium gradually decays into radon, by emitting alpha particles (helium nuclei). The interatomic distances are large enough to justify the assumption that each radium atom disintegrates independently of others. Moreover, each of the n initially present radium atoms has the same small probability p of disintegration during a time unit interval. (For instance, one gram of radium containing $n \approx 10^{22}$ atoms emits about 10^{10} alpha particles per second; hence the corresponding $p \approx 10^{10}/10^{22} = 10^{-12}$.) Call the disintegration of a radium atom a 'success'. Then the number of emitted alpha particles equals the number of 'successes' in a series of n Bernoulli trials with the 'success' probability p. The values of n and p being such, we have a very accurate agreement with a Poisson distribution, i.e., the probability that exactly m alpha particles are emitted during the time interval is given by

$$\mathbf{P}\{m\} = \frac{a^m}{m!}\,\mathrm{e}^{-a}, \quad k = 0, 1, 2, \ldots,$$

where $a = np$ is the average number of emitted alpha particles.

2.4. FINITE NUMBER OF EVENTS

Consider events A_k, $k = 1, \ldots, n$, whose possible outcomes can be jointly described by $\omega = (\omega_1, \ldots, \omega_n)$, with $\omega_k = 1$ or 0 depending on whether A_k or the complementary event A_k^c occur, respectively ($k = 1, \ldots, n$). Such outcomes $\omega = (\omega_1, \ldots, \omega_n)$ form the direct product

$$\Omega = \{1, 0\}^n.$$

Note that, according to the general model (2.1), (2.2), probability $\mathbf{P}(A)$ of an *arbitrary* event $A \subseteq \Omega$ is determined by the probabilities

$$\mathbf{P}\big(A_{i_1} \ldots A_{i_m}\big) = \mathbf{P}\big\{\omega_{i_1} = 1, \ \ldots, \ \omega_{i_m} = 1\big\} \tag{2.8}$$

for any $m = 1, \ldots, n$ and any $i_1, \ldots, i_m = 1, \ldots, n$. Indeed, by the general formula (2.2), for any events $B \subseteq A$ we have the equality

$$\mathbf{P}(A) = \mathbf{P}(B) + \mathbf{P}(A \backslash B),$$

where $A \backslash B = AB^c$. Hence, first we can find probabilities

$$\begin{aligned}&\mathbf{P}\{\omega_{i_1} = 1, \ldots, \omega_{i_m} = 1,\ \omega_{i_{m+1}} = 0\}\\ &\quad = \mathbf{P}\{\omega_{i_1} = 1,\ \ldots,\ \omega_{i_m} = 1\} - \mathbf{P}\{\omega_{i_1} = 1,\ \ldots,\ \omega_{i_{m+1}} = 1\}\end{aligned}$$

for any $i_1, \ldots, i_{m+1} = 1, \ldots, n$ $(m < n)$, then the probabilities

$$\begin{aligned}&\mathbf{P}\{\omega_{i_1} = 1,\ \ldots,\ \omega_{i_m} = 1,\ \omega_{i_{m+1}} = 0,\ \omega_{i_{m+2}}\\ &\qquad = \mathbf{P}\{\omega_{i_1} = 1,\ \ldots,\ \omega_{i_m} = 1,\ \omega_{i_{m+2}} = 0\}\\ &\qquad\quad - \mathbf{P}\{\omega_{i_1} = 1,\ \ldots,\ \omega_{i_m} = 1,\ \omega_{i_{m+1}} = 1,\ \omega_{i_{m+2}} = 0\}\end{aligned}$$

for any $i_1, \ldots, i_{m+1}, i_{m+2} = 1, \ldots, n$ $(m < n - 1)$.

In this way, we successively find all probabilities

$$\mathbf{P}\{\omega_{i_1} = 1,\ \ldots,\ \omega_{i_m} = 1,\ \omega_{i_{m+1}} = 0,\ \ldots,\ \omega_{i_n} = 0\}$$

for any $i_1, \ldots, i_n = 1,\ \ldots,\ n$ $(m < n)$. Obviously, events of the form $\{\omega_{i_1} = 1,\ \ldots,\ \omega_{i_m} = 1,\ \omega_{i_{m+1}} = 0,\ \ldots,\ \omega_{i_n} = 0\}$ represent all 'elementary events' $\omega \subseteq \Omega$, and for any event $A \subseteq \Omega$, its probability $\mathbf{P}(A)$ is given by the general formula (2.2).

For example, in the case $n = 2$, given

$$\mathbf{P}(A_1),\ \mathbf{P}(A_2),\ \mathbf{P}(A_1A_2),$$

we find

$$\mathbf{P}(A_1A_2^c) = \mathbf{P}(A_1) - \mathbf{P}(A_1A_2),\ \mathbf{P}(A_1^cA_2) = \mathbf{P}(A_2) - \mathbf{P}(A_1A_2),$$

$$\mathbf{P}(A_1^cA_2^c) = 1 - \big[\mathbf{P}(A_1) + \mathbf{P}(A_2)\big] + \mathbf{P}(A_1A_2).$$

2.5. THE GENERAL MODEL OF PROBABILITY THEORY

Let us first describe relations between events.

Events A_1 and A_2 are *equal* if the occurrence of A_1 implies the occurrence of A_2, and vice versa. A_1 and A_2 are called *disjoint* if the occurrence of one of them excludes the occurrence of the other one, in other words, if A_1 and A_2 cannot occur simultaneously.

The event A which occurs if and only if one of the events A_1 and A_2 occur, is called the *union* (*sum*) of A_1, A_2, and is denoted by $A = A_1 \cup A_2$. The union of several events $A_1, A_2, \ldots$ is defined analogously, and denoted by

$$A = \bigcup_k A_k.$$

The event A which occurs if and only if both A_1 and A_2 occur, is called the *intersection* (*product*) of A_1, A_2 and is denoted by $A = A_1 \cap A_2$. The product of several events $A_1, A_2, \ldots$ is defined analogously, and denoted by

$$A = \bigcap_k A_k,$$

or $A = A_1 \cdot A_2 \cdots$. The *difference* of A_1 and A_2 is the event A which occurs if and only if A_1 occurs whereas A_2 does not occur, and is denoted by $A = A_1 \backslash A_2$. The event A^c which occurs if and only if A does not occur, is called the *complementary event* to A.

Suppose that, among all possible events A which could occur in the given experiment, one can choose a set of *elementary events* with the following properties. Firstly, elementary events exclude each other (or, are disjoint) and, moreover, at least one of them certainly occurs during the experiment. Secondly, for any event A, the occurring elementary outcome decides whether A occurs or not. An elementary event is usually denoted by the Greek letter ω. The set Ω of all ω's is called the *space of elementary events*.

Let Ω be the space of elementary events ω of the considered experiment (phenomenon). With any event A connected with the experiment, we can associate the set of all possible outcomes ω whose occurrence implies A. We denote this set by the same symbol A, and identify it with the corresponding event.

The *certain event*, which occurs with every elementary outcome ω, equals the entire space Ω. The *impossible event* which never occurs, coincides with the empty set $\varnothing \subset \Omega$.

The notions of union, intersection etc. of events, introduced above, now become the corresponding relations between sets: $A_1 \cup A_1$ is the union of sets A_1 and A_2, $A_1 \cap A_2$ is their intersection, $A^c = \Omega \backslash A$ is the complement to A in the space Ω.

In particular, note that event A_1 implies the occurrence of event A_2, denoted by $A_1 \subseteq A_2$ or $A_2 \supseteq A_1$, if and only if A_1 is contained in A_2. The following properties of relations between events are useful. If $A_1 \subseteq A_2$, then $A_1^c \supseteq A_2^c$; if $A = A_1 \cup A_2$, then $A^c = A_1^c \cap A_2^c$; finally, if $A = A_1 \cap A_2$, then $A^c = A_1^c \cup A_2^c$. In general, if a certain relation among events is true, then the relation obtained by changing to complementary events and by replacing the symbols $\cup$, $\cap$, $\subseteq$, by the symbols $\cap$, $\cup$, $\supseteq$, respectively, is also true.

Often, one has to deal with events that are unions, intersections (products) etc. of other events. A family $\mathfrak{A}$ of events is called an *algebra* if it contains *finite* unions/intersections and complements of its elements (recall that

$$\Big(\bigcup_k A_k\Big)^c = \bigcap_k A_k^c,$$

for example). If, in addition, $\mathfrak{A}$ contains any *countable* unions/intersections, then it is called a *σ-algebra.*

The general model of the probability theory is given by a space Ω of elementary events ω equipped with probabilities $\mathbf{P}(A)$, $A \in \mathfrak{A}$, of all events $A \subseteq \Omega$ from a σ-algebra $\mathfrak{A}$, which satisfy the following conditions: for any event $A \in \mathfrak{A}$

$$0 \leqslant \mathbf{P}(A) \leqslant 1,$$

$\mathbf{P}(\varnothing) = 0$ for the impossible event $A = \varnothing$, and $\mathbf{P}(\Omega) = 1$ for the certain event $A = \Omega$; moreover, for any sequence $A_k \in \mathfrak{A}$, $k = 1, 2, \ldots$, of *disjoint* events,

$$\mathbf{P}\Big(\bigcup_k A_k\Big) = \sum_k \mathbf{P}(A_k). \tag{2.9}$$

(2.9) is called the *countable additivity* (or *σ-additivity*) property of the probability $\mathbf{P}(A)$, $A \in \mathfrak{A}$. This property is clearly satisfied in the probability model (2.1), (2.2) with countable number of elementary events; in particular, in (2.2) we just sum over all 'chances' $\omega \in A$ in favour of the event A.

For example, from (2.9) for any events $A \supseteq B$ we have

$$\mathbf{P}(A) = \mathbf{P}(B) + \mathbf{P}(A \backslash B),$$

or

$$\mathbf{P}(A) \geqslant \mathbf{P}(B)$$

whenever the occurrence of B implies A.

For a *finite* number of (disjoint) events, equality (2.9) seems rather obvious. In the case of *infinite* number of A_k, $k = 1, 2, \ldots$, we deal with increasing events

$$\bigcup_{k=1}^{n} A_k, \quad n = 1, 2, \ldots,$$

and the limit

$$\lim_{n\to\infty} \mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right)$$

of increasing bounded sequence

$$\mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right), \quad n = 1, 2, \ldots,$$

exists. The countable additivity property (2.9) says that this limit is exactly the probability $\mathbf{P}(A)$ of the *limit event*

$$A = \lim_{n\to\infty} \bigcup_{k=1}^{n} A_n = \bigcup_{k=1}^{\infty} A_k,$$

i.e.,

$$\lim_{n\to\infty} \mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right) = \mathbf{P}(A). \tag{2.10}$$

One can consider (2.10) as the *continuity* property of the probability, and apply it to *any* events A_k, $k = 1, 2, \ldots$, since

$$\bigcup_{k=1}^{n} A_k = \bigcup_{k=1}^{n} B_k, \qquad \bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} B_k = A$$

with *disjoint*

$$B_k = A_k \setminus \bigcup_{j=1}^{k-1} A_j, \quad k = 1, 2, \ldots.$$

In particular, (2.10) implies

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(A) \tag{2.11}$$

for any *increasing events* $A_1 \subseteq A_2 \subseteq \ldots$,

$$\bigcup_{k=1}^{n} A_k = A_n,$$

with the *limit event*

$$A = \lim_{n\to\infty} A_n \left(= \lim_{n\to\infty} \bigcup_{k=1}^{n} A_k \right).$$

The limit equality (2.11) holds also for any *decreasing events* $A_1 \supseteq A_2 \supseteq \ldots$, with the *limit event*

$$A = \lim_{n\to\infty} A_n \left(= \lim_{n\to\infty} \bigcap_{k=1}^{n} A_k \right).$$

Indeed, it is equivalent to

$$\lim_{n\to\infty} \left[1 - \mathbf{P}(A_n)\right] = 1 - \mathbf{P}(A),$$

where

$$1 - \mathbf{P}(A_n) = \mathbf{P}(A_n^c), \qquad 1 - \mathbf{P}(A) = \mathbf{P}(A^c)$$

and the complements $A_1^c \subseteq A_2^c \subseteq \ldots$ *increase*; moreover,

$$A^c = \left(\bigcap_{k=1}^{\infty} A_k \right)^c = \bigcup_{k=1}^{\infty} A_k^c.$$

In the sequel, we often use the following simple inequality:

$$\mathbf{P}\left(\bigcup_k A_k \right) \leqslant \sum_k \mathbf{P}(A_k), \tag{2.12}$$

where $A_k \in \mathfrak{A}$, $k = 1, 2, \ldots$, are *arbitrary* events. To prove (2.12), write

$$\bigcup_k A_k = \bigcup_k B_k,$$

where

$$B_k = A_k \Big\backslash \bigcup_{j=1}^{k-1} A_j, \quad k = 1, 2, \ldots,$$

are *disjoint* and $\mathbf{P}(B_k) \leqslant \mathbf{P}(A_k)$. Therefore,

$$\mathbf{P}\left(\bigcup_k B_k\right) = \sum_k \mathbf{P}(B_k) \leqslant \sum_k \mathbf{P}(A_k).$$

EXAMPLE. Suppose, each event A_k, $k = 1, 2, \ldots$, occurs with probability 1; what is the probability that they all occur simultaneously? The question concerns the event

$$A = \bigcap_k A_k,$$

with the complementary event

$$A^c = \bigcup_k A_k^c$$

satisfying

$$\mathbf{P}(A^c) \leqslant \sum_k \mathbf{P}(A_k^c) = 0,$$

since $\mathbf{P}(A_k^c) = 1 - \mathbf{P}(A_k) = 0$ for any $k = 1, 2, \ldots$. Hence

$$\mathbf{P}(A) = 1.$$

The following statement will serve later on as a powerful tool in our discussion.

LEMMA (First Borel–Cantelli lemma). *Let $A_1, A_2, \ldots$ be a sequence of events, with probabilities $p_k = \mathbf{P}(A)$, $k = 1, 2, \ldots$, such that*

$$\sum_{k=1}^{\infty} p_k < \infty. \tag{2.13}$$

Then, with probability 1 *only finitely many of the events $A_1, A_2, \ldots$ occur.*

Proof. Let B be the event that infinitely many of the events $A_1, A_2, \ldots$ occur. Put

$$B_n = \bigcup_{k \geqslant n} A_k,$$

so that B_n occurs if and only if at least one of the events $A_n, A_{n+1}, \ldots$ occurs. Clearly B occurs if and only if B_n occurs, for every $n = 1, 2, \ldots$. Therefore,

$$B = \bigcap_n B_n = \bigcap_n \left(\bigcup_{k \geqslant n} A_k \right).$$

Moreover, $B_1 \supset B_2 \supset \ldots$, hence

$$\mathbf{P}(B) = \lim_{n \to \infty} \mathbf{P}(B_n).$$

But

$$\mathbf{P}(B_n) \leqslant \sum_{k \geqslant n} \mathbf{P}(A_k) = \sum_{k \geqslant n} p_k \longrightarrow 0, \quad \text{as } n \to \infty,$$

because of (2.13). Therefore

$$\mathbf{P}(B) = \lim_{n \to \infty} \mathbf{P}(B_n) = 0,$$

i.e., the probability that infinitely many of the events $A_1, A_2, \ldots$ occur is 0. Equivalently, the probability that only finitely many of the events $A_1, A_2, \ldots$ occur is 1. □

In the general model of the probability theory, given by an abstract set $\Omega \ni \omega$ and a probability $\mathbf{P}(A)$, $A \in \mathfrak{A}$, defined on a σ-algebra of events $A \subseteq \Omega$, independence of

events has no 'physical' meaning. In such a case, we call events A_k, $k = 1, 2, \ldots$, *independent* (or *mutually independent*) if

$$\mathbf{P}(A_{i_1} \cdots A_{i_m}) = \mathbf{P}(A_{i_1}) \cdots \mathbf{P}(A_{i_m}) \tag{2.14}$$

for any mutually different $i_1, \ldots, i_m = 1, 2, \ldots$. The above definition is justified by our earlier discussion of the model with a *finite number* of equiprobable outcomes (see (1.11)). In another simple model with a finite number of events A_k, $k = 1, \ldots, n$ (see (2.8)), their independence in the sense of (2.14) means that A_k can be associated with independent trials Ω_k with two possible outcomes $\omega_k = 0$ or $\omega_k = 1$ corresponding to the occurrence of A_k^c or A_k, respectively, $k = 1, \ldots, n$, so that

$$\mathbf{P}(\omega_1, \ldots, \omega_n) = \mathbf{P}(\omega_1) \cdots \mathbf{P}(\omega_n),$$

see (2.4). In particular, the last model, with independent A_k, shows that any events which are equal either to A_k, or to its *complement* A_k^c ($k = 1, 2, \ldots$) are *mutually independent.*

LEMMA (Second Borel–Cantelli lemma). *Let $A_1, A_2, \ldots$ be a sequence of independent events, with probabilities $p_k = \mathbf{P}(A_k)$, $k = 1, 2, \ldots$, such that*

$$\sum_{k=1}^{\infty} p_k = \infty. \tag{2.15}$$

Then, with probability 1 *infinitely many of the events $A_1, A_2, \ldots$ occur.*

Proof. As in the proof of the first Borel–Cantelli lemma, let

$$B_n = \bigcup_{k \geqslant n} A_k, \qquad B = \bigcap_n B_n = \bigcap_n \Big(\bigcup_{k \geqslant n} A_k \Big),$$

so that B occurs if and only if infinitely many of the events $A_1, A_2, \ldots$ occur. By taking complements, we have

$$B_n^c = \bigcap_{k \geqslant n} A_k^c, \qquad B^c = \bigcup_n B_n^c.$$

In particular, for every $m = 0, 1, 2, \ldots$

$$B_n^c \subseteq \bigcap_{k=n}^{n+m} A_k^c.$$

Therefore,

$$\mathbf{P}\Big(B_n^c\Big) \leqslant \mathbf{P}\Big(\bigcap_{k=n}^{n+m} A_k^c\Big) = \mathbf{P}\Big(A_n^c\Big)\cdots\mathbf{P}\Big(A_{n+m}^c\Big)$$
$$= (1-p_n)\cdots(1-p_{n+m})$$
$$\leqslant \exp\Big(-\sum_{k=n}^{n+m} p_k\Big),$$

where we use the inequality $1 - x \leqslant \mathrm{e}^{-x}$, $x \geqslant 0$, and the fact that, if the events $A_1, A_2, \ldots$ are independent, then the complementary events $A_1^c, A_2^c, \ldots$ are also independent. But

$$\sum_{k=n}^{n+m} p_k \longrightarrow \infty \quad \text{as } m \to \infty$$

because of (2.15). Therefore, passing to the limit as $m \to \infty$, we find that, for every $n = 1, 2, \ldots$ $\mathbf{P}(B_n^c) = 0$. Consequently,

$$\mathbf{P}(B^c) \leqslant \sum_n \mathbf{P}(B_n^c) = 0,$$

or

$$\mathbf{P}(B) = 1 - \mathbf{P}(B^c) = 1,$$

i.e., the probability that infinitely many of the events $A_1, A_2, \ldots$ occur, is 1. □

Conditional probability. In the general model of the probability theory, it is assumed that the occurrence of an event B, $\mathbf{P}(B) > 0$, affects another event A in such a way that its *a posteriori* probability (i.e., the probability after B has occurred) becomes

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

The above probability is called the *conditional probability* of A given the event B. Of course, if the probabilities $\mathbf{P}(B)$ and $\mathbf{P}(A \mid B)$ are known, then we can find the probability of the event AB:

$$\mathbf{P}(AB) = \mathbf{P}(A \mid B)\mathbf{P}(B). \tag{2.16}$$

Suppose, we are given $\mathbf{P}(B_k)$, $\mathbf{P}(A \mid B_k)$ for some *disjoint* B_k, $k = 1, 2, \ldots$, and

$$A \subseteq \bigcup_k B_k.$$

Then, as

$$A = \bigcup_k (AB_k), \qquad \mathbf{P}(A) = \sum_k \mathbf{P}(AB_k),$$

we obtain the *total probability formula*:

$$\mathbf{P}(A) = \sum_k \mathbf{P}(A \mid B_k)\mathbf{P}(B_k). \tag{2.17}$$

As an application of the notion of conditional probability, consider the following problem.

Forecasting of events. Consider a random quantity $\xi = 1, 2, \ldots$ taking a finite number of integer values, depending on the outcome of another quantity (experiment) $\eta = 1, 2, \ldots$ which we observe. We want to forecast ξ given an observation of η. More precisely, we want to find an appropriate function $\widehat{\xi} = \varphi(\eta)$ of η which would serve as the forecast of ξ. Of course, the forecast can be wrong, which happens with the probability

$$\mathbf{P}\{\varphi(\eta) \neq \xi\}.$$

Let us find the *best forecast* $\varphi_0(\eta)$ such that

$$\mathbf{P}\{\varphi_0(\eta) \neq \xi\} \leqslant \mathbf{P}\{\varphi(\eta) \neq \xi\}$$

for any forecast $\varphi(\eta)$. We have

$$\begin{aligned}
\mathbf{P}\{\varphi(\eta) \neq \xi\} &= 1 - \mathbf{P}\{\varphi(\eta) = \xi\} \\
&= 1 - \sum_k \mathbf{P}\{\eta = k,\ \xi = \varphi(k)\} \\
&= 1 - \sum_k \mathbf{P}\{\xi = \varphi(k) \mid \eta = k\}\mathbf{P}\{\eta = k\}.
\end{aligned}$$

For any $k = 1, 2, \ldots$, define $j_0 = \varphi_0(k)$ as the *maximum* point of the conditional probability:

$$\mathbf{P}\{\xi = j_0 \mid \eta = k\} = \max_j \mathbf{P}\{\xi = j \mid \eta = k\}. \tag{2.18}$$

Then, for any φ,

$$\mathbf{P}\{\xi = \varphi_0(k) \mid \eta = k\} \geqslant \mathbf{P}\{\xi = \varphi(k) \mid \eta = k\},$$

and we immediately obtain the following result.

THEOREM. *The best forecast of* ξ *is given by* $\hat{\xi} = \varphi_0(\eta)$.

2.6. SOME EXAMPLES

Gambler's ruin problem. Consider the game of 'heads or tails', in which a coin is tossed and a player wins 1, say, if he successfully calls the side of the coin which lands upward, but otherwise loses 1. Suppose the player's initial capital is x, and he intends to play until he wins m but no longer. In other words, suppose the game continues until the player either wins the amount of m, stipulated in advance, or else loses all his capital and is ruined. What is the probability that the player will be ruined?

The probability of ruin clearly depends on both the initial capital x and the final amount m. Let $p(x)$ be the probability of the player's being ruined if he starts with a capital x. Then the probability of ruin, given that the player wins the first call, is just $p(x+1)$, since the player's capital becomes $x+1$ if he wins the first call. Similarly, the probability of ruin, given that the player loses the first call, is $p(x-1)$, since the player's capital becomes $x-1$ if he loses the first call. In other words, if B_1 is the event that the player wins the first call and B_2 the event that he loses the first call, while A is the event of ruin, then

$$\mathbf{P}(A \mid B_1) = p(x+1), \qquad \mathbf{P}(A \mid B_2) = p(x-1).$$

The mutually exclusive events B_1 and B_2 form a 'full set', since the player either wins or loses the first call. Moreover, we have

$$\mathbf{P}(B_1) = \frac{1}{2}, \qquad \mathbf{P}(B_2) = \frac{1}{2},$$

assuming fair tosses of an unbiased coin. Hence, by the total probability formula

$$\mathbf{P}(A) = \mathbf{P}(A \mid B_1)\mathbf{P}(B_1) + \mathbf{P}(A \mid B_2)\mathbf{P}(B_2),$$

we get that $p(x) = \varphi(x)$, as a function of $x = 0, \ldots, m$, satisfies the equation

$$\varphi(x) = \frac{1}{2}\left[\varphi(x+1) + \varphi(x-1)\right], \quad 1 \leqslant x \leqslant m-1,$$

where obviously

$$\varphi(0) = p(0) = 1, \qquad \varphi(m) = p(m) = 0.$$

The solution of the above equation is a *linear* function

$$\varphi(x) = c_1 + c_2(x),$$

where the coefficients c_1 and c_2 are determined by the boundary conditions:

$$c_1 = 1, \qquad c_1 + c_2 m = 0.$$

We finally find that the probability of ruin, given the initial capital of x, is just

$$p(x) = 1 - \frac{x}{m}, \quad 0 \leqslant x \leqslant m. \tag{2.19}$$

In a very similar way, one can find the corresponding probability not to be ruined but to win the final amount m

$$q(x) = \frac{x}{m},$$

which appears as the solution $\varphi(x) = q(x)$ of the functional equation considered above with the boundary conditions

$$\varphi(0) = q(0) = 0, \qquad \varphi(m) = q(m) = 1.$$

In total, the two probabilities (to be ruined or to win) give us

$$p(x) + q(x) = \left(1 - \frac{x}{m}\right) + \frac{x}{m} = 1, \tag{2.20}$$

which shows that there is *no chance* to play infinitely with the capital $0 < x < m$ always strictly between 0 and m, and not hitting these edge points in a series of *infinite* tosses. (Could you have guessed that result in advance, prior to the above calculations?)

Random walk. Imagine a particle which randomly 'walks' along the x-axis, visiting integer points $x = 0, \pm 1, \ldots$ only; once at point x, it shifts either to the point $x + 1$, or to the point $x - 1$, with probabilities p and $q = 1 - p$, respectively.

How often the particle can return to the initial point ($x = 0$, say)? Of course, the particle can be again at $x = 0$ after an even number ($= 2n$) of steps only, as the *total* number of steps to the left and to the right has to be the same. For a given n, the probability of such event equals

$$\mathbf{P}(n) = \binom{2n}{n} p^n q^n = \frac{(2n)!}{(n!)^2} p^n q^n,$$

According to the well-known *Stirling formula*

$$n! \sim \sqrt{2\pi n} n^n \mathrm{e}^{-n},$$

we find that, as $n \to \infty$,

$$\mathbf{P}(n) \sim \frac{1}{\sqrt{\pi n}} (4pq)^n.$$

In the case $p \neq q$, as $4pq = 1 - (p - q)^2 < 1$, we see that

$$\mathbf{P}(n) < \infty.$$

This shows, according to the Borel–Cantelli lemma, that after *infinitely many* steps with *probability* 1 the particle returns to initial point ($x = 0$) only a *finite* number of times. (One can guess that if $p > q$, say, then the particle moves to the right to $+\infty$, as time increases.)

What happens in the symmetric case $p = q = 1/2$? Let $\mathbf{P}^0(m)$ be the probability that the particle returns to 0 at $t = 2m$ for the first time. It is clear that if the first return occurs at $t = 2m$, then the conditional probability that the particle visits $x = 0$ at time $t = 2n$, is the same as the probability $\mathbf{P}(n - m)$ of visiting $x = 0$ at time $t = 2(n - m)$ from the very beginning. Hence, by the total probability formula,

$$\mathbf{P}(n) = \sum_{m=1}^{n} \mathbf{P}^0(m)\mathbf{P}(n - m), \quad n = 1, 2, \ldots, \quad \mathbf{P}(0) = 1,$$

which gives the following equation

$$F(z) - 1 = F^0(z)F(z), \qquad F(z) = \frac{1}{1 - F^0(z)}$$

for the *generating functions*

$$F(z) = \sum_{n=0}^{\infty} \mathbf{P}(n) z^n, \qquad F^0(z) = \sum_{m=1}^{\infty} \mathbf{P}^0(m) z^m, \quad |z| < 1.$$

We immediately see that

$$F^0(1) = \sum_{m=1}^{\infty} \mathbf{P}^0(m) = \mathbf{P}^0 = \lim_{z \to 1} F^0(z)$$

is the probability that the particle returns, at least once, to the origin. The equation

$$F(z) = \frac{1}{1 - F^0(z)}$$

shows that

$$\mathbf{P}^0 = \lim_{z \to \infty} F^0(z) = 1$$

if and only if

$$\lim_{z \to 1} F(z) = \sum_{n=0}^{\infty} \mathbf{P}(n) = \infty. \tag{2.21}$$

This is exactly the case of the symmetric random walk with $p = q = 1/2$, since in this case, $\mathbf{P}(n) \sim 1/\sqrt{\pi n}$, and (2.21) holds. Therefore, we can conclude that the particle returns to the initial point with probability 1. Obviously, after the first return, the situation will be exactly the same as at the very beginning and the second return occurs with probability 1, too, then, surely, will be the next one etc. Thus, with *probability* 1, the particle will return to the initial point infinitely many times.

If $p \neq q$, then condition (2.21) fails, and the return probability $\mathbf{P}^0 < 1$. What is $\mathbf{P}_0$? More generally, what is the probability for the random walk to hit a point $x = a$?

One can find it in a very similar way to the 'gambler's ruin' problem, see p. 26.

Namely, assume for a while that there are stopping barriers at points $x = a$, b ($a > 0 > b$), say, so when the particle comes to any of them, it will remain there forever. Consider the probability to hit the point a at some time, as a function $\varphi(x)$ of the

starting point x, of the random walk, $a \geqslant x \geqslant b$. According to the total probability formula,

$$\varphi(x) = p\varphi(x+1) + q\varphi(x-1), \quad a > x > b,$$

with the obvious boundary conditions

$$\varphi(a) = 1, \qquad \varphi(b) = 0.$$

In the case of $p \neq q$, one obtains

$$\varphi(x) = \frac{1-(q/p)^{x-b}}{1-(q/p)^{a-b}}, \quad a \geqslant x \geqslant b.$$

One can assume that the influence of the barrier at the left point b is negligible as $b \to -\infty$. Passing to the limit as $b \to -\infty$, we obtain the probability

$$\varphi(x) = \begin{cases} (q/p)^{x-a}, & p < q, \\ 1, & p > q, \end{cases}$$

for the particle, starting at $x \leqslant a$, to hit the point a at some time. Substituting a, p by b, q, respectively, we get the probability

$$\varphi(x) = \begin{cases} (p/q)^{x-b}, & p > q, \\ 1, & p < q \end{cases}$$

for the particle to hit b, starting from $x \geqslant b$. Now, we can find the probability

$$\mathbf{P}^0 = \left\{ \begin{matrix} p + q(q/p)^{-1}, & p < q \\ p(p/q)^{-1} + q, & p > q \end{matrix} \right\} = 1 - |p-q| \tag{2.22}$$

to return to the initial point, using the observation that the particle can return to 0 either from $x = 1$, or from $x = -1$, where it surely comes after the first step, with the probability p and q, respectively. □

Time distribution of radioactive decay. Let us return to the process of radioactive decay (see p. 15), with the probability p for a radium atom to disintegrate during a

time interval of length t. More precisely, a radium atom, existing at a moment t_0, will disintegrate at a random moment $t_0 + \tau \in (t_0, t_0 + t]$ with the probability

$$p = p(t) = \mathbf{P}\{\tau \leqslant t\},$$

which depends on $t \geqslant 0$. Consider the function

$$\varphi(t) = 1 - p(t) = \mathbf{P}\{\tau > t\}, \quad t \geqslant 0,$$

which is *decreasing* and

$$\varphi(0) = \mathbf{P}\{\tau > 0\} = 1.$$

Suppose we know that $\tau > s$, then we still have the radium atom at the moment $t_1 = t_0 + s$ and, according to our assumption, the corresponding *a posteriori* probability of $\tau > s + t$, given $\tau > s$, is the same as $\mathbf{P}\{\tau > t\}$. In other words, the conditional probability

$$\mathbf{P}\{\tau > s + t \mid \tau > s\} = \mathbf{P}\{\tau > t\}, \tag{2.23}$$

which implies

$$P\{\tau > s + t\} = P\{\tau > s\}P\{\tau > t\}.$$

This brings us the following functional equation:

$$\varphi(s + t) = \varphi(s)\varphi(t); \quad s, t \geqslant 0. \tag{2.24}$$

The probability $\varphi(t) = \mathbf{P}\{\tau > t\}$ is *continuous* at $t = 0$, since $\{\tau > 0\}$ is the limit of increasing events $\{\tau > t\}$, $t \to 0$. Equation (2.24) implies that $\varphi(s+t)$ is continuous at every point $s \geqslant 0$, i.e., $\varphi(t)$, $t \geqslant 0$, is a *continuous* function. Moreover, (2.24) implies, together with $\varphi(0) = 1$, that $\varphi(t)$, $t \geqslant 0$, is strictly *positive*, and one can check that $\log \varphi(t)$ is a *linear* function:

$$\log \varphi(t) = -\lambda t, \quad t \geqslant 0,$$

where $\lambda \geqslant 0$ is a constant. Finally, we obtain

$$\varphi(t) = e^{-\lambda t}, \quad t \geqslant 0. \tag{2.25}$$

Let be given at time t_0 some amount of radium, containing n atoms. Then

$$np = np(t), \qquad p(t) = 1 - e^{-\lambda t}$$

(see p. 15) is the average number of α-particles emitted during time interval (t_0, t_0+t), hence

$$n(t) = n - np(t) = ne^{-\lambda t}$$

is the average number of radium atoms left at time $t_0 + t$. According to this exponential law, one obtains

$$\frac{1}{2} = \frac{n(T)}{n} = e^{-\lambda T}$$

for the ratio of the initial amount of radium, and the amount left after time

$$T = \frac{\log 2}{\lambda}.$$

The half-life constant T (which does not depend on n) is experimentally known.

3. Random Variables

3.1. PROBABILITY DISTRIBUTIONS

We have already encountered numerous random variables in our discussion; in particular, the number of 'successes' in the Bernoulli trials, the number of α-particles emitted in a time interval, the time up to the moment of disintegration of a radium atom in the radioactive decay process, etc.

Roughly speaking, a *random variable* ξ is a quantity which takes its values 'at random' from a set of all possible values of ξ. A more precise meaning can be given at once in the (discrete) case of ξ taking a countable number of values $x \in \mathbb{R}$, $\mathbb{R} = (-\infty, \infty)$, with corresponding probabilities

$$\mathbf{P}(x) = \mathbf{P}_\xi(x), \qquad \sum_{-\infty}^{\infty} \mathbf{P}_\xi(x) = 1. \tag{3.1}$$

Here, 'randomness' of ξ is characterized by the *probability distribution* $\mathbf{P}_\xi = \mathbf{P}_\xi(x)$, $-\infty < x < \infty$, which determines the probability

$$\mathbf{P}_\xi(B) = \mathbf{P}\{\xi \in B\} = \sum_{x \in B} \mathbf{P}_\xi(x) \tag{3.2}$$

of an arbitrary event $\{\xi \in B\}$. One can recall here the hypergeometric distribution (1.3), the Bernoulli distribution (2.5) and the Poisson distribution (2.6) – all of them are discrete distributions over a corresponding set of integers $x = 0, 1, \ldots$.

In our example of the radioactive decay, we have actually met a random variable ξ of a different kind, taking values in any interval $x' \leqslant \xi \leqslant x''$ of the time axis $x \geqslant 0$ with the corresponding probability

$$\mathbf{P}\{x' < \xi \leqslant x''\} = \mathbf{P}\{\xi > x'\} - \mathbf{P}\{\xi > x''\} = \int_{x'}^{x''} \lambda \mathrm{e}^{-\lambda x}\, \mathrm{d}x.$$

Here, 'randomness' is characterized by the *probability density*

$$p(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x > 0, \\ 0, & x \leqslant 0, \end{cases}$$

cf. (2.25). In general, a *probability density* on the real lime $\mathbb{R} = (-\infty, \infty)$ is given by a function $p(x) \geqslant 0$ with

$$\int_{-\infty}^{\infty} p(x)\, \mathrm{d}x = 1,$$

and we say that ξ is a *random variable* with the *probability density* $p(x) = p_\xi(x)$, $-\infty < x < \infty$, if, for any interval $(x', x'']$, the probability

$$\mathbf{P}\{x' < \xi \leqslant x''\} = \int_{x'}^{x''} p_\xi(x)\, \mathrm{d}x. \tag{3.3}$$

EXAMPLE (*The uniform distribution*). Imagine that a point ξ is thrown 'at random' onto an interval $(a, b]$, as it happens e.g., in the roulette game, with $(a, b] = (-\pi, \pi]$ corresponding to the roulette circle. Then

$$\mathbf{P}\{x' < \xi \leqslant x''\} = \frac{x'' - x'}{b - a} = \int_{x'}^{x''} p_\xi(x)\, \mathrm{d}x,$$

with the probability density

$$p_\xi(x) = \begin{cases} \dfrac{1}{b-a}, & a < x \leqslant b, \\ 0, & x \leqslant a,\ x > b. \end{cases}$$

By the help of a probability density, we define the probability of an arbitrary event $\{\xi \in B\}$ as

$$\mathbf{P}_\xi(B) = \mathbf{P}\{\xi \in B\} = \int_B p_\xi(x)\,\mathrm{d}x. \tag{3.4}$$

In general, when speaking of a random variable $\xi \in \mathbb{R}$, we have in mind its *probability distribution*

$$\mathbf{P}_\xi\colon\ \mathbf{P}_\xi(B) = \mathbf{P}\{\xi \in B\}, \quad B \subseteq \mathbb{R}, \tag{3.5}$$

or the probabilities $\mathbf{P}_\xi(B)$ for ξ to belong to certain sets $B \subseteq \mathbb{R}$, including all intervals $B = (x', x'']$. From the latter, one can form many other $B \subseteq \mathbb{R}$ and determine the corresponding probabilities $\mathbf{P}_\xi(B)$, according to the known properties of countable additivity and continuity. For example, one has

$$[x', x''] = \lim_{n\to\infty} \left(x' - \frac{1}{n}, x''\right],$$

$$\mathbf{P}\{x' \leqslant \xi \leqslant x''\} = \lim_{n\to\infty} \mathbf{P}\left\{x' - \frac{1}{n} < \xi \leqslant x''\right\}$$

for closed intervals,

$$[x] = \lim_{n\to\infty} \left(x - \frac{1}{n}, x\right], \qquad \mathbf{P}\{\xi = x\} = \lim_{n\to\infty} \mathbf{P}\left\{x - \frac{1}{n} < \xi \leqslant x\right\}$$

for single points,

$$(-\infty, x] = \lim_{x'\to-\infty} (x', x], \qquad \mathbf{P}\{\xi \leqslant x\} = \lim_{x'\to-\infty} \mathbf{P}\{x' < \xi \leqslant x\},$$

etc.

On the other hand, for $B = (x', x'']$ we can define $\mathbf{P}\{\xi \in B\}$ as

$$\mathbf{P}\{x' < \xi \leqslant x''\} = F_\xi(x'') - F_\xi(x'),$$

where

$$F_\xi(x) = \mathbf{P}\{\xi \leqslant x\}, \quad -\infty < x < \infty;$$

is the so-called *distribution function*. Obviously, it is increasing, right-continuous, and

$$\lim_{x \to -\infty} F_\xi(x) = 0, \qquad \lim_{x \to +\infty} F_\xi(x) = 1$$

(why?).

3.2. JOINT PROBABILITY DISTRIBUTION

Dealing with *discrete* random variables ξ_k, $k = 1, \ldots, n$, we assume that there exists their *joint probability distribution*

$$\mathbf{P}_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_2) = \mathbf{P}\{\xi_1 = x_1, \ \ldots, \ \xi_n = x_n\}, \quad -\infty < x_1, \ldots, x_n < \infty,$$

where $x_1, \ldots, x_n$ range over a *countable* number of all possible values, and

$$\sum_{-\infty}^{\infty} \cdots \sum_{-\infty}^{\infty} \mathbf{P}_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) = 1.$$

Summing up over $(x_1, \ldots, x_n) \in B$, we get the probability

$$\begin{aligned} \mathbf{P}_{\xi_1,\ldots,\xi_n}(B) &= \mathbf{P}\{(\xi_1, \ldots, \xi_n) \in B\} \\ &= \sum_B \cdots \sum \mathbf{P}_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) \end{aligned} \tag{3.6}$$

of an arbitrary event $\{(\xi_1, \ldots, \xi_n) \in B\}$. Obviously, the probability distribution of $\xi_1, \ldots, \xi_m$ $(m < n)$ alone can be obtained from (3.6) as

$$\mathbf{P}_{\xi_1,\ldots,\xi_m}(x_1, \ldots, x_m) = \underbrace{\sum_{-\infty}^{\infty} \cdots \sum_{-\infty}^{\infty}}_{n-m} \mathbf{P}_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n), \tag{3.7}$$

where we sum over all possible $(x_{m+1}, \dots, x_n) \in \mathbb{R}^{n-m}$.

Another type of random variables $\xi_k \in \mathbb{R}$, $k = 1, \dots, n$, correspond to probabilities of the form

$$\begin{aligned} \mathbf{P}_{\xi_1,\dots,\xi_n}(B) &= \mathbf{P}\{(\xi_1, \dots, \xi_n) \in B\} \\ &= \int_B \cdots \int p_{\xi_1,\dots,\xi_n}(x_1, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n, \end{aligned} \tag{3.8}$$

for various sets $B \subseteq \mathbb{R}^n$ we are interested in, in particular, to probabilities

$$\begin{aligned} &\mathbf{P}\{x_1' < \xi_1 \leqslant x_1'', \dots, x_n' < \xi_n \leqslant x_n''\} \\ &\quad = \int_{x_1'}^{x_1''} \cdots \int_{x_n'}^{x_n''} p_{\xi_1,\dots,\xi_n}(x_1, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n \end{aligned}$$

corresponding to $B = (x_1', x_1''] \times \cdots \times (x_n', x_n'']$.*

The function

$$\begin{aligned} &p_{\xi_1,\dots,\xi_n}(x_1, \dots, x_n) \geqslant 0, \quad -\infty < x_1, \dots, x_n < \infty, \\ &\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\xi_1,\dots,\xi_n}(x_1, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n = 1, \end{aligned} \tag{3.9}$$

is called the *joint probability density* of $\xi_1, \dots, \xi_n$. In this case, the probability density of $\xi_1, \dots, \xi_m$ $(m < n)$, can be written as

$$\begin{aligned} &p_{\xi_1,\dots,\xi_m}(x_1, \dots, x_m) \\ &\quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\xi_1,\dots,\xi_n}(x_1, \dots, x_n)\, \mathrm{d}x_{m+1} \dots \mathrm{d}x_n. \end{aligned} \tag{3.10}$$

EXAMPLE (*Buffon's needle problem*). Suppose a needle is tossed at random onto a plane ruled with parallel lines a distance L apart, where by a 'needle' we mean a line segment of length $l \leqslant L$. What is the probability of the needle intersecting one of the parallel lines?

* It is worthwhile to mention that sets $B \subseteq \mathbb{R}^n$ look like 'boxes'. Other sets $B \subseteq \mathbb{R}^n$ can be formed by means of their unions and corresponding limits.

Let ξ_1 be the angle between the needle and the direction of the rulings, and let ξ_2 be the distance between the bottom point of the needle and the nearest line above this point (see Figure 1). Then, if the conditions

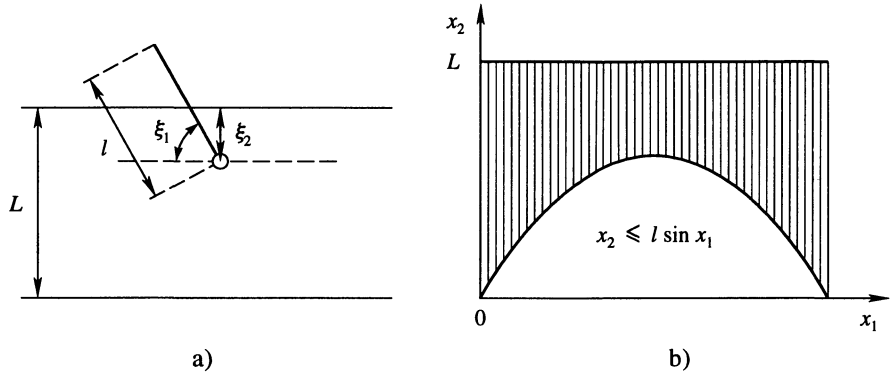

Fig. 1.

of the 'needle tossing experiment' are such that the random variable ξ_1 is uniformly distributed in the interval $(0, \pi]$, while the random variable ξ_2 is uniformly distributed in the interval $(0, L]$ and, moreover, ξ_1, ξ_2 is uniformly distributed over the rectangle $(0, \pi] \times (0, L]$, we find that their joint probability density is

$$p_{\xi_1,\xi_2}(x_1, x_2) = \frac{1}{\pi L}, \quad 0 < x_1 \leqslant \pi, \quad 0 < x_2 \leqslant L.$$

The event consisting of the needle intersecting one of the rulings occurs if and only if

$$\xi_2 \leqslant l \sin \xi_1,$$

i.e., if and only if the corresponding point $\xi = (\xi_1, \xi_2)$ falls in the region B, where B is the part of the rectangle $0 \leqslant x_1 \leqslant \pi$, $0 \leqslant x_2 \leqslant L$ lying between the x_1-axis and the curve $x_2 = \sin x_1$ [B is the unshaded region in Figure 1 (b)]. Hence, by the general formula (3.8),

$$\mathbf{P}\{(\xi_1, \xi_2) \in B\} = \int\int_B \frac{dx_1\, dx_2}{\pi L} = \frac{l}{\pi L} \int_0^{\pi} \sin x_1\, dx_1 = \frac{2l}{\pi L}.$$

This can be tested experimentally; in fact, if the needle is repeatedly tossed onto the ruled plane, then the frequency of the event A, consisting of the needle intersecting

one of the rulings, must be approximately $2l/(\pi L)$. Suppose the needle is tossed n times, and let $n(A)$ be the number of times A occurs, so that $n(A)/n$ is the relative frequency of the event A. Then, for large n,

$$\frac{n(A)}{n} \sim \frac{2l}{\pi L}.$$

Hence

$$\frac{2l}{L} \cdot \frac{n}{n(A)}$$

should be a good approximation to $\pi = 3.14\ldots$, for large n. This actually turns out to be the case.*

3.3. INDEPENDENT RANDOM VARIABLES

Discrete random variables ξ_k, $k = 1, \ldots, n$, are said *(mutually) independent* if

$$\mathbf{P}_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) = \mathbf{P}_{\xi_1}(x_1) \cdots \mathbf{P}_{\xi_n}(x_n), \quad -\infty < x_1, \ldots, x_n < \infty, \tag{3.11}$$

i.e., if their joint distribution is the product of (marginal) probability distributions of these random variables. Similarly, random variables ξ_k, $k = 1, \ldots, n$ with a joint probability density are called *(mutually) independent* if

$$p_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) = p_{\xi_1}(x_1) \cdots p_{\xi_n}(x_n), \quad -\infty < x_1, \ldots, x_n < \infty, \tag{3.12}$$

i.e., if the joint density is the product of corresponding marginal densities of these random variables.

EXAMPLE (*Normal, or Gaussian, distribution*). Let us imagine shooting at a target which is located at the origin of the $\mathbb{R}^2$ plane. The marks can be expected to be random points (ξ_1, ξ_2), with distribution which is centrally symmetric around the origin. Moreover, we can assume that 'errors' ξ_1, ξ_2, along orthogonal coordinates in $\mathbb{R}^2$, are independent, and jointly distributed according to a *probability density*

$$p_{\xi_1,\xi_2}(x_1, x_2) = p(x_1) \cdot p(x_2), \quad -\infty < x_1, x_2 < \infty,$$

* J.U. Uspensky, *Introduction to Mathematical Probability*, McGraw-Hill, New York, 1937, p. 113.

where $p(x_1), p(x_2)$ represent corresponding probability densities of ξ_1, ξ_2, respectively. In view of the central symmetry, we have

$$p_{\xi_1,\xi_2}(x_1, x_2) = f\left(x_1^2 + x_2^2\right)$$

as a function of $r^2 = x_1^2 + x_2^2$. Hence, with $x_1 = 0, x_2 = x$, we obtain

$$f(x^2) \equiv p(0)p(x), \quad -\infty < x < \infty,$$

or

$$p(0) \neq 0, \qquad f(0) = p(0)^2 \neq 0.$$

By taking $x_1^2 = s$, $x_2^2 (= x^2) = t$, one easily obtains from the above equations that

$$\varphi(t) = \frac{f(t)}{f(0)} = \frac{p(x)}{p(0)}, \quad t \geqslant 0,$$

satisfies the known equation

$$\varphi(s + t) = \varphi(s)\varphi(t), \quad s, t \geqslant 0.$$

Hence,

$$\varphi(t) = \mathrm{e}^{-\lambda t}, \quad t \geqslant 0,$$

see (2.25), and, consequently,

$$p(x) = p(0)\mathrm{e}^{-\lambda x^2}, \quad -\infty < x < \infty.$$

To find the constants $p(0)$ and $\lambda > 0$, we shall need the equality

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\, \mathrm{d}x = 1,$$

which follows from

$$\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)^2$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x_1^2+x_2^2)/2}\,dx_1\,dx_2$$
$$= \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_0^{\infty} e^{-r^2/2} r\,dr = 1.$$

Substituting x by x/σ gives

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2\sigma^2}\,dx \equiv 1, \quad \sigma > 0. \tag{3.13}$$

Moreover, differentiation of the identity

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-ux^2/2}\,dx \equiv \frac{1}{\sqrt{u}}, \quad u > 0,$$

with respect to $u > 0$ gives

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-x^2/2\sigma^2}\,dx = \sigma^2. \tag{3.14}$$

Returning to $p > 0$ and $\lambda > 0$, put $\lambda = 1/2\sigma^2$, then, using the condition

$$\int_{-\infty}^{\infty} p(x)\,dx = 1$$

for the probability density $p(x)$, $-\infty < x < \infty$, we obtain from (3.13) that

$$p(0) = \frac{1}{\sigma\sqrt{2\pi}}.$$

Thus, the marginal probability density of $\xi = \xi_1, \xi_2$ is

$$p_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \quad -\infty < x < \infty, \tag{3.15}$$

and is called *normal* (or *Gaussian*); the corresponding parameter $\sigma^2 > 0$ is given by

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 p_\xi(x)\,\mathrm{d}x,$$

see (3.14), and is called the *variance* of the random variable ξ. The joint probability density

$$p_{\xi_1,\xi_2}(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2\sigma^2}(x_1^2 + x_2^2)\right\}, \quad -\infty < x < \infty, \tag{3.16}$$

of independent Gaussian random variables ξ_1, ξ_2, is also called *normal* (or *Gaussian*).

3.4. CONDITIONAL DISTRIBUTIONS

The dependence between two *discrete* random variables ξ and η can be characterized by the *conditional probability distribution*

$$\mathbf{P}_\xi(x \mid y), \quad -\infty < x < \infty,$$

of ξ given an outcome $\{\eta = y\}$; it is assumed that for any y, $-\infty < y < \infty$, the identity

$$\mathbf{P}_\xi(x \mid y)\mathbf{P}_\eta(y) = \mathbf{P}_{\xi,\eta}(x, y) \tag{3.17}$$

holds true. Here, the so-called *Bayes formula* applies:

$$\mathbf{P}_\xi(x \mid y) = \mathbf{P}_\xi(x)\frac{\mathbf{P}_\eta(y \mid x)}{\mathbf{P}_\eta(y)}.$$

In a similar way, for random variables ξ, η having a joint probability density $p_{\xi,\eta}(x, y)$, one can define the *conditional probability density*

$$p_\xi(x \mid y), \quad -\infty < x < \infty,$$

of ξ given an outcome $\{\eta = y\}$; it is assumed that for all y, $-\infty < y < \infty$, the identity

$$p_\xi(x \mid y)p_\eta(y) = p_{\xi,\eta}(x, y) \tag{3.18}$$

holds true, together with the *Bayes formula*:

$$p_\xi(x \mid y) = p_\xi(x) \frac{p_\eta(y \mid x)}{p_\eta(y)}.$$

In the same way, one can define conditional probabilities and densities for any random vectors $(\xi_1, \ldots, \xi_m)$ and $(\eta_1, \ldots, \eta_n)$; in the corresponding formulas (3.17), (3.18) one has to replace ξ, η, x, y by $(\xi_1, \ldots, \xi_m)$, $(\eta_1, \ldots, \eta_n)$, $(x_1, \ldots, x_m)$, $(y_1, \ldots, y_n)$, respectively.

To check the probabilistic intuition, consider the following question: what is the conditional probability distribution of the sum $\xi = \xi_1 + \xi_2$ of independent random variables ξ_1, ξ_2 given $\eta = \xi_2$? One can guess that, in the case of *discrete* ξ_1, ξ_2, the conditional distribution of $\xi = \xi_1 + \xi_2$ given $\eta(= \xi_2) = y$ is

$$\mathbf{P}_\xi(x \mid y) = \mathbf{P}_{\xi_1}(x - y), \quad -\infty < x < \infty,$$

just like $\eta(= \xi_2) \equiv y$ is being constant and, in the case of ξ_1, ξ_2 having a probability density, the corresponding conditional probability density is

$$p_\xi(x \mid y) = p_{\xi_1}(x - y), \quad -\infty < x < \infty.$$

3.5. FUNCTIONS OF RANDOM VARIABLES

Given two independent random variables ξ_1 and ξ_2 with probability densities $p_{\xi_1}(x)$ and $p_{\xi_2}(x_2)$, what can we say about the distribution of $\xi = \xi_1 + \xi_2$? The answer to this simple question, which is part of a general problem concerning functions of random variables, is that the probability density of ξ is given by the *convolution*

$$p_\xi = p_{\xi_1} * p_{\xi_2},$$

i.e.,

$$p_\xi(x) = \int_{-\infty}^{\infty} p_{\xi_1}(x - y) p_{\xi_2}(y)\, dy, \quad -\infty < x < \infty, \tag{3.19}$$

since

$$\mathbf{P}\{x' < \xi \leqslant x''\} = \iint\limits_{x' < x_1 + x_2 \leqslant x''} p_{\xi_1}(x_1) p_{\xi_2}(x_2)\, dx_1\, dx_2$$

$$= \int_{x'}^{x''} \left\{ \int_{-\infty}^{\infty} p_{\xi_1}(x - y) p_{\xi_2}(y)\, dy \right\} dx$$

follows by the substitution $x = x_1 + x_2$, $y = x_2$.

EXAMPLE (*Triangular distribution*). This is the distribution of $\xi = \xi_1 + \xi_2$, where ξ_1, ξ_2 are independent and uniformly distributed in $(-a, 0)$ and $(0, a)$, respectively; its probability density is

$$p_\xi(x) = \frac{1}{a}\int_0^a p_{\xi_1}(x-y)\,dy$$

$$= \begin{cases} \frac{1}{a^2}\int_0^{x+a} dy = \frac{1}{a}\left(1+\frac{x}{a}\right), & -a < x \leqslant 0, \\ \frac{1}{a^2}\int_x^a dy = \frac{1}{a}\left(1-\frac{x}{a}\right), & 0 < x \leqslant a, \\ 0, & x < -a,\ x > a \end{cases} \tag{3.20}$$

(see Figure 2).

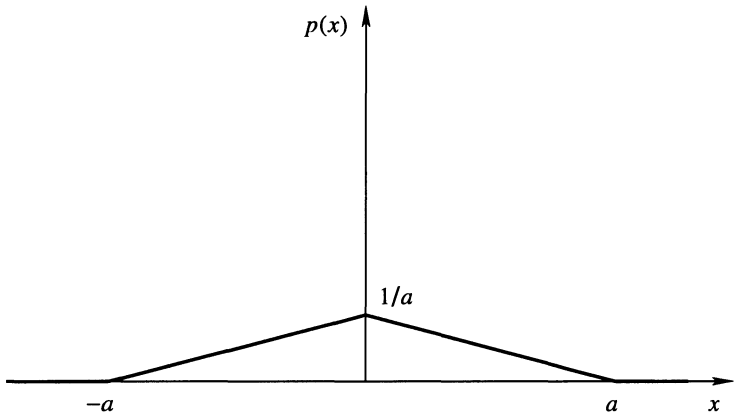

Fig. 2.

EXAMPLE (*Gamma-distribution*). Let $\xi_1, \ldots, \xi_n$ be independent random variables having the same exponential distribution, or probability density

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0, \\ 0, & x < 0. \end{cases}$$

The sum $\xi = \xi_1 + \cdots + \xi_n$ has the so-called *gamma-distribution*, with the probability density

$$p_\xi(x) = \begin{cases} \lambda \frac{(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x}, & x \geqslant 0, \\ 0, & x < 0, \end{cases} \tag{3.21}$$

which is the n-fold convolution

$$p(x)^{*n} = \int_{-\infty}^{\infty} p(x-y)^{*(n-1)} p(y)\,dy, \quad -\infty < x < \infty \quad (n = 2, 3, \ldots),$$

$$p(x)^{*1} \equiv p(x), \quad -\infty < x < \infty.$$

Let a joint probability density of $\xi_1, \ldots, \xi_n$ be given. We want to find the joint density of the random variables

$$\eta_1 = \varphi_1(\xi_1, \ldots, \xi_n), \ldots, \eta_n = \varphi_n(\xi_1, \ldots, \xi_n),$$

where

$$y_1 = \varphi_1(x_1, \ldots, x_n), \ldots, y_n = \varphi_n(x_1, \ldots, x_n)$$

is a one-to-one differentiable mapping $\mathbb{R}^n \to \mathbb{R}^n$ with a non-degenerate Jacobi determinant

$$J(x_1, \ldots, x_n) = \begin{vmatrix} \frac{\partial \varphi_1}{\partial x_1} & \cdots & \frac{\partial \varphi_1}{\partial x_n} \\ \frac{\partial \varphi_n}{\partial x_1} & \cdots & \frac{\partial \varphi_n}{\partial x_n} \end{vmatrix} \neq 0.$$

One can verify that the joint probability density of $\eta_1, \ldots, \eta_n$ is given by

$$p_\eta(y_1, \ldots, y_n) = p_\xi(x_1, \ldots, x_n)\big|J(x_1, \ldots, x_n)\big|^{-1}, \quad (y_1, \ldots, y_n) \in \mathbb{R}^n. \tag{3.22}$$

3.6. RANDOM VARIABLES IN THE GENERAL MODEL OF PROBABILITY THEORY

Given a family of random variables ξ, one can assume that all of them are associated with some probability model $(\Omega, \mathfrak{A}, \mathbf{P})$, where Ω is a space of elementary events $\omega \in \Omega$, equipped with probabilities $\mathbf{P}(A)$ of all events $A \subseteq \Omega$ belonging to a σ-algebra $\mathfrak{A}$. Any random variable ξ can be considered as a function

$$\xi = \xi(\omega), \quad \omega \in \Omega, \tag{3.23}$$

on the space Ω, under the implicit assumption that, for $\xi \in \mathbb{R}$,

$$\{x' < \xi \leqslant x''\} \in \mathfrak{A}, \quad -\infty < x' < x'' < \infty.$$

Then, every ξ can be *approximated* by an appropriate *discrete* random variable, for example

$$\xi^h = kh, \quad (k-1)h < \xi \leqslant kh,$$

with the probabilities

$$\mathbf{P}\{\xi^h = kh\} = \mathbf{P}\{(k-1)h < \xi < kh\}, \quad k = 0, \pm 1, \ldots .$$

Obviously,

$$|\xi - \xi^h| \leqslant h \tag{3.24}$$

for *all* possible outcomes, hence we have a *uniform approximation* of ξ with the uniform convergence $\xi^h \to \xi$, when $h \to 0$.

The above approximation helps to characterize various properties of random variables by means of corresponding properties of discrete random variables; for example, a very intuitive definition of *independent* random variables ξ_k, $k = 1, \ldots, n$, can be given, in the sense that they take their values *independently* from each other, by requiring the corresponding approximations ξ_k^h, $k = 1, \ldots, n$, to be independent according to definition (3.11). In particular, we call random variables $\xi_1, \ldots, \xi_n$ (*mutually*) *independent* if any events of the type

$$\{x_1' < \xi_1 \leqslant x_1''\}, \ldots, \{x_n' < \xi_n \leqslant x_n''\} \tag{3.25}$$

are (mutually) independent; c.f. (3.11), (3.12), using the definition of independent events given in (2.14).

4. Mathematical Expectation

4.1. MEAN VALUE OF DISCRETE VARIABLE

Consider a discrete random variable ξ, taking value $\xi = x$ with probability

$$\mathbf{P}_\xi(x) = \mathbf{P}\{\xi = x\}, \quad -\infty < x < \infty.$$

The sum

$$\mathbf{E}\xi = \sum_{-\infty}^{\infty} x\mathbf{P}_\xi(x) = \sum_{-\infty}^{\infty} x\mathbf{P}\{\xi = x\} \tag{4.1}$$

is called the *mathematical expectation* (or the *mean value*) of ξ, assuming that it absolutely converges, i.e.,

$$\sum_{-\infty}^{\infty} |x|\mathbf{P}_\xi(x) < \infty,$$

and we sum over the countable set of all possible values x of the *discrete* variable ξ. The term 'mean value' has a very explicit meaning in the case when ξ takes a finite number N of values $x = x_1, \ldots, x_N$, with equal probabilities $\mathbf{P}_\xi(x) = 1/N$, as

$$\mathbf{E}\xi = \frac{1}{N}\sum_{k=1}^{N} x_k.$$

Recall that we have already discussed the *mean value* of Bernoulli and Poisson distributions; see (2.5), (2.6).

Speaking about the general case, note at once that if $\xi = a$ takes a constant value $x = a$ with probability 1, then

$$\mathbf{E}\xi = a. \tag{4.2}$$

Next, if $\xi = 1_A$ is the *indicator* of an event A ($1_A = 1$ if A occurs, $1_A = 0$ otherwise), then

$$\mathbf{E}1_A = \mathbf{P}(A).$$

If $\eta = \varphi(\xi)$ is a function of a random variable ξ with probability distribution $\mathbf{P}_\xi(x)$, $-\infty < x < \infty$, then

$$\mathbf{E}\eta = \sum_{-\infty}^{\infty} y\mathbf{P}\{\eta = y\} = \sum_{-\infty}^{\infty} y \sum_{x:\ \varphi(x)=y} \mathbf{P}_\xi(x) = \sum_{-\infty}^{\infty} \varphi(x)\mathbf{P}_\xi(x),$$

where we sum over all x, $-\infty < x < \infty$, assuming the absolute convergence of the series, i.e.,

$$\sum_{-\infty}^{\infty} |\varphi(x)|\mathbf{P}_{\xi}(x) < \infty.$$

In a similar way, given $\eta = (\xi_1, \ldots, \xi_n)$ as a function of discrete random variables $\xi_1, \ldots, \xi_n$, we obtain

$$\mathbf{E}\varphi(\xi_1, \ldots, \xi_n) = \sum_{-\infty}^{\infty} \cdots \sum_{-\infty}^{\infty} \varphi(x_1, \ldots, x_n)\mathbf{P}_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n),$$

assuming

$$\sum_{-\infty}^{\infty} \cdots \sum_{-\infty}^{\infty} |\varphi(x_1, \ldots, x_n)|\mathbf{P}_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) < \infty.$$

In particular, formula (4.3) implies

$$\mathbf{E}|\xi| = \sum_{-\infty}^{\infty} |x|\mathbf{P}_{\xi}(x). \tag{4.3}$$

The last sum is always well defined, although sometimes it can be infinite, and $\mathbf{E}|\xi| < \infty$ is just the condition which was assumed in definition (4.1) of the mathematical expectation. Sometimes it will be convenient to use this condition, in order to indicate the very existence of $\mathbf{E}\xi$. □

Formula (4.3) helps to reveal some remarkable properties of the mathematical expectation $\mathbf{E}\xi$, concerning its dependence on ξ. Namely, $\mathbf{E}\xi$ is *linear* in the sense that, for any linear combination

$$\xi = \sum_{k=1}^{n} c_k \xi_k$$

of random variables $\xi_1, \ldots, \xi_n$, we have

$$\mathbf{E}\left(\sum_{k=1}^{n} c_k \xi_k\right) = \sum_{k=1}^{n} c_k \mathbf{E}\xi_k. \tag{4.4}$$

$\mathbf{E}\xi$ is *multiplicative* in the sense that, for any product $\xi = \xi_1 \cdots \xi_n$ of (mutually) *independent* ξ_k, $k = 1, \ldots, n$, we have

$$\mathbf{E}(\xi_1 \cdots \xi_n) = \mathbf{E}\xi_1 \cdots \mathbf{E}\xi_n. \tag{4.5}$$

These properties can be verified by an application of the general formula (4.3). For example,

$$\begin{aligned}
\mathbf{E}(c_1\xi_1 + c_2\xi_2) &= \sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty}(c_1x_1 + c_2x_2)\mathbf{P}_{\xi_1,\xi_2}(x_1, x_2) \\
&= c_1\sum_{-\infty}^{\infty}x_1\sum_{-\infty}^{\infty}\mathbf{P}_{\xi_1,\xi_2}(x_1, x_2) + c_2\sum_{-\infty}^{\infty}x_2\sum_{-\infty}^{\infty}\mathbf{P}_{\xi_1,\xi_2}(x_1, x_2) \\
&= c_1\sum_{-\infty}^{\infty}x_1\mathbf{P}_{\xi_1}(x_1) + c_2\sum_{-\infty}^{\infty}x_2\mathbf{P}_{\xi_1}(x_2) \\
&= c_1\mathbf{E}\xi_1 + c_2\mathbf{E}\xi_2,
\end{aligned}$$

and, for *independent* ξ_1, ξ_2 with $\mathbf{E}|\xi_1|, \mathbf{E}|\xi_2| < \infty$, the absolute convergence of

$$\begin{aligned}
&\sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty}|x_1|\,|x_2|\mathbf{P}_{\xi_1,\xi_2}(x_1, x_2) \\
&\quad = \sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty}|x_1|\,|x_2|\mathbf{P}_{\xi_1}(x_1)\mathbf{P}_{\xi_2}(x_2) \\
&\quad = \sum_{-\infty}^{\infty}|x_1|\mathbf{P}_{\xi_1}(x_1)\sum_{-\infty}^{\infty}|x_2|\mathbf{P}_{\xi_2}(x_2) < \infty
\end{aligned}$$

implies the existence of

$$\begin{aligned}
\mathbf{E}(\xi_1 \cdot \xi_2) &= \sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty}x_1x_2\mathbf{P}_{\xi_1,\xi_2}(x_1, x_2) \\
&= \sum_{-\infty}^{\infty}x_1\mathbf{P}_{\xi_1}(x_1)\sum_{-\infty}^{\infty}x_2\mathbf{P}_{\xi_2}(x_2) = \mathbf{E}\xi_1 \cdot \mathbf{E}\xi_2.
\end{aligned}$$

Moreover, if $\xi_1, \ldots, \xi_n$ are *independent*, then, for arbitrary functions $\varphi_1, \ldots, \varphi_n$ such that $\mathbf{E}\varphi_k(\xi_k)$, $k = 1, \ldots, n$, exist, we have

$$\mathbf{E}\big[\varphi_1(\xi_1) \cdots \varphi_n(\xi_n)\big] = \mathbf{E}\varphi_1(\xi_1) \cdots \mathbf{E}\varphi_n(\xi_n). \tag{4.5}'$$

□

Let us consider random variables as functions of $\omega \in \Omega$, i.e.,

$$\xi = \xi(\omega), \quad \omega \in \Omega,$$

in the framework of the general model $(\Omega, \mathfrak{A}, \mathbf{P})$ of the probability theory; see (3.23). Obviously, any *discrete* random variable ξ is a *discrete function* of $\omega \in \Omega$, taking a countable number of values

$$\xi(\omega) = x_k, \quad \omega \in A_k, \quad k = 1, 2, \ldots,$$

on *disjoint* sets $A_k \in \mathfrak{A}$ forming a partition

$$\Omega = \bigcup_k A_k.$$

Then

$$\mathbf{E}\xi = \sum_k x_k \mathbf{P}(A_k), \tag{4.6}$$

since

$$\sum_k x_k \mathbf{P}(A_k) = \sum_{-\infty}^{\infty} x \sum_{k:\ x_k = x} \mathbf{P}(A_k)$$

$$= \sum_{-\infty}^{\infty} x \mathbf{P}\{\xi = x\} = \sum_{-\infty}^{\infty} x \mathbf{P}_\xi(x)$$

with

$$\sum_k |x_k| \mathbf{P}(A_k) = \sum_{-\infty}^{\infty} |x| \mathbf{P}_\xi(x) < \infty.$$

Representation (4.6) permits us to prove some properties of the mean value $\mathbf{E}\xi$ with respect to ξ, by taking *the same* partition

$$\Omega = \bigcup_k A_k$$

for different ξ. Actually, this enables to consider random variables as functions of the corresponding *disjoint* events A_k, $k = 1, 2, \ldots$, alone. Using this observation, one can easily prove again the linearity and multiplicativity properties of the mean value. Another important property is

$$\mathbf{E}\xi_1 \leqslant \mathbf{E}\xi_2 \tag{4.7}$$

for $\xi_1 \leqslant \xi_2$ (i.e., for ξ_1, ξ_2 such that $\xi_1(\omega) \leqslant \xi_2(\omega)$ for every $\omega \in \Omega$). Indeed, let ξ_1, ξ_2 take values $x_{1k} \leqslant x_{2k}$ on A_k, $k = 1, 2, \ldots$, respectively, then from (4.6) we get

$$\mathbf{E}\xi_1 = \sum_{-\infty}^{\infty} x_{1k}\mathbf{P}(A_k) \leqslant \sum_{-\infty}^{\infty} x_{2k}\mathbf{P}(A_k) = \mathbf{E}\xi_2,$$

which also implies

$$|\mathbf{E}\xi| \leqslant \mathbf{E}|\xi|, \tag{4.8}$$

since $-|\xi| \leqslant \xi \leqslant |\xi|$.

Condition $\mathbf{E}|\xi| < \infty$ guarantees the existence of $\mathbf{E}\xi$; one can check that if $|\xi| \leqslant \eta$ and $\mathbf{E}\eta < \infty$, then

$$\mathbf{E}|\xi| \leqslant \mathbf{E}\eta < \infty \tag{4.9}$$

so that $\mathbf{E}\xi$ exists and, of course, $|\mathbf{E}\xi| \leqslant \mathbf{E}\eta$.

Obviously, $\mathbf{E}|\xi| < \infty$ is a restriction on probabilities of large values of $|\xi|$; it implies

$$\mathbf{E}\left(1_{\{|\xi|>a\}}|\xi|\right) = \sum_{|x|>a} |x|\mathbf{P}_\xi(x) \to 0 \tag{4.10}$$

as $a \to \infty$, where the last expectation represents the mean value of $|\xi|$ over all outcomes with $|\xi| > a$.

4.2. LIMIT MEAN VALUES

We know (see (3.24)) that *any* random variable ξ can be written as the limit

$$\xi = \lim_{h \to 0} \xi^h$$

of the corresponding *discrete* approximations

$$\xi^h = kh, \qquad (k-1)h < \xi \leqslant kh, \quad k = 0, \pm 1, \ldots .$$

Assuming the existence of $\mathbf{E}\xi^h$, let us show that the limit

$$\mathbf{E}\xi = \lim_{h \to 0} \mathbf{E}\xi^h \tag{4.11}$$

exists. Indeed, according to (4.3), (4.8),

$$\begin{aligned} &\left|\mathbf{E}\xi^{h_1} - \mathbf{E}\xi^{h_2}\right| = \left|\mathbf{E}\left(\xi^{h_1} - \xi^{h_2}\right)\right| \\ &\leqslant \mathbf{E}\left|\xi^{h_1} - \xi^{h_2}\right| \leqslant 2\max(h_1, h_2) \to 0, \end{aligned}$$

so that the limit (4.11) exists, which is called the *mean value* (*mathematical expectation*) of the (limit) random variable $\xi (= \lim \xi^h)$.

Similarly to (4.3), one can obtain the mean value of a function $\eta = \varphi(\xi_1, \ldots, \xi_n)$ of random variables $\xi_1, \ldots, \xi_n$ with a given probability density,

$$\mathbf{E}\varphi(\xi_1, \ldots, \xi_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \varphi(x_1, \ldots, x_n) p_\xi(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n, \tag{4.12}$$

where

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left|\varphi(x_1, \ldots, x_n)\right| p_{\xi_1, \ldots, \xi_n}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n < \infty.$$

Indeed, (4.12) holds for a *discrete function* φ, taking values y_k on sets B_k, $k = 1, 2, \ldots$, from the corresponding partition

$$\mathbb{R}^n = \bigcup_k B_k,$$

since

$$\begin{aligned}\mathbf{E}\eta &= \sum_k y_k \mathbf{P}\{\eta = y_k\} \\ &= \sum_k y_k \int_{B_k} \cdots \int p_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \sum_k \int_{B_k} \cdots \int \varphi(x_1,\ldots,x_n) p_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \int_{\mathbb{R}^n} \cdots \int \varphi(x_1,\ldots,x_n) p_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n.\end{aligned}$$

Next, writing φ as the *uniform* limit

$$\varphi = \lim_{h\to 0} \varphi^{(h)}$$

of the corresponding *discrete approximations* φ^h,

$$\varphi^h = kh, \qquad (k-1)h < \varphi \leqslant kh, \quad k = 0, \pm 1, \ldots,$$

we obtain

$$\begin{aligned}&\mathbf{E}\varphi(\xi_1,\ldots,\xi_n) = \lim \mathbf{E}\varphi^h(\xi_1,\ldots,\xi_n) \\ &= \lim_{h\to 0} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \varphi^h(x_1,\ldots,x_n) p_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \int_{\mathbb{R}^n} \cdots \int \varphi(x_1,\ldots,x_n) p_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n,\end{aligned}$$

having in mind that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n = 1.$$

For a single random variable, (4.12) gives

$$\mathbf{E}\xi = \int_{-\infty}^{\infty} p_\xi(x)\,\mathrm{d}x.$$

EXAMPLE. If ξ is *uniformly distributed* on $(a, b]$, then

$$\mathbf{E}\xi = \frac{1}{b-a}\int_a^b x\,\mathrm{d}x = \frac{b-a}{2}.$$

EXAMPLE. If ξ is *exponentially distributed* with parameter λ, then

$$\mathbf{E}\xi = \int_0^\infty x\left(\lambda \mathrm{e}^{-\lambda x}\right)\mathrm{d}x = \frac{1}{\lambda}.$$

EXAMPLE. If ξ is *normal*, see (3.15), then

$$\mathbf{E}\xi = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x\mathrm{e}^{-x^2/2\sigma^2}\,\mathrm{d}x = 0,$$

since the probability density is symmetric with respect to $x = 0$.

The limit approach (4.11) helps to prove properties (4.4), (4.8), for general random variables. Actually, (4.5)–(4.5)′ and (4.7)–(4.10) hold as well. Let us verify (4.5), taking $n = 2$ for simplicity. As ξ_1, ξ_2 are *independent*, the corresponding discrete variables ξ_1^h, ξ_2^h are also independent, and

$$\begin{aligned}\mathbf{E}(\xi_1 \cdot \xi_2) &= \lim_{h\to 0}\mathbf{E}\left(\xi_1^h \cdot \xi_2^h\right)\\ &= \lim_{h\to 0}\mathbf{E}\xi_1^h \cdot \mathbf{E}\xi_2^h = \mathbf{E}\xi_1 \cdot \mathbf{E}\xi_2,\end{aligned}$$

since

$$\begin{aligned}&\left|\mathbf{E}(\xi_1\cdot\xi_2) - \mathbf{E}(\xi_1^h\cdot\xi_2^h)\right|\\ &\leqslant \mathbf{E}\left|\xi_1\xi_2 - \xi_1^h\xi_2^h\right|\\ &= \mathbf{E}\left|\xi_1\xi_2 - \xi_1^h\xi_2 + \xi_1^h\xi_2 - \xi_1^h\xi_2^h\right|\\ &\leqslant \mathbf{E}\left|\xi_1 - \xi_1^h\right|\left|\xi_2\right| + \mathbf{E}\left|\xi_1^h\right|\left|\xi_2 - \xi_2^h\right|\\ &\leqslant h\mathbf{E}\left|\xi_2\right| + h\mathbf{E}\left(\left|\xi_1\right| + h\right) \to 0.\end{aligned}$$

The multiplicative formula (4.5)′ for *independent* variables can be obtained in a similar way, by using discrete approximations φ^h of the corresponding functions φ. Next, if $\xi_1 \leqslant \xi_2$, then $\xi_1^h \leqslant \xi_2^h + h$ and

$$\mathbf{E}\xi_1 = \lim_{h\to 0}\mathbf{E}\xi_1^h \leqslant \lim_{h\to 0}\left[\mathbf{E}\xi_2^h + h\right] = \mathbf{E}\xi_2$$

which proves (4.7); (4.8)–(4.10) follow in a similar way. □

In addition to (4.10), it is worthwhile to mention that

$$\mathbf{E}\xi = \lim_{a\to\infty} \mathbf{E}1_{\{|\xi|\leqslant a\}}\xi \tag{4.10$'$}$$

provided $\mathbf{E}\xi$ exists and is finite; and the latter can be justified with the help of

$$\lim_{a\to\infty} \mathbf{E}1_{\{|\xi|\leqslant a\}}|\xi| \longrightarrow \mathbf{E}|\xi| < \infty.$$

Note here that, according to (4.1), (4.11), $\mathbf{E}\xi$ is well-defined for any random variable $\xi \geqslant 0$, although it can be infinite (in particular, this concerns the absolute value $|\xi| \geqslant 0$ of arbitrary random variable ξ).

We call random variables ξ, $\widetilde{\xi}$ *equivalent* if $\xi = \widetilde{\xi}$ with probability 1, i.e.,

$$\mathbf{P}\{\xi - \widetilde{\xi} = 0\} = 1.$$

Obviously, for equivalent ξ, $\widetilde{\xi}$, we have

$$\mathbf{E}(\xi - \widetilde{\xi}) = \mathbf{E}\xi - \mathbf{E}\widetilde{\xi} = 0.$$

Hence, for example, ξ has a finite mean value $\mathbf{E}\xi$, if there is a *majorant* $\eta \geqslant 0$, $\mathbf{E}\eta < \infty$, such that

$$|\xi| \leqslant \eta$$

with probability 1. Namely, consider an equivalent variable $\widetilde{\xi}$ such that $|\widetilde{\xi}| \leqslant \eta$ for all possible outcomes, then, as $\mathbf{E}|\widetilde{\xi}| \leqslant \mathbf{E}\eta$ and $\mathbf{E}|\widetilde{\xi}| = \mathbf{E}|\xi|$,

$$\mathbf{E}|\xi| \leqslant \mathbf{E}\eta. \tag{4.7$'$}$$

4.3. SOME LIMIT PROPERTIES

It is clear that if a random variable

$$\xi = \lim_{n\to\infty} \xi_n$$

is a *uniform* limit of ξ_n, $\mathbf{E}|\xi_n| < \infty$, i.e.,

$$|\xi - \xi_n| \leqslant h_n \longrightarrow 0, \quad n \to \infty,$$

then the limit

$$\lim_{n\to\infty} \mathbf{E}\xi_n = \mathbf{E}\xi$$

exists, which gives the mathematical expectation of ξ.

Moreover, in such a case we have

$$\mathbf{E}|\xi_n - \xi| \longrightarrow 0, \tag{4.13}$$

or the *convergence in mean*, which implies, of course,

$$\mathbf{E}\xi_n \longrightarrow \mathbf{E}\xi,$$

since

$$|\mathbf{E}\xi_n - \mathbf{E}\xi| \leqslant \mathbf{E}|\xi_n - \xi|.$$

Suppose we only know that (4.13) holds. How can one estimate the distance between the random variables ξ_n and ξ? How can one be sure that $|\xi_n - \xi| \leqslant \varepsilon$, or how small is the probability of $|\xi_n - \xi| > \varepsilon$? The answer is given by the inequality

$$\mathbf{P}\{|\xi_n - \xi| > \varepsilon\} \leqslant \frac{1}{\varepsilon}\mathbf{E}|\xi_n - \xi|,$$

which is a particular case of the *Chebyshev inequality*:

$$\mathbf{P}\{|\eta| > \varepsilon\} \leqslant \frac{1}{\varepsilon}\mathbf{E}|\eta| \tag{4.14}$$

valid for any random variable η and any constant $\varepsilon > 0$. (4.14) is clear from

$$\mathbf{E}\varphi(\eta) = \varepsilon\mathbf{P}\{|\eta| > \varepsilon\} \leqslant \mathbf{E}|\eta|,$$

where

$$\varphi(\eta) = \begin{Bmatrix} 0, & |\eta| \leqslant \varepsilon \\ \varepsilon, & |\eta| > \varepsilon \end{Bmatrix} \leqslant |\eta|.$$

For bounded random variables $|\eta| \leqslant a$, there is some kind of converse of (4.14), namely,

$$\mathbf{E}|\eta| \leqslant a\mathbf{P}\{|\eta| > \varepsilon\} + \varepsilon, \tag{4.15}$$

which follows from

$$\mathbf{E}\varphi(\eta) = \varepsilon\mathbf{P}\{|\eta| \leqslant \varepsilon\} + a\mathbf{P}\{|\eta| > \varepsilon\} \geqslant \mathbf{E}|\eta|,$$

with

$$\varphi(\eta) = \left\{ \begin{matrix} \varepsilon, & |\eta| \leqslant \varepsilon \\ a, & |\eta| > \varepsilon \end{matrix} \right\} \geqslant |\eta|.$$

We see that convergence in mean implies *convergence in probability*:

$$\mathbf{P}\{|\xi_n - \xi| > \varepsilon\} \longrightarrow 0 \tag{4.16}$$

for any $\varepsilon > 0$, as $n \to \infty$. Moreover, the two types of convergence are equivalent to each other in the case of *bounded* random variables $|\xi_n|$, $|\xi| \leqslant a$. Indeed, for any $\varepsilon > 0$,

$$\mathbf{E}|\xi_n - \xi| \leqslant \varepsilon + 2a\mathbf{P}\{|\xi_n - \xi| > \varepsilon\}$$

according to (4.15), where $\mathbf{P}\{|\xi_n - \xi| > \varepsilon\} \to 0$ because of the convergence in probability. □

Consider the random variables ξ_n, ξ as functions $\xi_n(\omega), \xi(\omega)$ of elementary outcome $\omega \in \Omega$. The set of outcomes $\omega \in \Omega$ such that $\xi_n(\omega) \to \xi(\omega)$ can be written as

$$\begin{aligned} A &= \{\omega\colon\ \xi_n(\omega) \to \xi(\omega)\} \\ &= \bigcap_r \bigcup_m \bigcap_{n \geqslant m} \left\{\omega\colon\ |\xi_n(\omega) - \xi(\omega)| \leqslant \frac{1}{r}\right\}, \end{aligned}$$

where, by definition, the set on the right-hand side consists of $\omega \in \Omega$ such that, for any $r = 1, 2, \ldots$, there is $m = 1, 2, \ldots$ such that

$$|\xi_n(\omega) - \xi(\omega)| \leqslant \frac{1}{r}, \quad n \geqslant m.$$

Thus, we have the convergence $\xi_n(\omega) \to \xi(\omega)$ with the probability $P(A)$, which, for $\mathbf{P}(A) = 1$, gives the *convergence with probability* 1.

LEMMA. *Convergence with probability* 1 *is equivalent to the condition*: *for any* $\varepsilon > 0$

$$\mathbf{P}\left\{ \sup_{n \geqslant m} |\xi_n - \xi| > \right\} \varepsilon \longrightarrow 0 \quad (m \to \infty). \tag{4.17}$$

Note that condition (4.17) implies (4.16).

Proof. It suffices to take $\varepsilon = 1/r$, $r = 1, 2, \ldots$, in (4.17). The complementary event to the event

$$A = \{\omega\colon\ \xi_n(\omega) \to \xi(\omega)\}$$

is

$$B = \bigcup_r \bigcap_m B_{rm},$$

where

$$B_{rm} = \bigcup_{n \geqslant m} \left\{ |\xi_n - \xi| > \frac{1}{r} \right\} = \left\{ \sup_{n \geqslant m} |\xi_n - \xi| > \frac{1}{r} \right\}.$$

As B_{rm}, $m = 1, 2, \ldots$, are *decreasing*, from (4.17) we obtain

$$\lim_{m \to \infty} \mathbf{P}(B_{rm}) = \mathbf{P}(B_r) = 0, \qquad B_r = \bigcap_r B_{rm}.$$

Obviously, B_r *increase* with $r = 1, 2, \ldots$, hence $\mathbf{P}(B_r) = 0$ for all r is equivalent to

$$\mathbf{P}(B) = \lim_{r \to \infty} \mathbf{P}(B_r) = \sup_r \mathbf{P}(B_r) = 0.$$

EXAMPLE. Suppose, for any $\varepsilon > 0$,

$$\sum_n \mathbf{P}\{|\xi_n - \xi| > \varepsilon\} < \infty. \tag{4.18}$$

Then

$$\mathbf{P}\Big\{\sup_{n\geqslant m}|\xi_n-\xi|>\varepsilon\Big\}=\mathbf{P}\Big(\bigcup_{n\geqslant m}\{|\xi_n-\xi|>\varepsilon\}\Big)$$
$$\leqslant\sum_{n\geqslant m}\mathbf{P}\{|\xi_n-\xi|>\varepsilon\}\longrightarrow 0,\qquad m\to\infty,$$

which shows that $\xi_n\to\xi$ with probability 1. This conclusion is very much consistent with the first Borel–Cantelli lemma, which says that, under the condition (4.18), only a finite number of the events $\{|\xi_n-\xi|>\varepsilon\}$ occur. In other words, starting with some $n_\varepsilon=n_\varepsilon(\omega)$, which depends on $\omega\in\Omega$, for $n\geqslant n_\varepsilon(\omega)$ we have the inequality $|\xi_n-\xi|\leqslant\varepsilon$. □

Now we are ready to prove the following result.

THEOREM. *Suppose,* $\xi_n\to\xi$ *with probability* 1 *and* $|\xi_n|\leqslant\eta$ *for some* (*majorant*) η, $\mathbf{E}\eta<\infty$, *then*

$$\lim_{n\to\infty}\mathbf{E}\xi_n=\mathbf{E}\xi.$$

Proof. We have

$$\lim_{n\to\infty}|\xi_n(\omega)|=|\xi(\omega)|\leqslant\eta(\omega)$$

with probability 1, in particular, $\mathbf{E}\xi$ exists. Then, we can define bounded variables

$$\xi_n'=1_{\{\eta\leqslant a\}}\xi_n,\qquad \xi'=1_{\{\eta\leqslant a\}}\xi,$$

$|\xi_n'|,|\xi'|\leqslant a$, such that $\xi_n'\to\xi'$ with probability 1 and, consequently,

$$\mathbf{E}|\xi_n'-\xi'|\longrightarrow 0.$$

Clearly,

$$\mathbf{E}|\xi_n-\xi|=\mathbf{E}|\xi_n'-\xi'|+\mathbf{E}\big(1_{\{\eta>a\}}|\xi_n-\xi|\big)$$
$$\leqslant\mathbf{E}|\xi_n'-\xi'|+2\mathbf{E}\big(1_{\{\eta>a\}}\eta\big),$$

where

$$\mathbf{E}\big(1_{\{\eta>a\}}\eta\big) \longrightarrow 0$$

as $a \to \infty$; see (4.10). Hence, we conclude that

$$\mathbf{E}\big|\xi_n - \xi\big| \longrightarrow 0.$$

Next, consider the following situation. Suppose, we deal with *increasing* random variables $\xi_n \geqslant 0$, $\mathbf{E}\xi_n \leqslant C$, $n = 1, 2, \ldots$. Then, there is the limit

$$\lim_{n\to\infty} \xi_n = \xi, \quad 0 \leqslant \xi \leqslant \infty.$$

Since $\{\xi_n > x\}$, $n = 1, 2, \ldots$, *increase*,

$$\mathbf{P}\{\xi > x\} = \lim_{n\to\infty} \mathbf{P}\{\xi_n > x\} \leqslant \lim_{n\to\infty} \frac{1}{x}\,\mathbf{E}\xi_n \leqslant \frac{C}{x} \longrightarrow 0$$

as $x \to \infty$, in particular,

$$\mathbf{P}\{\xi < \infty\} = 1 - \lim_{x\to\infty} \mathbf{P}\{\xi > x\} = 1.$$

As $0 \leqslant \xi < \infty$, so $\mathbf{E}\xi$ is well defined although, possibly, $\mathbf{E}\xi = \infty$, as $\xi \geqslant \xi_n$ with probability 1 and

$$\mathbf{E}\xi \geqslant \mathbf{E}\xi_n, \qquad \mathbf{E}\xi \geqslant \lim_{n\to\infty} \mathbf{E}\xi_n.$$

Actually, $\mathbf{E}\xi < \infty$ and, moreover,

$$\mathbf{E}\xi = \lim_{n\to\infty} \mathbf{E}\xi_n. \tag{4.19}$$

Indeed, with probability 1

$$\xi_n 1_{\{\xi<a\}} \longrightarrow \xi 1_{\{\xi<a\}}, \qquad \big|\xi_n 1_{\{\xi<a\}}\big| \leqslant a,$$

which implies

$$\lim_{n\to\infty} \mathbf{E}\xi_n \geqslant \lim_{n\to\infty} \mathbf{E}\xi_n 1_{\{\xi<a\}} = \mathbf{E}\xi 1_{\{\xi<a\}}$$

for any $a > 0$, so that

$$\mathbf{E}\xi = \lim_{a\to\infty} \mathbf{E}\xi 1_{\{\xi<a\}} \leqslant \lim_{n\to\infty} \mathbf{E}\xi_n,$$

which yields (4.19). We obtain the following result:

For any increasing sequence $\xi_n \geqslant 0$, $\mathbf{E}\xi_n \leqslant C$, $n = 1, 2, \ldots$, *of bounded random variables, with probability* 1 *there exists the limit random variable*

$$\xi = \lim_{n\to\infty} \xi_n$$

and

$$\mathbf{E}\xi = \lim_{n\to\infty} \mathbf{E}\xi_n.$$

4.4. CONDITIONAL EXPECTATION

Let ξ and $\eta_1, \ldots, \eta_n$ be discrete random variables. Consider the conditional distribution

$$\mathbf{P}_\xi(x \mid y_1, \ldots, y_n) = \frac{\mathbf{P}_{\xi,\eta_1,\ldots,\eta_n}(x, y_1, \ldots, y_n)}{\mathbf{P}_{\eta_1,\ldots,\eta_n}(y_1, \ldots, y_n)}, \quad -\infty < x < \infty,$$

of ξ given $\eta_1 = y_1, \ldots, \eta_n = y_n$, then we can define the corresponding *conditional expectation*

$$\mathbf{E}(\xi \mid y_1, \ldots, y_n) = \sum_{-\infty}^{\infty} x\mathbf{P}_\xi(x \mid y_1, \ldots, y_n). \tag{4.20}$$

One can easily verify that the following *total expectation formula*

$$\mathbf{E}\xi = \sum_{-\infty}^{\infty} \cdots \sum_{-\infty}^{\infty} \mathbf{E}(\xi \mid y_1, \ldots, y_n) \times \mathbf{P}_{\eta_1,\ldots,\eta_n}(y_1, \ldots, y_n) \tag{4.21}$$

holds, giving $\mathbf{E}\xi$ as the mean value of $\varphi(\eta_1, \ldots, \eta_n) = \mathbf{E}(\xi \mid \eta_1, \ldots, \eta_n)$.

EXAMPLE (*Exit time*). Consider the symmetric random walk with stopping barriers at points $a > 0 > b$, see p. 29. Let τ be the first time when the particle comes either to a, or to b; what is the mean value $\mathbf{E}\tau$? To find $\mathbf{E}\tau$, consider it as a function $\varphi(x)$ of the initial point x of the random walk. Using the total mean formula, we obtain the equation

$$\frac{1}{2}\left[\varphi(x+1) - 2\varphi(x) + \varphi(x-1)\right] = 1, \qquad a > x > b,$$

$$\varphi(a) = \varphi(b) = 0.$$

Indeed, after the first step the particle comes either to $x+1$, or to $x-1$ (with equal probability $p = q = 1/2$), using one time unit to do it, hence

$$\varphi(x) - 1 = \frac{1}{2}\,\varphi(x+1) + \frac{1}{2}\,\varphi(x-1).$$

The same equation can be obtained in a more formal way, by introducing a random variable $\eta = \pm 1$ which measures the first step, and then using the total mean formula

$$\mathbf{E}\tau = \mathbf{E}\big[\mathbf{E}(\tau \mid \eta)\big] = p\mathbf{E}(\tau \mid 1) + q\mathbf{E}(\tau \mid -1), \qquad p = q = 1/2,$$

with

$$\mathbf{E}\tau = \varphi(x), \qquad \mathbf{E}(\tau \mid 1) = 1 + \varphi(x+1), \qquad \mathbf{E}(\tau \mid -1) = 1 + \varphi(x-1).$$

One can verify that the solution of our equation is a quadratic polynomial, which is uniquely determined by the boundary conditions, namely

$$\varphi(x) = (a - x)(x - b), \quad a \geqslant x \geqslant b.$$

In particular, taking $x = 0$ as the initial point, we obtain

$$\mathbf{E}\tau = -ab, \quad a \geqslant 0 \geqslant b.$$

Note that $\mathbf{E}\tau \to \infty$ when $b \to -\infty$, which reflects the fact that for the symmetric random walk with a *single barrier*,

$$\mathbf{E}\tau = \infty. \tag{4.22}$$

□

Consider now ξ and $\eta_1, \ldots, \eta_n$ having a joint probability density. Then we can define the conditional probability density

$$p_\xi(x \mid y_1, \ldots, y_n) = \frac{p_{\xi,\eta_1,\ldots,\eta_n}(x, y_1, \ldots, y_n)}{p_{\eta_1,\ldots,\eta_n}(y_1, \ldots, y_n)}, \quad -\infty < x < \infty,$$

and the corresponding *conditional mean value*

$$\mathbf{E}(\xi \mid y_1, \ldots, y_n) = \int_{-\infty}^{\infty} x p_\xi(x \mid y_1, \ldots, y_n)\, \mathrm{d}x \tag{4.23}$$

of ξ given $\eta_1 = y_1, \ldots, \eta_n = y_n$. One can easily verify that the following *total mean value* formula

$$\mathbf{E}\xi = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{E}(\xi \mid y_1, \ldots, y_n) p_\eta(y_1, \ldots, y_n)\, \mathrm{d}y_1 \ldots \mathrm{d}y_n \tag{4.24}$$

holds, which gives $\mathbf{E}\xi$ as the mean value of the function $\varphi(\eta_1, \ldots, \eta_n) = \mathbf{E}(\xi \mid \eta_1, \ldots, \eta_n)$ of $\eta_1, \ldots, \eta_n$ given in (4.23).

5. Correlation

5.1. VARIANCE AND CORRELATION

Here, we consider random variables ξ with $\mathbf{E}|\xi|^2 < \infty$. For any such ξ_1, ξ_2 there is $\mathbf{E}(\xi_1 \cdot \xi_2)$ since $|\xi_1 \cdot \xi_2|$ is dominated by $\eta = (|\xi_1|^2 + |\xi_2|^2)/2$ with $\mathbf{E}\eta < \infty$. In particular, for any ξ, $\mathbf{E}|\xi|^2 < \infty$, the mean value $\mathbf{E}\xi$ is finite (this follows from above with $\xi_1 = \xi$ and $\xi_2 = 1$). The quantity

$$\mathbf{D}\xi = \mathbf{E}|\xi - a|^2, \quad a = \mathbf{E}\xi, \tag{5.1}$$

is called the *variance* of the random variable ξ. In our earlier discussion, see (3.12), we have already met the *variance* $\sigma^2 = \mathbf{E}\xi^2$ as the parameter of a *normal distribution* with *zero* mean $\mathbf{E}\xi = 0$.

For ξ_1, ξ_2 with $\mathbf{E}\xi_1 = a_1$, $\mathbf{E}\xi_2 = a_2$, one can define the *correlation coefficient*

$$r = \frac{\mathbf{E}(\xi_1 - a_1)(\xi_2 - a_2)}{\sqrt{\mathbf{D}\xi_1} \cdot \sqrt{\mathbf{D}\xi_2}}. \tag{5.2}$$

Let random variables $\xi_0 = 1,\ \xi_1, \ldots, \xi_n$ be given; their linear combinations

$$\xi = \sum_{k=0}^{n} c_k \xi_k$$

form a *linear* space H. The *positive bilinear* form

$$(\xi, \eta) = \mathbf{E}\xi\eta, \quad \xi, \eta \in H, \tag{5.3}$$

is the *inner product* in H with the corresponding *square mean norm*

$$\|\xi\| = (\xi, \xi)^{1/2} = \left(\mathbf{E}|\xi|^2\right)^{1/2}, \quad \xi \in H, \tag{5.4}$$

and the *square mean* distance

$$\|\xi - \eta\|, \quad \xi, \eta \in H.$$

(Note that $\|\xi\| = 0$ if and only if $\xi = 0$ with probability 1, i.e., is equivalent to $\tilde{\xi} \equiv 0$.) The well known inequality

$$\left|(\xi, \eta)\right| \leqslant \|\xi\| \cdot \|\eta\|$$

in our case becomes

$$|\mathbf{E}\xi\eta| \leqslant \left(\mathbf{E}|\xi|^2\right)^{1/2} \left(\mathbf{E}|\eta|^2\right)^{1/2}. \tag{5.5}$$

With the square mean distance, the *variance*

$$\mathbf{D}\xi = \|\xi - a\|$$

measures the difference between ξ and $a = \mathbf{E}\xi$. For normalized random variables ξ and η with *zero* expectations $\mathbf{E}\xi = \mathbf{E}\eta = 0$ and $\mathbf{D}\xi = \mathbf{D}\eta = 1$, the *correlation* coefficient becomes very simple:

$$r = (\xi, \eta);$$

in particular,

$$-1 \leqslant r \leqslant 1, \tag{5.6}$$

since

$$|r| = |(\xi, \eta)| \leqslant \|\xi\| \cdot \|\eta\| = 1,$$

$r = \pm 1$ if and only if $\xi = \pm\eta$. In general, the correlation coefficient measures the dependence between ξ and η, which are *linearly* dependent in the extreme case $|r| = 1$.

Random variables ξ, η are called *uncorrelated* if their correlation coefficient $r = 0$; for ξ, η with $\mathbf{E}\xi = \mathbf{E}\eta = 0$, this means that elements $\xi, \eta \in H$ are *orthogonal*. For example, *independent* ξ and η are uncorrelated since

$$\mathbf{E}(\xi - \mathbf{E}\xi)(\eta - \mathbf{E}\eta) = \mathbf{E}(\xi - \mathbf{E}\xi) \cdot \mathbf{E}(\eta - \mathbf{E}\eta) = 0.$$

□

In the framework of our *Euclidean* space H, consider the following problem: find the *best forecast* $\widehat{\xi}$ of a random variable $\xi \in H$, as a linear combination

$$\widehat{\xi} = \sum_{k=1}^{m} \widehat{c}_k \eta_k$$

of random variables $\eta_1, \ldots, \eta_m$, which we can observe. More precisely, the corresponding *square mean error* $\|\xi - \widehat{\xi}\|$ has to be *minimal*:

$$\|\xi - \widehat{\xi}\| = \min_{\eta} \|\xi - \eta\|, \tag{5.7}$$

where the minimum is taken over all linear combinations

$$\eta = \sum_{k=1}^{m} c_k \eta_k.$$

Of course, the solution is given by the orthogonal projection of $\xi \in H$ onto the subspace of H consisting of all η's, and, for *orthonormal* $\eta_1, \ldots, \eta_m$, we have

$$\widehat{\xi} = \sum_{k=1}^{m} (\xi, \eta_k)\eta_k. \tag{5.8}$$

EXAMPLE. Consider *independent* measurements $\xi_k = \theta + \Delta_k$, $k = 1, \ldots, n$, of a quantity θ, such that

$$\mathbf{E}\Delta_k = 0, \qquad \mathbf{E}\Delta_k^2 = \sigma^2 < \infty.$$

We want to find the best estimate of the unknown θ given the observations $\xi_1, \ldots, \xi_n$. To solve the problem, one can proceed as follows. First, we exclude θ by changing to

$$\eta_k = \xi_k - \xi_n = \Delta_k - \Delta_n, \quad k = 1, \ldots, n-1.$$

Next, take the projection $\widehat{\Delta}_n$ of Δ_n onto the linear span H_0 of $\eta_1, \ldots, \eta_{n-1}$, and define the estimator

$$\widehat{\theta} = \xi_n - \widehat{\Delta}_n = \theta + (\Delta_n - \widehat{\Delta}_n)$$

of θ, having the *minimal* square mean error

$$\|\Delta_n - \widehat{\Delta}_n\| = \min_{\eta \in H_0} \|\Delta_n - \eta\|.$$

It is easy to verify that the *best linear estimate* $\widehat{\theta}$ defined above, is given by

$$\widehat{\theta} = \frac{1}{n} \sum_{k=1}^{n} \xi_k, \tag{5.9}$$

with

$$\|\widehat{\theta} - \theta\| = \frac{\sigma}{\sqrt{n}}.$$

In general, one can even find a *better non-linear* estimate of θ, by defining $\widehat{\Delta}_n$ as the corresponding *conditional expectation*

$$\widehat{\Delta}_n = \mathbf{E}(\Delta_n \mid \eta_1, \ldots, \eta_{n-1}).$$

For example, in the case of *uniformly distributed* Δ_k, $-a \leqslant \Delta_k \leqslant a$, $k = 1, \ldots, n$, such non-linear estimate is given by

$$\widehat{\theta} = \xi_n - \widehat{\Delta}_n = \frac{\xi_{(1)} + \xi_{(n)}}{2},$$

where

$$\xi_{(1)} = \min(\xi_1, \ldots, \xi_n), \qquad \xi_{(n)} = \max(\xi_1, \ldots, \xi_n).$$

We leave to the reader to verify that in such a case,

$$\|\widehat{\theta} - \theta\| = ?$$

5.2. NORMAL CORRELATIONS

Consider $\xi_0 \equiv 1$ and independent *normal* (*Gaussian*) random variables $\xi_1, \ldots, \xi_n$ with $\mathbf{E}\xi_k = 0$ and $\mathbf{E}\xi_k^2 = \sigma^2 = 1$, $k = 1, \ldots, n$, see (3.14), (3.15). The variables $\xi_0, \xi_1, \ldots, \xi_n$ form an orthonormal basis in the space H of linear combinations

$$\eta = \sum_{k=0}^{n} c_k \xi_k,$$

since

$$(\xi_i, \xi_j) = \mathbf{E}\xi_i \xi_j = \begin{cases} \sigma^2 = 1, & i = j, \\ \mathbf{E}\xi_i \mathbf{E}\xi_j = 0, & i \neq j. \end{cases}$$

According to (3.15), the joint probability density of $\xi_1, \ldots, \xi_n$ is

$$p_{\xi_1, \ldots, \xi_n}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{k=1}^{n} x_k^2 \right\}, \quad -\infty < x_1, \ldots, x_n < \infty. \tag{5.10}$$

Consider a *linear* transformation

$$\eta_k = a_k + \sum_{j=1}^{n} \sigma_{kj} \xi_j, \quad k = 1, \ldots, n,$$

where $\sigma = \{\sigma_{kj}\}$ is a non-degenerate matrix. Obviously,

$$\mathbf{E}\eta_k = a_k, \quad k = 1, \ldots, n. \tag{5.11}$$

The mapping $\mathbb{R}^n \to \mathbb{R}^n$,

$$y_k = a_k + \sum_{j=1}^{n} \sigma_{kj} x_j, \quad k = 1, \ldots, n,$$

with the determinant $|J| = |\sigma|$, maps the quadratic form

$$\sum_{k=1}^{n} x_k^2$$

into

$$\sum_{i,j=1}^{n} b_{ij}(y_i - a_i)(y_j - a_j),$$

where $\{b_{ij}\} = B^{-1}$ is the inverse of the product-matrix $B = \sigma \cdot \sigma^*$, with entries

$$B_{ij} = \sum_{k=1}^{n} \sigma_{ik}\sigma_{jk};$$

moreover,

$$B_{ij} = \mathbf{E}(\eta_i - a_i)(\eta_j - a_j), \quad i, j = 1, \ldots, n, \tag{5.12}$$

since

$$\mathbf{E}(\eta_i - a_i)(\eta_j - a_j) = \sum_{k=1}^{n}\sum_{l=1}^{n} \sigma_{ik}\sigma_{jl}\mathbf{E}\xi_k\xi_l = \sum_{k=1}^{n} \sigma_{ik}\sigma_{jk}.$$

$B = \{B_{ij}\}$ is called the *covariance matrix* of $\eta_1, \ldots, \eta_n$. The determinant $|B|$ of $B = \sigma\sigma^*$ (σ^* is the transposed matrix) is

$$|B| = |\sigma|^2 = |J|^2.$$

Hence, according to the general formula (3.22), we obtain the probability density of $\eta_1, \ldots, \eta_n$:

$$p_{\eta_1,\ldots,\eta_n}(y_1, \ldots, y_n) = \frac{1}{|B|^{1/2}(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2}\sum_{j=1}^{n} b_{ij}(y_i - a_i)(y_j - a_j) \right\}, \tag{5.13}$$

where $a_k = \mathbf{E}\eta_k$, $k = 1, \ldots, n$, and

$$\{b_{ij}\} = B^{-1}.$$

The probability density (5.13), as well as random variables $\eta_1, \ldots, \eta_n$ themselves, is called *normal* (or *Gaussian*).

One can immediately see that if normal variables $\eta_1, \ldots, \eta_n$ are *uncorrelated*,

$$B_{ij} = \begin{cases} B_{ii} = \sigma_i^2, & i = j, \\ 0, & i \neq j, \end{cases}$$

then

$$p_{\eta_1,\ldots,\eta_n}(y_1, \ldots, y_n) = \frac{1}{\left(\prod_{k=1}^{n} \sigma_k\right)(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2}\sum_{k=1}^{n} (y_i - a_i)^2/2\sigma_i^2 \right\} \tag{5.14}$$

is the product of probability densities

$$p_\eta(y) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left\{ -\frac{1}{2}(y - a)^2/2\sigma^2 \right\}$$

of $\eta = \eta_k$, with $a = a_k$, $\sigma^2 = \sigma_k^2$ $(k = 1, \ldots, n)$, i.e., $\eta_1, \ldots, \eta_n$ are *independent*.

We conclude here with the simple remark that any non-degenerate *linear* mapping $\mathbb{R}^n \to \mathbb{R}^n$ maps *normal* variables into normal variables again.

5.3. PROPERTIES OF THE VARIANCE AND THE LAW OF LARGE NUMBERS

By the definition of the variance,

$$\mathbf{D}\xi = \mathbf{E}\xi^2 - (\mathbf{E}\xi)^2.$$

In particular, the variance does not depend on a constant shift:

$$\mathbf{D}(\xi + c) = \mathbf{D}\xi,$$

and

$$\mathbf{D}(c\xi) = c^2\mathbf{D}\xi,$$

where c is a constant. If random variables ξ_1, ξ_2 are uncorrelated, then

$$\mathbf{D}(\xi_1 + \xi_2) = \mathbf{D}\xi_1 + \mathbf{D}\xi_2, \tag{5.15}$$

since for $\mathbf{E}\xi_1 = \mathbf{E}\xi_2 = 0$, say,

$$\mathbf{E}(\xi_1 + \xi_2)^2 = \mathbf{E}\xi_1^2 + 2\mathbf{E}\xi_1\xi_2 + \mathbf{E}\xi_2^2,$$

and $\mathbf{E}\xi_1\xi_2 = 0$.

It was mentioned earlier that $\mathbf{D}\xi$ measures the *dispersion* of a random variable ξ around its mean value $a = \mathbf{E}\xi$. One can roughly estimate the corresponding probability

$$\mathbf{P}\{|\xi - a| > \varepsilon\} \leqslant \frac{1}{\varepsilon^2}\mathbf{D}\xi, \tag{5.16}$$

using the *Chebyshev inequality* (4.14). □

Consider the *empirical mean*

$$\frac{1}{n}\sum_{k=1}^{n}\xi_k$$

of *uncorrelated* random variables ξ_k, $k = 1, \ldots, n$,

$$\mathbf{E}\left(\frac{1}{n}\sum_{k=1}^{n}\xi_k\right) = \frac{1}{n}\sum_{k=1}^{n}\mathbf{E}\xi_k = a^{(n)}.$$

Suppose

$$\mathbf{D}\xi_k \leqslant b, \quad k = 1, \ldots, n;$$

then

$$\mathbf{D}\left(\frac{1}{n}\sum_{k=1}^{n}\xi_k\right) = \frac{1}{n^2}\sum_{k=1}^{n}\mathbf{D}\xi_k \leqslant \frac{b}{n} \longrightarrow 0$$

as $n \to \infty$. In particular,

$$\frac{1}{n}\sum_{k=1}^{n}\xi_k - a^{(n)} \longrightarrow 0 \tag{5.17}$$

in probability; the rate of the convergence

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k - a^{(n)}\right| > \varepsilon\right\} \longrightarrow 0, \quad n \to \infty,$$

can be roughly estimated by the Chebyshev inequality:

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k - a^{(n)}\right| > \varepsilon\right\} \leqslant \frac{1}{n}\frac{b}{\varepsilon^2}. \tag{5.18}$$

□

Let random variables ξ_k, $k = 1, \ldots, n$, be independent and *identically* distributed (i.i.d. for short), with

$$\mathbf{E}\xi_k = a, \qquad \mathbf{D}\xi_k = b = \sigma^2, \quad k = 1, \ldots, n.$$

Then (5.17) implies the convergence

$$\frac{1}{n}\sum_{k=1}^{n}\xi_k \longrightarrow a, \quad n \to \infty, \tag{5.19}$$

of the empirical mean to the corresponding mean value. This remarkable phenomenon is known as the *law of large numbers*.

For example, let

$$\xi_k = 1_A, \quad k = 1, \ldots, n,$$

be the indicators of an event $A(= A_k)$ in the corresponding Bernoulli trials, which occurs with the same probability

$$p = \mathbf{P}(A) = \mathbf{E}1_A.$$

Then

$$\sigma^2 = \mathbf{E}1_A^2 - (\mathbf{E}1_A)^2 = p(1-p).$$

Here, the empirical mean gives the *frequency*

$$\frac{n(A)}{n} = \frac{1}{n}\sum_{k=1}^{n} 1_{A_k}$$

of the event A, and the *law of large numbers* says that

$$\frac{n(A)}{n} \longrightarrow \mathbf{P}(A), \quad n \to \infty. \tag{5.20}$$

□

Let us show that in (5.17) and (5.19), we have the convergence with probability 1.

Note first, using the first Borel–Cantelli lemma, that this true for the subsequence $n = m^2$, $m \to \infty$, since for any $\varepsilon > 0$

$$\sum_{m=1}^{\infty}\mathbf{P}\left\{\left|\frac{1}{m^2}\sum_{k=1}^{m^2}\xi_k - a^{(n)}\right| > \varepsilon\right\} \leqslant \frac{b}{\varepsilon^2}\sum_{m=1}^{\infty}\frac{1}{m^2} < \infty$$

according to (5.18).

Assume, for simplicity, that $\mathbf{E}\xi_n = 0$, $k = 1, \ldots, n$, and set

$$\eta_m = \max_{m^2<n<(m+1)^2} \frac{1}{m^2}\left|\sum_{k=m^2+1}^{n} \xi_k\right|.$$

Obviously, for $m^2 < n < (m+1)^2$, we have

$$\frac{1}{n}\left|\sum_{k=1}^{n}\xi_k\right| \leqslant \frac{1}{m^2}\left|\sum_{k=1}^{n}\xi_k\right| \leqslant \frac{1}{m^2}\left|\sum_{k=1}^{m^2}\xi_k\right| + \eta_m$$

where

$$\frac{1}{m^2}\left|\sum_{k=1}^{m^2}\xi_k\right| \longrightarrow 0$$

with probability 1, according to the observation above. Therefore, if suffices to prove that $\eta_m \to 0$ with probability 1. Write

$$\{|\eta_m| > \varepsilon\} = \bigcup_{m^2<n<(m+1)^2} \left\{\frac{1}{m^2}\left|\sum_{k=m^2+1}^{n}\xi_k\right| > \varepsilon\right\},$$

then

$$\begin{aligned}
\mathbf{P}\{|\eta_m| > \varepsilon\} &\leqslant \sum_{m^2<n<(m+1)^2} \mathbf{P}\left\{\frac{1}{m^2}\left|\sum_{k=m^2+1}^{n}\xi_k\right| > \varepsilon\right\} \\
&\leqslant \sum_{m^2<n<(m+1)^2} \frac{1}{(m^2)^2\varepsilon^2}\mathbf{D}\left(\sum_{k=m^2+1}^{n}\xi_k\right) \\
&\leqslant 2m \times \frac{2m}{(m^2)^2}\frac{b}{\varepsilon^2} = \frac{1}{m^2}\frac{4b}{\varepsilon^2}
\end{aligned}$$

according to (5.18), as $n - m^2 \leqslant (m+1)^2 - m^2 - 1 = 2m$. Consequently,

$$\sum_{m=1}^{\infty} \mathbf{P}\left\{|\eta_m| > \varepsilon\right\} \leqslant \frac{4b}{\varepsilon^2}\sum_{m=1}^{\infty}\frac{1}{m^2} < \infty,$$

and the desired convergence follows from the first Borel–Cantelli lemma.

Thus, we obtain the following result.

THEOREM (The law of large numbers). *Let* ξ_k, $k = 1, 2, \ldots$, *be a sequence of uncorrelated random variables,* $\mathbf{D}\xi_k \leqslant b$. *Then*

$$\frac{1}{n}\sum_{k=1}^{n}(\xi_k - \mathbf{E}\xi_k) \longrightarrow 0 \tag{5.21}$$

with probability 1.

6. Characteristic Functions

6.1. SOME EXAMPLES

Let ξ be a random variable taking only integer values $x = k$ with probabilities

$$\mathbf{P}(k) = \mathbf{P}\{\xi = k\}, \quad k = 0, \pm 1, \ldots .$$

The corresponding *Fourier series*

$$f(u) = \sum_{-\infty}^{\infty} \mathbf{P}(k)\mathrm{e}^{iuk}, \quad -\infty < u < \infty, \tag{6.1}$$

defines a function $f(u)$ with the period 2π and $\mathbf{P}(k)$ as the Fourier coefficients

$$\mathbf{P}(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-iuk} f(u)\,\mathrm{d}u, \quad k = 0, \pm 1, \ldots . \tag{6.2}$$

EXAMPLE (*Binomial distribution*). For

$$\mathbf{P}(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n,$$

we have

$$f(u) = \sum_{k=0}^{n} \binom{n}{k} \left(p\mathrm{e}^{iu}\right)^k q^{n-k} = \left(p\mathrm{e}^{iu} + q\right)^n, \quad q = 1 - p. \tag{6.3}$$

EXAMPLE (*Poisson distribution*). For

$$\mathbf{P}(k) = \frac{a^k}{k!}\mathrm{e}^{-a}, \quad k = 0, 1, \ldots,$$

we have

$$f(u) = \sum_{k=0}^{\infty} \frac{(ae^{iu})^k}{k!} e^{-a} = e^{a(e^{iu}-1)}. \tag{6.4}$$

□

Next, let ξ be a random variable with probability density $p(x)$, $-\infty < x < \infty$. The *Fourier integral*

$$f(u) = \int_{-\infty}^{\infty} e^{iux} p(x)\,dx, \quad -\infty < u < \infty, \tag{6.5}$$

uniquely determines the function $p(x)$, $-\infty < x < \infty$; in particular, if $f(u)$, $-\infty < u < \infty$, is integrable, then $p(x)$ is given by the *inverse Fourier transform*

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iux} f(u)\,du. \tag{6.6}$$

EXAMPLE (*Normal distribution*). For the probability density

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty,$$

the integral

$$\int_{-\infty}^{\infty} e^{zx} p(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{zx - x^2/2}\,dx$$

exists for all complex z and is an analytic function, which for real z coincides with

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{zx - x^2/2}\,dx = e^{z^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x+z)^2/2}\,d(x+z) = e^{z^2/2}.$$

Therefore, it coincides with the analytic function $e^{z^2/2}$ for all *complex* z; in particular, for $z = e^{iu}$ we have

$$f(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{iux - x^2/2}\,dx = e^{-u^2/2}. \tag{6.7}$$

Using a linear transform we obtain the characteristic function of the normal random variable with mean value a and variance σ^2

$$f(u) = \mathbf{E}e^{iu(a+\sigma\xi)} = e^{iua} f(\sigma u) = e^{iua-(\sigma^2 u^2)/2}, \quad -\infty < u < \infty. \tag{6.7}'$$

Dealing with complex-valued functions, let us introduce complex random variables

$$\xi = \xi' + i\xi''$$

with *real* ξ', ξ'' and $i = \sqrt{-1}$, and the corresponding *mathematical expectation (mean value)*

$$\mathbf{E}\xi = \mathbf{E}\xi' + i\mathbf{E}\xi''$$

which is *linear* and satisfies other properties discussed above. In particular, one can verify the multiplicative formula (4.5)′ for real independent random variables $\xi_1, \ldots, \xi_n$ and *complex* functions $\varphi_1, \ldots, \varphi_n$, namely

$$\begin{aligned}
\mathbf{E}[\varphi_1(\xi_1)\varphi_2(\xi_2)] &= \mathbf{E}(\varphi_1' + i\varphi_1'')(\varphi_2' + i\varphi_2'') \\
&= \mathbf{E}[(\varphi_1'\varphi_2' - \varphi_1''\varphi_2'') + i(\varphi_1''\varphi_2' + \varphi_1'\varphi_2'')] \\
&= (\mathbf{E}\varphi_1' \cdot \mathbf{E}\varphi_2' - \mathbf{E}\varphi_1'' \cdot \mathbf{E}\varphi_2'') + i(\mathbf{E}\varphi_1'' \cdot \mathbf{E}\varphi_2' + \mathbf{E}\varphi_1' \cdot \mathbf{E}\varphi_2'') \\
&= (\mathbf{E}\varphi_1' + i\mathbf{E}\varphi_1')(\mathbf{E}\varphi_2' + i\mathbf{E}\varphi_2'') = \mathbf{E}\varphi_1(\xi_1)\mathbf{E}\varphi_2(\xi_2).
\end{aligned}$$

In accordance with general formulas (4.3), (4.12), we see that (6.2), (6.5) is nothing else but

$$f(u) = \mathbf{E}e^{iu\xi}, \quad -\infty < u < \infty, \tag{6.8}$$

or the mean value of the complex-valued function $\varphi(\xi) = e^{iu\xi}$, which is well-defined for any real random variable ξ since $|e^{iu\xi}| \leqslant 1$, and is called the *characteristic function*

$$f(u) = f_\xi(u), \quad -\infty < u < \infty,$$

of ξ. □

According to the multiplicative property, for any *independent* (real) random variables $\xi_1, \ldots, \xi_n$ we have

$$\mathbf{E}e^{iu(\xi_1+\cdots+\xi_n)} = \mathbf{E}e^{iu\xi_1} \ldots \mathbf{E}e^{iu\xi_n}.$$

In other words, *the characteristic function of the sum* $\xi = \xi_1 + \cdots + \xi_n$ *of independent random variables is the product of the characteristic functions of the summands*:

$$f_\xi(u) = f_{\xi_1}(u) \cdots f_{\xi_n}(u), \quad -\infty < u < \infty. \tag{6.9}$$

EXAMPLE (*Triangular distribution*). It is known that a triangular distribution corresponds to the sum $\xi = \xi_1 + \xi_2$ of independent uniformly distributed random variables $-a < \xi_1 \leqslant 0$ and $0 < \xi_2 \leqslant a$. Clearly,

$$f_{\xi_2}(u) = f_{\xi_1}(-u) = \overline{f_{\xi_1}(u)} = \frac{1}{a}\int_0^a e^{iux}\,dx = \frac{e^{iua}-1}{iua},$$

so for the triangular probability density

$$p(x) = \frac{1}{a}\left(1 - \frac{|x|}{a}\right), \quad -a < x < a,$$

we obtain

$$\begin{aligned} f_\xi(u) &= \frac{1}{a}\int_{-a}^a e^{iux}\left(1 - \frac{|x|}{a}\right)dx \\ &= \left|\frac{e^{iua}-1}{iua}\right|^2 = \left(\frac{\sin\frac{1}{2}au}{\frac{1}{2}au}\right)^2. \end{aligned} \tag{6.10}$$

EXAMPLE (*Chi-square distribution*). A chi-square distribution is given by the probability density

$$p(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)}\, x^{(n/2)-1}e^{-(x/2)}, & x > 0, \\ 0, & x \leqslant 0, \end{cases} \tag{6.11}$$

where

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x}\, dx$$

is the gamma-function and the integer n is called the degree of freedom. For $n = 1$, $p(x)$ is the probability density of the square ξ^2 of a standard *normal* variable ξ with $\mathbf{E}\xi = 0$, $\mathbf{D}\xi = 1$. Let us show that for any $n \geqslant 1$, (6.11) is the probability density of the sum

$$\sum_{k=1}^{n} \xi_k^2$$

of squares of *independent* standard normal variables ξ_k, $k = 1, \ldots, n$. Consider the integral

$$f(u) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty x^{(n/2)-1} e^{-((1/2)-iu)x}\, dx$$

which is an analytic function of the complex variable u in the upper half plane $\operatorname{Im} u > -1/2$. On the half line

$$1/2 - iu = \lambda > 0$$

it coincides with

$$\begin{aligned} f(u) &= \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty x^{(n/2)-1} e^{-\lambda x}\, dx \\ &= \frac{1}{2^{n/2}} \lambda^{-n/2} \times \frac{1}{\Gamma(n/2)} \int_0^\infty x^{(n/2)-1} e^{-x}\, dx \\ &= \frac{1}{2^{n/2}} \lambda^{-n/2} \frac{\Gamma(n/2)}{\Gamma(n/2)} = (1 - 2iu)^{-n/2}. \end{aligned}$$

Therefore, $f(u)$ coincides with the analytic function $(1 - 2iu)^{-n/2}$ for all u, $\operatorname{Im} u > -1/2$. By taking u real, $-\infty < u < \infty$, we find that the corresponding characteristic function is

$$f(u) = (1 - 2iu)^{-n/2}. \tag{6.12}$$

As $(1 - 2iu)^{-1/2}$ is the characteristic function of the square of a standard normal variable, from (6.12) we obtain that $f(u)$ is the characteristic function of the sum of n squares of independent standard normal variables.

6.2. ELEMENTARY ANALYSIS OF CHARACTERISTIC FUNCTIONS

Note that any characteristic function

$$f(u) = \mathbf{E}e^{iu\xi}, \quad -\infty < u < \infty,$$

is *continuous*, since the convergence

$$e^{i(u+h)\xi} \longrightarrow e^{iu\xi}, \quad h \to 0,$$

and boundedness of the exponents (ξ is real) implies

$$f(u+h) = \mathbf{E}e^{i(u+h)\xi} \longrightarrow \mathbf{E}e^{iu\xi} = f(u).$$

Of course,

$$f(0) = 1. \tag{6.13}$$

Suppose there exist the *moments*

$$a_k = \mathbf{E}\xi^k, \qquad \mathbf{E}|\xi|^k < \infty, \quad k = 1, \ldots, m.$$

Then the convergence

$$\frac{1}{h}\left[(i\xi)^{k-1}e^{i(u+h)\xi} - (i\xi)^{k-1}e^{iu\xi}\right] \longrightarrow (i\xi)^k e^{iu\xi}, \quad h \to 0,$$

together with the bound

$$\left|\frac{1}{h}\left[(i\xi)^{k-1}e^{i(u+h)\xi} - (i\xi)^{k-1}e^{iu\xi}\right]\right| \leqslant \eta = |\xi|^k, \qquad \mathbf{E}\eta < \infty \quad (k = 1, \ldots, m)$$

gives us

$$\frac{1}{h}\left[f^{(k-1)}(u+h) - f^{(k-1)}(u)\right] \longrightarrow \mathbf{E}(i\xi)^k e^{iu\xi} = f^{(k)}(u),$$

$k = 1, \ldots, m$. Moreover,

$$f^{(m)}(u+h) - f^{(m)}(u) = \mathbf{E}\left[(i\xi)^m e^{i(u+h)\xi} - (i\xi)^m e^{iu\xi}\right] \longrightarrow 0, \quad h \to 0,$$

i.e., the m-th derivative $f^{(m)}(u)$ is *continuous*. This leads to the well-known expansion

$$f(u) = \sum_{k=0}^{m} \frac{i^k a_k}{k!} u^k + o(u^m), \tag{6.14}$$

at the point $u = 0$, with

$$f^{(k)}(0) = i^k a_k = \mathbf{E}(i\xi)^k, \quad k = 0, \ldots, m.$$

Moreover, in the case $\mathbf{E}|\xi|^{m+1} < \infty$ we have the following estimate of the remainder term in (6.14):

$$|o(u^m)| \leqslant \frac{\mathbf{E}|\xi|^{m+1}}{(m+1)!} |u|^{m+1},$$

since

$$|f^{(m+1)}(u)| = |\mathbf{E}(i\xi)^{m+1} e^{iu\xi}| \leqslant \mathbf{E}|\xi|^{m+1}, \quad -\infty < u < \infty.$$

EXAMPLE (*Moments of a normal distribution*). For a normal variable ξ with

$$a_1 = \mathbf{E}\xi = 0, \qquad a_2 = \mathbf{E}\xi^2 = \sigma^2,$$

we have

$$a_3 = \mathbf{E}\xi^3 = 0,$$

since the corresponding probability density is symmetric with respect to the origin, and

$$a_4 = 3\sigma^4$$

which follows from the general formula (6.14) with the characteristic function

$$f(u) = e^{-\sigma^2 u^2/2}, \quad -\infty < u < \infty.$$

6.3. THE INVERSE FORMULA OF PROBABILITY DISTRIBUTIONS

Let ξ be a real random variable. Let

$$1_{(x',x'']}(x) = \begin{cases} 1, & x' < x \leqslant x'', \\ 0, & x \leqslant x' \text{ or } x > x'', \end{cases}$$

be the *indicator* of a finite interval $(x', x'']$. Then

$$\mathbf{P}\{x' < \xi \leqslant x''\} = \mathbf{E}1_{(x',x'']}(\xi) = \lim \mathbf{E}\varphi(\xi), \tag{6.15}$$

where $\varphi(x)$ are bounded continuous functions and

$$\varphi(x) \longrightarrow 1_{(x',x'']}(x), \quad -\infty < x < \infty,$$

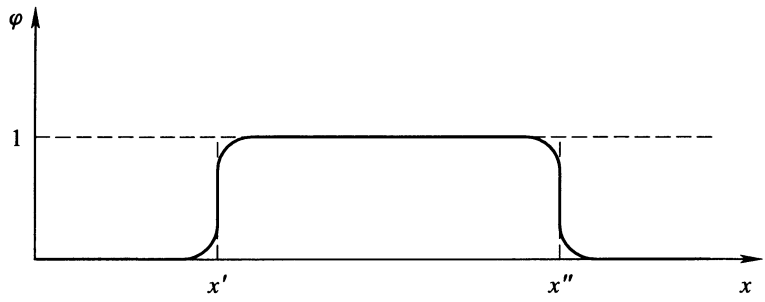

Fig. 3.

as shown in Figure 3. Moreover, we can take $\varphi(x)$ to be *infinitely differentiable* and vanishing for sufficiently large $|x|$, $-\infty < x < \infty$; the class of all such functions is denoted C_0^∞. □

A remarkable property of functions $\varphi \in C_0^\infty$ is that their *Fourier transform*

$$\tilde{\varphi}(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iux} \varphi(x)\, dx, \quad -\infty < x < \infty,$$

is *integrable* which follows, for example, from the well known relationship

$$\begin{aligned}\widetilde{\varphi^{(k)}}(u) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iux}\varphi^{(k)}(x)\,dx \\ &= -\int_{-\infty}^{\infty}(-iu)e^{-iux}\varphi^{(k-1)}(x)\,dx = (iu)\widetilde{\varphi^{(k-1)}}(u) \\ &= \cdots = (iu)^k\tilde{\varphi}(u)\end{aligned}$$

and the boundedness of $\widetilde{\varphi^{(k)}}(u)$, $-\infty < u < \infty$, $k = 1, 2, \ldots$. Applying the *inverse Fourier transform*, we have

$$\varphi(x) = \int_{-\infty}^{\infty} e^{iux}\tilde{\varphi}(u)\,du = \lim_{n\to\infty}\varphi_n(x)$$

as the limit of sums

$$\varphi_n(x) = \sum_{k=1}^{n} e^{iu_{kn}x}\tilde{\varphi}(u_{kn})h_{kn}, \quad -\infty < x < \infty,$$

which are *bounded*, according to

$$|\varphi_n(x)| \leqslant \sum_{k=1}^{n}|\tilde{\varphi}(u_{kn})|h_{kn} \longrightarrow \int_{-\infty}^{\infty}|\tilde{\varphi}(u)|\,du.$$

Hence

$$\begin{aligned}\mathbf{E}\varphi(\xi) &= \lim_{n\to\infty}\mathbf{E}\varphi_n(\xi) \\ &= \lim_{n\to\infty}\sum_{k=1}^{n}\tilde{\varphi}(u_{kn})\left[\mathbf{E}e^{iu_{kn}\xi}\right]h_{kn} \\ &= \lim_{n\to\infty}\sum_{k=1}^{n}\tilde{\varphi}(u_{kn})f_\xi(u_{kn})h_{kn} = \int_{-\infty}^{\infty}\tilde{\varphi}(u)f_\xi(u)\,du,\end{aligned}$$

where $f_\xi(u)$, $-\infty < u < \infty$, is the characteristic function. Thus, we obtain the following result.

THEOREM. *The characteristic function f_ξ uniquely determines the probability distribution of a random variable ξ, by the formula*

$$\mathbf{E}\varphi(\xi) = \int_{-\infty}^{\infty} \widetilde{\varphi}(u) f_\xi(u)\,du, \quad \varphi \in C_0^\infty. \tag{6.16}$$

6.4. WEAK CONVERGENCE OF DISTRIBUTIONS

Let us consider the *weak convergence* $\mathbf{P}_{\xi_n} \Rightarrow \mathbf{P}_\xi$ of the probability distributions of random variables ξ_n, ξ, respectively, in the sense that

$$\mathbf{E}\varphi(\xi_n) \longrightarrow \mathbf{E}\varphi(\xi) \tag{6.17}$$

for every $\varphi \in C_0^\infty$ (recall that C_0^∞ is the space of all infinitely differentiable functions $\varphi(x)$, $-\infty < x < \infty$, which vanish for sufficiently large $|x|$).

Let $f_n = f_{\xi_n}$ and $f = f_\xi$ be the characteristic functions of $\mathbf{P}_n = \mathbf{P}_{\xi_n}$ and $\mathbf{P} = \mathbf{P}_\xi$, respectively.

THEOREM. *The convergence*

$$f_n(u) \longrightarrow f(u), \tag{6.18}$$

which is uniform on any finite interval $u' \leqslant u \leqslant u''$, implies the weak convergence $\mathbf{P}_n \Rightarrow \mathbf{P}$.

Proof. Applying the inverse formula (6.16) with $\varphi \in C_0^\infty$ and using the integrability of $\widetilde{\varphi}$, we obtain

$$\mathbf{E}\varphi(\xi_n) = \int_{-\infty}^{\infty} \widetilde{\varphi}(u) f_{\xi_n}(u)\,du \longrightarrow \int_{-\infty}^{\infty} \widetilde{\varphi}(u) f_\xi(u)\,du = \mathbf{E}\varphi(\xi),$$

since

$$\left| \int_{-\infty}^{\infty} \widetilde{\varphi}(u) f_{\xi_n}(u)\,du - \int_{-\infty}^{\infty} \widetilde{\varphi}(u) f_\xi(u)\,du \right| \leqslant$$

$$\leqslant \int_{-\infty}^{u'} \left|\widetilde{\varphi}(u)\right| du + \max_{u' \leqslant u \leqslant u''} \left|f_n(u) - f(u)\right| \int_{u'}^{u''} \left|\widetilde{\varphi}(u)\right| du + \int_{u''}^{\infty} \left|\widetilde{\varphi}(u)\right| du$$

and

$$\int_{-\infty}^{u'} \left|\widetilde{\varphi}(u)\right| du \longrightarrow 0, \quad u' \to -\infty; \qquad \int_{u''}^{\infty} \left|\widetilde{\varphi}(u)\right| du \longrightarrow 0, \quad u'' \to \infty.$$

7. The Central Limit Theorem

7.1. SOME LIMIT PROPERTIES OF PROBABILITIES

Let us discuss some consequences of the weak convergence $\mathbf{P}_{\xi_n} \Rightarrow \mathbf{P}_{\xi}$.

Consider an interval $(x', x'']$, and a functions $\varphi = \varphi_1$, $\varphi_2 \in C_0^{\infty}$, as shown in Figure 4, such that

$$\varphi_1(x) \leqslant 1_{(x',x'']}(x) \leqslant \varphi_2(x).$$

Then

$$\mathbf{E}\varphi_1(\xi) \leqslant \mathbf{E}1_{(x',x'']}(\xi) \leqslant \mathbf{E}\varphi_2(\xi).$$

Suppose x', x'' satisfy

$$\mathbf{P}\{\xi = x\} = 0, \quad x = x', \ x''. \tag{7.1}$$

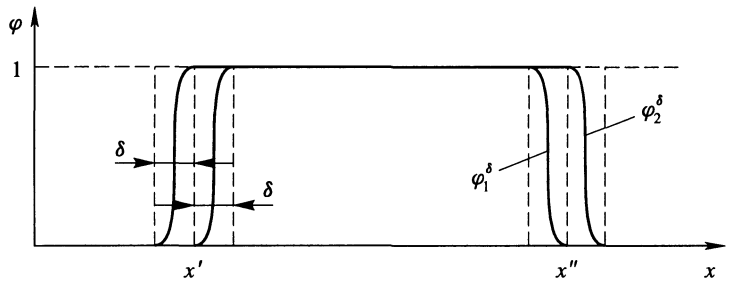

Fig. 4.

Then

$$\begin{aligned}&\left|\mathbf{E}\varphi_2(\xi) - \mathbf{E}\varphi_1(\xi)\right| \leqslant \mathbf{E}\left|\varphi_2(\xi) - \mathbf{E}\varphi_1(\xi)\right| \\ &\leqslant \mathbf{P}\{x' - \delta < \xi < x' + \delta\} + \mathbf{P}\{x'' - \delta < \xi < x'' + \delta\} \longrightarrow 0, \quad \delta \to 0,\end{aligned}$$

according to the continuity property of the probability distribution, and therefore

$$\begin{aligned}&\mathbf{P}\{x' < \xi \leqslant x''\} = \mathbf{E}1_{(x',x'']}(\xi) \\ &= \lim_{\delta \to 0} \mathbf{E}\varphi_1(x) = \lim_{\delta \to 0} \mathbf{E}\varphi_2(x).\end{aligned}$$

The weak convergence $\mathbf{P}_{\xi_n} \Rightarrow \mathbf{P}_\xi$ gives us the diagram

$$\begin{array}{ccc} \mathbf{E}\varphi_1(\xi_n) \leqslant \mathbf{E}1_{(x',x'']}(\xi_n) \leqslant \mathbf{E}\varphi_2(\xi_n) \\ \downarrow \qquad\qquad\qquad \downarrow \\ \mathbf{E}\varphi_1(\xi) \longrightarrow \mathbf{E}1_{(x',x'']}(\xi) \longleftarrow \mathbf{E}\varphi_2(\xi), \end{array}$$

which implies that, in the case of (7.1),

$$\mathbf{P}\{x' < \xi_n \leqslant x''\} = \mathbf{E}1_{(x',x'']}(\xi_n) \longrightarrow \mathbf{E}1_{(x',x'']}(\xi) = \mathbf{P}\{x' < \xi \leqslant x''\}.$$

Thus, we obtain the following result.

THEOREM. *The weak convergence* $\mathbf{P}_{\xi_n} \Rightarrow \mathbf{P}_\xi$ *implies*

$$\mathbf{P}\{x' < \xi_n \leqslant x''\} \longrightarrow \mathbf{P}\{x' < \xi \leqslant x''\} \tag{7.2}$$

for any x', x'' *satisfying* (7.1).

EXAMPLE. Suppose we have the weak convergence $\mathbf{P}_{\xi_n} \Rightarrow \mathbf{P}_\xi$ of integer-valued random variables, taking values $x = k$, $k = 0, \pm 1, \ldots$. Then

$$\mathbf{P}_{\xi_n}(k) = \mathbf{P}\{\xi_n = k\} \longrightarrow \mathbf{P}\{\xi = k\} = \mathbf{P}_\xi(k), \tag{7.3}$$

since (7.2) applies to any x', x'' such that $k - 1 < x' < k < x'' < k + 1$.

EXAMPLE. If ξ has a *probability density*, then the weak convergence $\mathbf{P}_{\xi_n} \Rightarrow \mathbf{P}_\xi$ implies

$$\mathbf{P}\{x' < \xi_n \leqslant x''\} \longrightarrow \int_{x'}^{x''} p_\xi(x)\,\mathrm{d}x \tag{7.4}$$

for any $x' < x''$, since (7.1) holds for any x, $-\infty < x < \infty$. □

Let us define the *weak convergence* of distribution functions

$$F_{\xi_n}(x) = \mathbf{P}\{\xi_n \leqslant x\}, \quad -\infty < x < \infty,$$

as

$$F_{\xi_n}(x) \longrightarrow F_\xi(x) \tag{7.5}$$

for every x, $-\infty < x < \infty$, which is a *continuity point* of the limit function $F_\xi(x)$, i.e., for every x such that

$$F_\xi(x) = F_\xi(x-0) = \lim_{h\to 0} F(x-h).$$

As $F_\xi(x)$ is increasing and right-continuous, $0 \leqslant F_\xi(x) \leqslant 1$, the number of discontinuity points is at most *countable*, since

$$\sum_x \left[F_\xi(x) - F_\xi(x-0)\right] \leqslant 1.$$

THEOREM. *The weak convergence* $\mathbf{P}_{\xi_n} \Rightarrow \mathbf{P}_\xi$ *implies the weak convergence* $F_{\xi_n} \Rightarrow F_\xi$.

Proof. For any n,

$$\mathbf{P}\{|\xi_n| \geqslant a\} \longrightarrow 0, \quad a \to \infty, \tag{7.6}$$

and the convergence is *uniform* in $n = 1, 2, \ldots$ because of $\mathbf{P}_{\xi_n} \Rightarrow \mathbf{P}_\xi$; indeed,

$$\mathbf{P}\{|\xi_n| \geqslant a\} \leqslant 1 - \mathbf{E}\varphi(\xi_n) \longrightarrow 1 - \mathbf{E}\varphi(\xi) \leqslant \mathbf{P}\{|\xi| > a - \delta\},$$

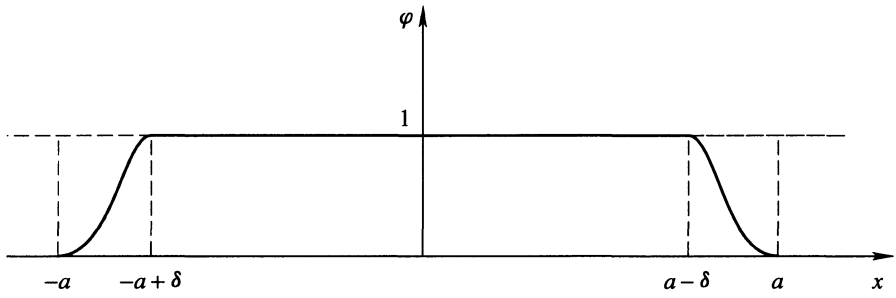

Fig. 5.

where $\varphi \in C_0^\infty$ is a function shown in Figure 5. (The (uniform) convergence of (7.6) is called the *compactness property*.) (7.2) can be written as

$$F_{\xi_n}(x'') - F_{\xi_n}(x') \longrightarrow F_\xi(x'') - F_\xi(x'),$$

x', x'' being continuity points of F_ξ; putting here $x' = -a$, $x'' = x$, with a suitable a, we get

$$F_{\xi_n}(x) - F_{\xi_n}(-a) \longrightarrow F_\xi(x) - F_\xi(a), \quad n \to \infty,$$

where

$$F_{\xi_n}(-a) \leqslant \mathbf{P}\{|\xi_n| \geqslant a\} \leqslant \varepsilon, \qquad F_{\xi}(-a) \leqslant \mathbf{P}\{|\xi| \geqslant a\} \leqslant \varepsilon,$$

$\varepsilon > 0$ being arbitrary small. Therefore, we conclude that

$$F_{\xi_n}(x) \longrightarrow F_{\xi}(x).$$

□

Convergence in probability. Suppose, $\xi_n \to \xi$ in probability; then, at every point x of continuity of the distribution function F_{ξ}, and for any $\varepsilon > 0$, we have

$$\begin{aligned}
&\left|\mathbf{P}\{\xi_n \leqslant x\} - \mathbf{P}\{\xi \leqslant x\}\right| \\
&\quad \leqslant \mathbf{P}\{\xi_n \leqslant x, \xi > x\} + \mathbf{P}\{\xi_n > x, \xi \leqslant x\} \\
&\quad \leqslant \mathbf{P}\{\xi_n \leqslant x, \xi > x + \delta\} + \mathbf{P}\{\xi_n > x, \xi \leqslant x - \delta\} + \varepsilon \\
&\quad \leqslant \mathbf{P}\{|\xi_n - \xi| > \delta\} + \varepsilon
\end{aligned}$$

provided $\delta > 0$ was chosen sufficiently small. Hence, the weak convergence $F_{\xi_n} \Rightarrow F_{\xi}$ immediately follows.

On the other hand, suppose we have the weak convergence

$$F_{\xi_n}(x) \Rightarrow F_0(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geqslant 0, \end{cases}$$

where the limit function corresponds to the random variable $\xi = 0$. Then

$$\xi_n \longrightarrow 0$$

in probability, since, for any $-\varepsilon_1 < 0 < \varepsilon_2$,

$$\mathbf{P}\{\xi_n \leqslant -\varepsilon_1\} = F_{\xi_n}(-\varepsilon_1) \longrightarrow 0, \qquad \mathbf{P}\{\xi_n > \varepsilon_2\} = 1 - F_{\xi_n}(\varepsilon_2) \longrightarrow 0.$$

EXAMPLE (*The law of large numbers*). Let ξ_k, $\mathbf{E}\xi_k = a$, $k = 1, 2, \ldots$, be i.i.d. random variables (we only assume the existence of their mean value). If $f(u)$, $-\infty < u < \infty$, is the characteristic function of $\xi_k - a$, then the characteristic function of

$$\frac{1}{n}\sum_{k=1}^{n} \xi_k - a = \frac{1}{n}\sum_{k=1}^{n} (\xi_k - a)$$

is

$$f_n(u) = f(u/n)^n, \quad -\infty < u < \infty;$$

see (6.9). Applying (6.14) with $m = 1$, $a_1 = 0$, we have

$$f\left(\frac{u}{n}\right) = 1 + o\left(\frac{1}{n}\right)$$

and

$$f_n(u) = \left[1 + o\left(\frac{1}{n}\right)\right]^n \longrightarrow 1, \quad n \to \infty,$$

uniformly in $u' \leqslant u \leqslant u''$ from any finite interval, with the limit $f(u) \equiv 1$ being the characteristic function of the *zero* random variable. Thus, we have the weak convergence of the corresponding distribution functions, which implies, as we already know, that

$$\frac{1}{n}\sum_{k=1}^{n} \xi_k - a \longrightarrow 0$$

in probability.

7.2. THE CENTRAL LIMIT THEOREM

Suppose, we deal with a random variable which can be written as a sum

$$S_n = \sum_{k=1}^{n} \xi_{kn} \tag{7.7}$$

of a large number n of *small* independent random variables ξ_{kn}. To be more precise, let us assume that ξ_{kn} are *normalized* in the sense that

$$\mathbf{E}\xi_{kn} = a_{kn} = 0, \qquad \mathbf{E}\xi_{kn}^2 = b_{kn} \leqslant b_n \to 0, \qquad \mathbf{E}\sum_{k=1}^{n} \left|\xi_{kn}\right|^2 = 1,$$

and, consequently,

$$\mathbf{E}S_n = 0, \qquad \mathbf{D}S_n = 1.$$

Assume, in addition, that the so-called *Lyapunov condition*

$$\mathbf{E}\sum_{k=1}^{n} |\xi_{kn}|^3 \longrightarrow 0 \tag{7.8}$$

is satisfied, which roughly says that ξ_{kn} are small enough so that $|\xi_{kn}|^3$ in the mean are much smaller than $|\xi_{kn}|^2$, $k = 1, \ldots, n$. In this case, we can approximate the distribution of (7.7) by the *standard normal distribution*:

$$\mathbf{P}\{x' < S_n \leqslant x''\} \sim \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} \mathrm{e}^{-x^2/2}\, \mathrm{d}x, \tag{7.9}$$

thanks to the famous *central limit theorem*.

To prove it, consider the characteristic function $f_n(u)$ of S_n,

$$f_n(u) = \prod_{k=1}^{n} f_{kn}(u), \quad -\infty < u < \infty,$$

which is the product of the characteristic functions $f_{kn}(u)$ of (*independent*) ξ_{kn}. According to (6.14),

$$f_{kn}(u) = 1 - \frac{b_{kn}}{2} u^2 + \frac{c_{kn}}{6} u^3, \quad |c_{kn}| \leqslant \mathbf{E}|\xi_{kn}|^3.$$

Moreover,

$$h_{kn} = -\frac{b_{kn}}{2} u^2 + \frac{c_{kn}}{6} u^3 \longrightarrow 0, \quad n \to \infty,$$

uniformly in $u' \leqslant u \leqslant u''$ from every finite interval $[u', u'']$. Therefore,

$$\log f_n(u) = \sum_{k=1}^{n} \log(1 + h_{kn}) \longrightarrow -\frac{1}{2} u^2$$

since

$$\sum_{k=1}^{n} h_{kn} = -\frac{1}{2}\left(\sum_{k=1}^{n} b_{kn}\right) u^2 + \frac{1}{6}\left(\sum_{k=1}^{n} c_{kn}\right) u^3 \longrightarrow -\frac{1}{2} u^2$$

with

$$\sum_{k=1}^{n} b_{kn} = 1, \qquad \left|\sum_{k=1}^{n} c_{kn}\right| \leqslant \sum_{k=1}^{n} \mathbf{E}\left|\xi_{kn}\right|^3 \longrightarrow 0.$$

Thus

$$f_n(u) \longrightarrow \mathrm{e}^{-u^2/2},$$

where the limit function

$$f(u) = \mathrm{e}^{-u^2/2}$$

is the well-known characteristic function of the standard normal distribution. Using a general property of the weak convergence, see (6.18), we obtain the following result.

THEOREM. *Under the above conditions,*

$$\mathbf{P}\{x' < S_n \leqslant x''\} \longrightarrow \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} \mathrm{e}^{-u^2/2}\, \mathrm{d}x, \quad -\infty < x' < x'' < \infty. \tag{7.10}$$

EXAMPLE. Let ξ_k, $k = 1, 2, \ldots$, be independent identically distributed random variables with

$$\mathbf{E}\xi_k = a, \qquad \mathbf{E}\left|\xi_k - a\right|^2 = b = \sigma^2, \qquad \mathbf{E}\left|\xi_k - a\right|^3 = c < \infty.$$

The corresponding normalized sum can be written as

$$S_n = \frac{1}{\sigma\sqrt{n}}\left(\sum_{k=1}^{n} \xi_k - na\right) = \sum_{k=1}^{n} \xi_{kn}, \tag{7.11}$$

with

$$\xi_{kn} = (\xi_k - a)/\sigma\sqrt{n}, \quad k = 1, \ldots, n,$$

satisfying the conditions of the theorem, since

$$\mathbf{E}\xi_{kn} = 0, \qquad \mathbf{E}|\xi_{kn}|^2 = \frac{1}{n}, \qquad \mathbf{E}|\xi_{kn}|^3 = \frac{c}{\sigma^3 n^{3/2}}, \quad k = 1, \ldots, n.$$

Thus, one can apply to (7.11) the standard normal approximation (7.9), (7.10). This approximation can be applied, in particular, to estimate the probability of deviation of the empirical mean

$$\frac{1}{n}\sum_{k=1}^{n}\xi_n$$

from its mean value $a = \mathbf{E}\xi_k$, by writing (7.11) as

$$S_n = \frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{k=1}^{n}\xi_k - a\right). \tag{7.12}$$

CHAPTER 2

Random Processes

1. Random Processes with Discrete State Space

1.1. THE POISSON PROCESS AND RELATED PROCESSES

Let us return to the process of radioactive decay discussed above, where radium Ra disintegrates into radon Rn, by emitting α-particles. Let $\xi(t)$ be the total number of α-particles emitted up to time t. Of course, for any $0 \leqslant s \leqslant t$, the difference $\xi(t) - \xi(s)$ is the number of α-particles emitted during the time interval $(s, t]$. As we already know, the random variable $\xi(t) - \xi(s)$ is distributed according to the Poisson law

$$\mathbf{P}\{\xi(t) - \xi(s) = k\} = \frac{[a(t-s)]^k}{k!} e^{-a(t-s)}, \quad k = 0, 1, \ldots, \tag{1.1}$$

with the mean value

$$a(t-s) = \mathbf{E}[\xi(t) - \xi(s)]$$

which depends on the difference $t - s$ only. We have

$$a(t) = a(s) - a(t-s), \quad 0 \leqslant s \leqslant t < \infty,$$

since

$$\xi(t) = \xi(s) + [\xi(t) - \xi(s)],$$

which implies that $a(t)$ is *linear*:

$$a(t) = at, \quad t \geqslant 0. \tag{1.2}$$

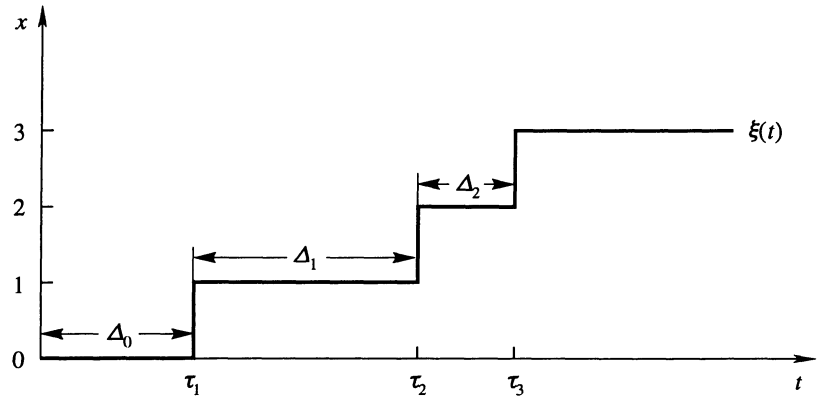

Fig. 6.

Here,

$$a = \mathbf{E}[\xi(t+1) - \xi(t)]$$

is the mean value of α-particles emitted during unit time, and we assume $\xi(0) = 0$.

A *trajectory* of the random process $\xi(t)$, $t \geqslant 0$, is shown in Figure 6, where τ_k are the moments of arrival (emittance) of α-particles, and

$$\tau = \tau_0,\ \tau_1 - \tau_0,\ \tau_2 - \tau_1, \ldots$$

are *waiting times*. Here, τ is *exponentially* distributed:

$$\mathbf{P}\{\tau > t\} = \mathrm{e}^{-\lambda t}, \quad t \geqslant 0 \tag{1.3}$$

(see p. 31); moreover, $\lambda = a$, since

$$\mathbf{P}\{\tau > t\} = \mathbf{P}\{\tau_0 > t\} = \mathbf{P}\{\xi(t) = 0\} = \mathrm{e}^{-at}$$

according to (1.1).

Consider $t \geqslant 0$ as the time axis. Suppose $\xi(s) = k$ at time s when we start our observation of the radioactive decay process. The waiting time for the transition $k \to k+1$ is distributed according to the same exponential law (1.3), which does not depend on the past behavior $\xi(t)$, $t \leqslant s$, up to the moment s. For $k = 0$, this follows from the known equality:

$$\mathbf{P}\{\tau > s + t \mid \tau > s\} = \mathbf{P}\{\tau > t\} \tag{1.4}$$

(see p. 31).

The above properties, including (1.1)–(1.4), define the *Poisson process* $\xi(t)$, $t \geqslant 0$, with intensity a. □

Homogeneous Markov property. Consider an arbitrary *random process* $\xi(t)$, $t \geqslant 0$, with a countable number of states $i = 0, 1, \ldots$. Starting at some initial state $\xi(0) = i$, $i = 0, 1, \ldots$, the process stays there for random time τ_0, after which it jumps to a new state $\xi(\tau_0) = j$ with probability π_{ij}, $j \neq i$, $j = 0, 1, \ldots$,

$$\sum_{j \neq i} \pi_{ij} = 1, \tag{1.5}$$

where it stays up to the random moment τ_1 when the next transition $j \to \xi(\tau_1)$ occurs etc. We assume that for any fixed time $s > 0$ and any 'current' state $\xi(s) = k$, the 'future' $\xi(t)$, $t \geqslant s$, obeys the same probability law as the process with $s = 0$ and $\xi(0) = k$, independently of the 'past' $\xi(t)$, $t \leqslant s$. (This is called the *homogeneous Markov property.*) In particular, given $\xi(s) = k$, the (waiting) time from s up to the next state transition is distributed according to the exponential law with the same parameter $\lambda = \lambda_k$ as the corresponding time $\tau = \tau_0$ with $\xi(0) = k$, i.e.

$$\mathbf{P}\{\tau > t\} = \mathrm{e}^{-\lambda_k t}, \quad t \geqslant 0. \tag{1.6}$$

Clearly, the above model gives the Poisson process in the case when $\pi_{ij} = 1$ for $j = i + 1$ ($\pi_{ij} = 0$ otherwise) and $\lambda_k = \lambda$ ($= a$) does not depend on $k = 0, 1, \ldots$.

EXAMPLE (*A single server system*). Imagine a service system which serves customers as follows: in absence of any customers, and independently of what happened before, a customer's service time is distributed exponentially with parameter λ. If a customer is being served, then other arriving customers are rejected. We assume that the probability of more than one simultaneous arrival is equal to 0, and that, having completed serving the customers, the system waits, independently of what happened before, for the next arrival for a random time, which has an exponential distribution with parameter μ. Obviously, if we consider two states: $\xi(t) = 0$ if there are no customers, and $\xi(t) = 1$ if a customer is being served at time t, then $\xi(t)$, $t \geqslant 0$, is a random process satisfying the above conditions with $\lambda_0 = \mu$, $\pi_{0,1} = 1$ and $\lambda_1 = \lambda$, $\pi_{1,0} = 1$.

Note that the above model describes $\xi(t)$ up to the random time

$$\tau = \lim_{n \to \infty} \tau_n \tag{1.7}$$

only, and it can happen that $\tau < \infty$, which means an *infinite* number transitions in *finite* time.

For example, this phenomenon occurs in the growth process

$$\xi(0) = 0 \longrightarrow \xi(\tau_0) = 1 \longrightarrow \cdots \longrightarrow \xi(\tau_k) = k + 1 \longrightarrow \cdots$$

with $\lambda_k \to \infty$ satisfying

$$\sum_{k=1}^{\infty} \frac{1}{\lambda_k} < \infty.$$

Here,

$$\frac{1}{\lambda_k} = \mathbf{E}(\tau_k - \tau_{k-1}), \quad k = 1, 2, \ldots, \tag{1.8}$$

are the mean values of the time intervals between consecutive transitions so that

$$\mathbf{E}\tau = \lim_{n\to\infty} \mathbf{E}\tau_n = \sum_k \frac{1}{\lambda_k} < \infty,$$

see Chapter 1, (4.19), which clearly implies

$$\mathbf{P}\{\tau < \infty\} = 1.$$

Transition probabilities. Often, one is interested in probabilities $\mathbf{P}\{\xi(t) = j\}$, which depend on t and the initial state $\xi(0) = i$, in particular, the *transition probabilities*

$$p_{ij}(t) = \mathbf{P}\{\xi(t) = j \mid \xi(0) = i\}, \quad i, j = 0, 1, \ldots, \tag{1.9}$$

of a *time-homogeneous Markov process* $\xi(t)$, $t \geqslant 0$, *with countable number of states.*

Here, we assume that the probabilities (1.9) concern only finite number of transitions from $\xi(0) = i$ to $\xi(t) = j$; more precisely, we assume that

$$\bigcup_j \{\xi(t) = j\} = \{\tau > t\}, \qquad \sum_j p_{ij}(t) = \mathbf{P}\{\tau > t \mid \xi(0) = i\}, \tag{1.10}$$

where τ is defined by (1.7).

Note first that

$$\{\xi(t) = j\} \subseteq \bigcup_k \{\xi(s) = k\}, \quad t \geqslant s,$$

and, thanks to the Markov property,

$$p_{ij}(t) = \sum_k p_{ik}(s) p_{kj}(t-s), \quad 0 \leqslant s \leqslant t, \tag{1.11}$$

since the conditional probability of $\xi(t) = j$ given $\xi(0) = i$, $\xi(s) = k$ does not depend on $\xi(0) = i$ and equals $p_{kj}(t-s)$, while the probability of $\xi(s) = k$ under the condition $\xi(0) = i$ is $p_{ik}(s)$, $k = 0, 1, \ldots$.

Next, as $h \to 0$,

$$\begin{aligned} &p_{ii}(h) = 1 - \lambda_i h + o(h), \\ &p_{ij}(h) = \lambda_{ij} h + o(h), \quad j \neq i, \quad o(h)/h \to 0. \end{aligned} \tag{1.12}$$

Indeed, according to (1.6),

$$\mathbf{P}\{\tau_0 > h \mid \xi(0) = i\} = \mathrm{e}^{-\lambda_i h} = 1 - \lambda_i h + o(h)$$

and

$$\mathbf{P}\{\tau_0 \leqslant h, \xi(\tau_0) = j \mid \xi(0) = i\} = \left(1 - \mathrm{e}^{-\lambda_i h}\right)\pi_{ij} = \lambda_i \pi_{ij} h + o(h),$$

as the probability of the transition $i \to j$ at time $\tau_0 = s$ does not depend on s. Moreover, the probability of $\xi(h) = j$ occurring in the result of more than one transition is less than

$$\begin{aligned} &\sum_{k \neq i} \mathbf{P}\{\tau_0 \leqslant h, \xi(\tau_0) = k, \ \tau_1 - \tau_0 \leqslant h \mid \xi(0) = i\} \\ &\quad = \left(1 - \mathrm{e}^{-\lambda_j h}\right) \sum_{k \neq i} \pi_{ik} \left(1 - \mathrm{e}^{-\lambda_k h}\right) = o(h), \end{aligned}$$

since $\tau_1 - \tau_0$ does not depend on what happened before the moment $\tau_0 = s$ when a transition $i \to k$, $k \neq i$, occurred. The above argument shows that the constants λ_{ij} in (1.12) satisfy

$$\lambda_{ij} = \lambda_i \pi_{ij}, \quad j \neq i, \tag{1.13}$$

$$\sum_{j \neq i} \lambda_{ij} = \lambda_i \sum_{j \neq i} \pi_{ij} = \lambda_i.$$

12. THE KOLMOGOROV EQUATIONS

Consider the transition probabilities $p_{ij}(t) \geqslant 0$, $i, j = 0, 1, \ldots$, $t \geqslant 0$, which satisfy $\sum_j p_{ij}(t) \leqslant 1$ together with (1.11)–(1.12). According to (1.12), $p_{ij}(t)$, as a function of $t \geqslant 0$, is continuous and differentiable at the initial point $t = 0$,

$$p_{ij}(0) = \begin{cases} 1, & j = i, \\ 0, & j \neq i, \end{cases} \tag{1.14}$$

and

$$p'_{ij}(0) = \lambda_{ij}, \quad i, j = 0, 1, \ldots .$$

Here, we put $\lambda_{ii} = -\lambda_i$, i.e.

$$-\lambda_{ii} = \lambda_i = \sum_{j \neq i} \lambda_{ij}, \tag{1.15}$$

see (1.13).

Let us show that $p_{ij}(t)$ are differentiable for all $t > 0$ and

$$p'_{ij}(t) = \sum_k \lambda_{ik} p_{kj}(t); \quad i, j = 0, 1, \ldots . \tag{1.16}$$

According to (1.11),

$$p_{ij}(s + h) - p_{ij}(s) = [p_{ii}(h) - 1]p_{ij}(s) + \sum_{k \neq i} p_{ik}(h) p_{kj}(s), \quad h, s \geqslant 0,$$

which shows at once that

$$|p_{ij}(s + h) - p_{ij}(s)| \leqslant [1 - p_{ii}(h)] + \sum_{k \neq i} p_{ik}(h) \leqslant 2[1 - p_{ii}(h)] \longrightarrow 0, \quad h \to 0,$$

where we can set $s = t$ or $s = t - h$, for any $t > 0$. Therefore, $p_{ij}(t)$, $t \geqslant 0$, is continuous. In a similar way,

$$\begin{aligned} \frac{p_{ij}(s + h) - p_{ij}(s)}{h} &= \frac{p_{ii}(h) - 1}{h} p_{ij}(s) + \sum_{k \neq i} \frac{p_{ik}(h)}{h} p_{kj}(s) \\ &= \frac{p_{ii}(h) - 1}{h} p_{ij}(s) + \sum_{k \neq i, k \leqslant n} \frac{p_{ik}(h)}{h} p_{kj}(s) + r_n(h), \end{aligned}$$

where, for an arbitrary small $\varepsilon > 0$,

$$0 \leqslant r_n(h) = \frac{1}{h}\sum_{k>n} p_{ik}(h)p_{kj}(s) \leqslant \frac{1}{h}\sum_{k>n} p_{ik}(h) \leqslant \frac{1}{h}\left[1 - \sum_{k\leqslant n} p_{ik}(h)\right]$$

$$= \frac{1 - p_{ii}(h)}{h} - \sum_{k\leqslant n, k\neq i} \frac{p_{ik}(h)}{h} \longrightarrow \lambda_i - \sum_{k\leqslant n, k\neq i} \lambda_{ik} \leqslant \varepsilon, \quad h \to 0,$$

provided n was chosen large enough, since according to (1.15),

$$\sum_{k\neq i} \lambda_{ik} = \lambda_i.$$

This shows that for any $t \geqslant 0$ there are limits

$$\lim_{h\to 0} \frac{p_{ij}(t+h) - p_{ij}(t)}{h} = p'_{ij}(t)$$

and

$$p'_{ij}(t) = p'_{ii}(0)p_{ij}(t) + \lim_{n\to\infty} \sum_{k\neq i, k\leqslant n} p'_{ik}(0)p_{kj}(t).$$

Using the definition of λ_{ij} and condition (1.15), we obtain the following result.

THEOREM. *Transition probabilities* $p_{ij}(t)$, $t \geqslant 0$, *satisfy the differential equations* (1.16).

Regularity of the process. Suppose, we have the equality

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_j p_{ij}(t) = \sum_j \frac{\mathrm{d}}{\mathrm{d}t} p_{ij}(t)$$

$$= \sum_j \left[\sum_k \lambda_{ik} p_{kj}(t)\right] = \sum_k \lambda_{ik}\left[\sum_j p_{kj}(t)\right].$$

Then, with the notation

$$f_i(t) = \sum_j p_{ij}(t) = \mathbf{P}\{\tau > t \mid \xi(0) = i\}, \quad t \geqslant 0, \ i = 0, 1, \ldots,$$

see (1.10), we obtain the following system of differential equations

$$f_i'(t) = \sum_k \lambda_{ik} f_k(t)$$

with the initial conditions

$$f_i(0) = p_{ii}(0) = 1, \quad i = 0, 1, \ldots .$$

The above system has the solution

$$f_i(t) \equiv 1,$$

since

$$\sum_k \lambda_{ik} = -\lambda_i + \sum_{k \neq i} \lambda_{ik} = 0, \quad i = 0, 1, \ldots .$$

If this solution is *unique*, then

$$\mathbf{P}\{\tau > t\} \equiv 1, \quad t \geqslant 0. \tag{1.17}$$

In other words, *with probability* 1 *the number of state transitions of the Markov process is finite in any finite time interval.* In particular, this is true when the number of states is finite.

Equations (1.16) are known as the *Kolmogorov backward differential equations.* There are the corresponding *Kolmogorov forward differential equations*:

$$p_{ij}'(t) = \sum_k p_{ik}(t) \lambda_{kj}, \quad i, j = 0, 1, \ldots . \tag{1.18}$$

Equations (1.18) are satisfied, in particular, if

$$\lambda_k \leqslant C, \quad k = 0, 1, \ldots . \tag{1.19}$$

Indeed, in this case from (1.11) we obtain

$$p_{ij}'(t) = \sum_k p_{ik}(s) p_{kj}'(t - s), \quad 0 \leqslant s \leqslant t,$$

since, according to (1.16), the derivatives

$$|p'_{kj}(t-s)| = \left| \sum_l \lambda_{kl} p_{lj}(t-s) \right| \leqslant 2\lambda_k \leqslant 2C$$

are bounded, and the derivative series above converges uniformly. By setting there $s = 0$, we get (1.18).

Assuming condition (1.19), we have the following result.

THEOREM. *Transition probabilities* $p_{ij}(t)$, $t \geqslant 0$, *satisfy the differential equations* (1.18).

EXAMPLE (*Poisson process*). Consider the Poisson process, which corresponds to our general model with parameters

$$\lambda_{ij} = \begin{cases} -\lambda, & j = i, \\ \lambda, & j = i, \\ 0, & j \neq i, i+1. \end{cases}$$

Setting

$$f_k(t) = e^{\lambda t} p_{ij}(t), \quad j - i = k = 0, 1, \ldots,$$

we obtain from (1.18), (1.13) the system of equations

$$f'_0(t) = 0,$$

$$f'_k(t) = \lambda f_{k-1}(t), \quad k = 1, 2, \ldots,$$

subject to the initial conditions

$$f_0(0) = 1, \ f_k(0) = 0, \quad k = 1, 2, \ldots.$$

It is easy to see that the above system has the unique solution

$$f_0(t) = 1, \ f_1(t) = \lambda t, \ldots, \ f_k(t) = \frac{(\lambda t)^k}{k!}, \ldots.$$

Finally, we obtain the transition probabilities

$$\mathbf{P}\{\xi(t) = j \mid \xi(s) = i\} = \frac{[\lambda(t-s)]^k}{k!} e^{-\lambda(t-s)}, \quad t \geqslant s, \ j - i = k = 0, 1, \ldots,$$

which actually describe the Poisson process with parameter $a = \lambda$; see (1.1), (1.12).

1.3. EXAMPLE (BRANCHING PROCESSES)

Consider a *branching process* $\xi(t)$, $t \geqslant 0$, which describes an evolution of particles such that each particle, existing at time s, and independently of its past and other particles, is transformed into n particles at time $t + s$, with probability $p_n(t)$, $n = 0, 1, \ldots$. The state of the process is characterized by the total number $\xi(t)$ of particles existing at time t, assuming that this number is finite.

Hence, if $\xi(s) = k$, the number of particles at time $t + s$ is

$$\xi(s + t) = \xi_1(t) + \ldots + \xi_k(t),$$

where $\xi_i(t)$ denotes the number of the descendents of the ith particles after time t. The random variables $\xi_1(t), \ldots, \xi_k(t)$ are independent and have the same probability distribution:

$$\mathbf{P}\{\xi_i(t) = n\} = p_n(t), \quad n = 0, 1, \ldots .$$

It follows that $\xi(t)$, $t \geqslant 0$, is a homogeneous Markov process with transition probabilities

$$\begin{aligned} p_{kn}(t) &= \mathbf{P}\{\xi_1(t) + \ldots + \xi_k(t) = n\} \\ &= \sum_{n_1 + \ldots + n_k = n} p_{n_1}(t) \ldots p_{n_k}(t), \quad k = 1, 2, \ldots ; \ n = 0, 1, \ldots , \end{aligned} \tag{1.20}$$

and, clearly,

$$p_{00}(t) \equiv 1, \qquad p_{0n} \equiv 0, \quad n = 1, 2, \ldots .$$

According to the Kolmogorov backward differential equations,

$$p'_{1n}(t) = \sum_k \lambda_{1k} p_{kn}(t), \quad n = 0, 1, \ldots ,$$

where

$$\lambda_{1n} = p'_{1n}(0), \quad n = 0, 1, \ldots ,$$

and

$$\sum_{n\neq 1}\lambda_{1n} = -\lambda_{11} = \lambda_1.$$

The method of generating functions. Introduce the generating function

$$F_k(t,z) = \sum_n p_{kn}(t)z^n,$$

of z, $0 \leqslant z < 1$, where the sum is taken over $n = 0, 1, \ldots$. From (1.20), one can easily obtain

$$F_k(t,z) = F_1(t,z)^k, \quad k = 0, 1, \ldots. \tag{1.21}$$

Using the Kolmogorov backward differential equation and the bound

$$|p'_{1n}(t)| \leqslant 2\lambda, \quad n = 0, 1, \ldots,$$

for every fixed z, $0 \leqslant z < 1$ we obtain

$$\sum_n p'_{1n}(t)z^n = \sum_k \lambda_{1k} \sum_n p_{kn}(t)z^n,$$

or the following differential equation for the generating function $F_1(t,z)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}F_1(t,z) = \sum_k \lambda_{1k}F_k(t,z).$$

With (1.21) in mind, the above equation for $F(t,z) = F_1(t,z)$ can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}F(t,z) = \sum_k \lambda_{1k}F(t,z)^k.$$

We introduce the function

$$f(x) = \sum_k \lambda_{1k}x^k, \quad 0 \leqslant x \leqslant 1.$$

Using the fact that $F(0, z) = z$, we see that for every z, $0 \leqslant z \leqslant 1$, the generating function $F(t, z)$ coincides with the solution $x = x(t)$ of the equation

$$\frac{dx}{dt} = f(x), \quad t \geqslant 0, \tag{1.22}$$

subject to the initial condition $x(0) = z$.

Instead of equation (1.22), it is convenient to consider an equivalent equation for the inverse function $t = t(x)$ of $x = x(t)$, i.e.,

$$\frac{dt}{dx} = \frac{1}{f(x)}, \quad 0 \leqslant x \leqslant 1.$$

The solution of the above equation can be written as

$$t(x) = \int_z^x \frac{du}{f(u)}, \quad 0 \leqslant x \leqslant 1.$$

Analysis of the differential equation (1.22) *for the generating function.* From $\lambda_{1k} \geqslant 0$, $k \neq 1$, it follows

$$f''(x) = \sum_{k \geqslant 2} k(k-1)\lambda_{1k} x^{k-2} \geqslant 0, \quad 0 \leqslant x < 1,$$

so that the function $f(x)$ is convex and its derivative monotonically increases in the interval $0 < x < 1$. Moreover, as

$$\sum_{k=0}^{\infty} \lambda_{1k} = 0,$$

so $x = 1$ is a root of the equation $f(x) = 0$. Apart from it, the last equation may have another root in $(0, 1)$ (see Figure 7).

Suppose first that there is a root $x = \alpha$, $0 < \alpha < 1$. Then, $x^0(t) \equiv \alpha$ is a solution of the differential equation (1.22). Let $x(t)$ be another solution, with $x(0) = z$, $0 \leqslant z < \alpha$. Since $f'(\alpha)$ is finite and since, for $x \sim \alpha$, $f(x)$ is approximately equal to $f'(\alpha)(x - \alpha)$, it follows that the corresponding inverse function

$$t(x) = \int_z^x \frac{du}{f(u)}$$

increases to $+\infty$ as $x \to \alpha$.

Note that $x(t)$ does not intersect $x^0(t) \equiv \alpha$ anywhere. Moreover, as $f(x)$ is positive in the interval $0 \leqslant x < \alpha$, $x(t)$ is monotone increasing as $t \to \infty$ and is bounded by α. In particular, $x(t)$ has a limit $\beta = \lim_{t\to\infty} x(t)$, $z \leqslant \beta < \alpha$. On the other hand, as $x \to \beta$, the continuous function $f(x)$ has the limit

$$f(\beta) = \lim_{t\to\infty} f\big(x(t)\big) = \lim_{t\to\infty} x'(t).$$

It is clear that $f(\beta) = 0$, otherwise

$$x(t) = z + \int_0^t f\big(x(s)\big)\mathrm{d}s$$

tends to $+\infty$ as $t \to \infty$. Hence β is a root of the equation $f(x) = 0$ and $\beta = \alpha$. Consequently, every solution $x = x(t)$, $x(0) = z$, $0 \leqslant z < \alpha$, is monotone increasing and

$$\lim_{t\to\infty} x(t) = \alpha. \tag{1.23}$$

Solutions starting at $z \in (\alpha, 1)$ $(0 \leqslant \alpha < 1)$ at $t = 0$, behave in an analogous way, with the only difference that $x(t)$ is monotone decreasing, as $x'(t) = f(x(t))$ is negative ($f(x) \leqslant 0$ for $\alpha \leqslant x < 1$). The corresponding graphs of $x(t)$ for different values of $x(0) = z$, $0 \leqslant z < 1$, are shown in Figure 8. Obviously, the situation is more simple if $\alpha = 0$.

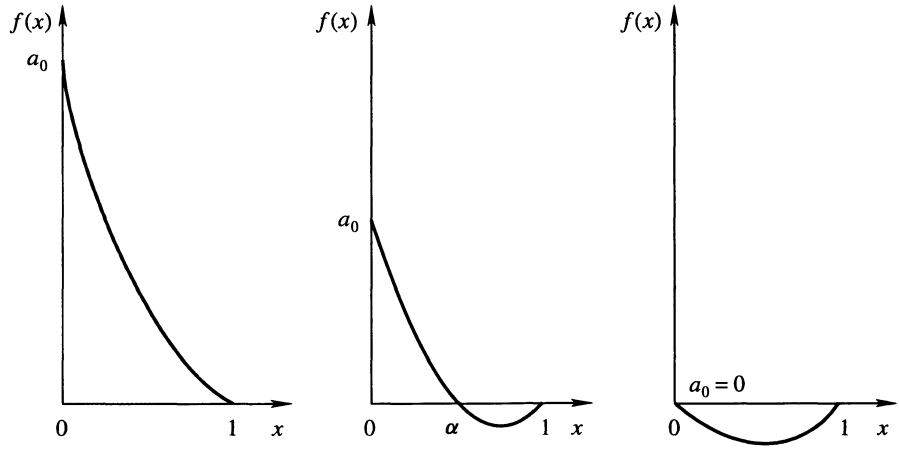

Fig. 7.

The case $x(0) = z = 1$ has to be considered separately. As $f(1) = 0$, so $x^1(t) \equiv 1$ is a solution of equation (1.22). Assume first that $1/f(x)$ is nonintegrable in a neighbourhood of $x = 1$, i.e., $\alpha < 1$ and

$$\int_{x_0}^{1} \frac{du}{f(u)} = -\infty, \qquad \alpha < x_0 < 1. \tag{1.24}$$

Take an arbitrary solution $x(t)$, $x(0) = 1$ of (1.22). Suppose that $x(t_0) = x_0$ for some $t_0 = t(x_0) \geqslant 0$. The corresponding inverse function can be written as

$$t(x) = t_0 + \int_{x_0}^{x} \frac{du}{f(u)}.$$

Note that $x(t)$ does not intersect $x^1(t) \equiv 1$ for $t \geqslant 0$ since

$$t(1) = t_0 + \int_{x_0}^{1} \frac{du}{f(u)} = -\infty.$$

In particular, $x(0) = z < 1$, which is a contradiction. Therefore, $x(t) \equiv 1$ is the unique solution going through the point $t = 0$, $x = 1$.

Next, consider the case

$$\int_{x_0}^{1} \frac{dx}{f(x)} > -\infty. \tag{1.25}$$

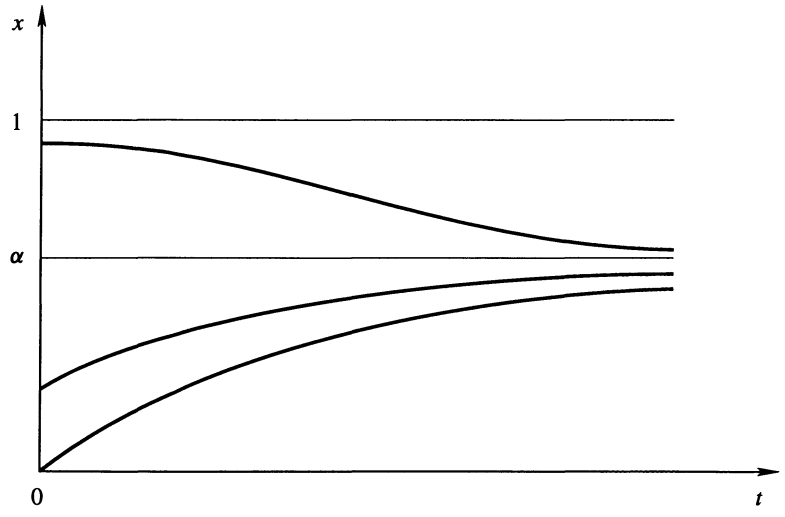

Fig. 8.

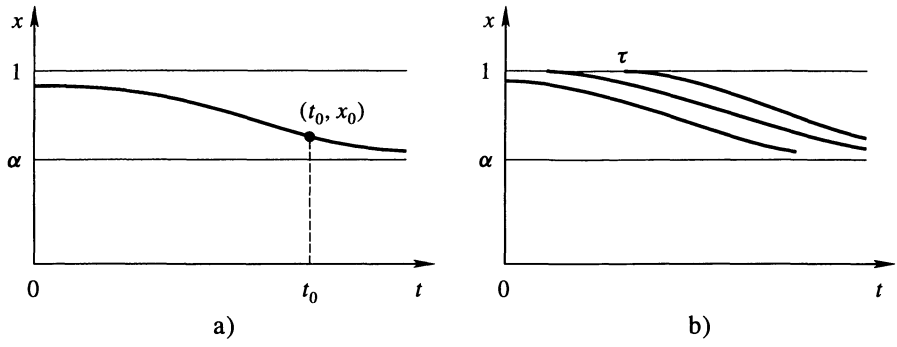

Fig. 9.

Then, for sufficiently large $t_0 > 0$, the corresponding inverse curve

$$t(x) = t_0 + \int_{x_0}^{x} \frac{du}{f(u)}$$

intersects tangentially $x^1(t) \equiv 1$ at a certain point $t = \tau$, $x = 1$, where

$$\tau = t_0 + \int_{x_0}^{1} \frac{du}{f(u)} \geqslant 0$$

(see Figure 9). In this case, we have an entire family $\{x_\tau(t)\}_{\tau \geqslant 0}$ of solutions going through the point $t = 0$, $x = 1$, each $x_\tau(t)$ corresponding to a particular choice of $\tau \geqslant 0$. In particular, the solution $x_0(t)$ corresponding to $\tau = 0$ has the property that it lies below all other solutions, namely

$$x_0(t) \leqslant x_\tau(t), \quad 0 < t < \infty.$$

This follows from the fact that, in the domain $0 \leqslant x < 1$, $0 \leqslant t < \infty$, a solution of the corresponding differential equation is unique so that different solutions do not intersect each other in this domain. It is easy to see that $x_0(t)$ is the (increasing) limit of solutions $x(t) = x(t, z)$, $x(0, z) = z \in [0, 1)$:

$$x_0(t) = \lim_{z \to 1} x(t, z). \tag{1.26}$$

The study of the differential equation (1.22) enables us to draw the following conclusions about the corresponding branching process $\xi(t)$, $t \geqslant 0$.

The probability of degeneration. In general, there is a positive probability that the number of particles at time t is zero. (Of course, this probability is 0 if $\lambda_{10} = 0$, i.e., if the number of particles does not decrease.) Given $k = 0, 1, 2, \ldots$ particles at time $t = 0$, the above probability is $p_{k0}(t) = F(t, 0)^k = p_0(t)^k$ (see (1.21)).

The function $p_0(t)$ is a solution of the differential equation (1.22) with the parameter $z = 0$:

$$p_0'(t) = f\big(p_0(t)\big), \qquad p_0(0) = 0.$$

We know that, as $t \to \infty$, this solution tends to α, or the smallest root of the equation $f(x) = 0$ (see (1.23)), i.e.,

$$\lim_{t\to\infty} p_0(t) = \alpha.$$

Thus, α is the probability of degeneration of the branching process $\xi(t)$, or the probability that after some time the number of particles is 0. More generally, if we are given k particles at time $t = 0$, then the probability of degeneration is

$$\lim_{t\to\infty} p_{k0}(t) = \alpha^k.$$

The probability of explosion. From (1.22) we see that, under the condition (1.24), with probability 1 every particle creates a finite number of particles in a finite time. Indeed,

$$\lim_{z\to 1} F(t, z) = \sum_n p_{1n}(t) = 1,$$

and

$$\mathbf{P}\{\tau \leqslant t\} = 1 - \sum_n p_{1n}(t) \equiv 0, \quad t \geqslant 0,$$

where τ is the time when the total population of particles becomes infinite, see (1.10). On the other hand, if condition (1.25) holds, then

$$\lim_{z\to 1} F(t, z) = x_0(t) = \mathbf{P}\{\tau > t\} < 1$$

according to (1.26), where $x_0(t) < 1$ $(t > 0)$. Thus, with a positive probability $\mathbf{P}\{\tau \leqslant t\}$, a given particle produces an infinite number of offsprings in a finite

time t. (This phenomenon is called *explosion*.) The corresponding probability given k particles at time $t = 0$ is

$$1 - x_0(t)^k,$$

see (1.21).

QUESTION. Is explosion possible for a population of particles which multiply by dividing into two new ones (i.e., with $\lambda_{12} = -\lambda_{11} = \lambda$ being the only non-zero coefficient)?

1.4. THE (LIMIT) STATIONARY PROBABILITY DISTRIBUTION

Let $\xi(t)$, $t \geqslant 0$, be a homogeneous Markov process having transition probabilities

$$p_{ij}(t); \quad i, j = 0, 1, \ldots,$$

$$\sum_j p_{ij}(t) \equiv 1, \quad t \geqslant 0, \tag{1.27}$$

cf. (1.10). Consider the probabilities

$$p_i(s) = \mathbf{P}\{\xi(s) = i\}, \quad i = 0, 1, \ldots,$$

at some time moment $s \geqslant 0$. According to the total probability formula,

$$p_j(t) = \mathbf{P}\{\xi(t) = j\} = \sum_i p_i(s) p_{ij}(t - s), \quad t \geqslant s, \quad j = 0, 1, \ldots. \tag{1.28}$$

At $s = 0$, we have the *initial probability distribution*

$$p_i(0) = \mathbf{P}\{\xi(0) = i\}, \quad i = 0, 1, \ldots,$$

$$\sum_i p_i(0) = 1.$$

We say that a probability distribution

$$p_i^* \geqslant 0, \quad i = 0, 1, \ldots,$$

$$\sum_i p_i^* = 1,$$

is *stationary*, if

$$p_j^* = \sum_i p_i^* p_{ij}(t), \quad t \geqslant 0, \quad j = 0, 1, \ldots .$$

According to (1.28), if the initial probability distribution is *stationary*: $p_i(0) = p_i^*$, then the probabilities

$$p_i(t) = \mathbf{P}\{\xi(t) = i\} \equiv p_i^*, \quad i = 0, 1, \ldots , \tag{1.29}$$

do not depend on $t \geqslant 0$.

Under condition (1.19) we can obtain from (1.28) the following *forward differential equations*

$$p_j'(t) = \sum_i p_i(t)\lambda_{ij}, \quad i = 0, 1, \ldots , \tag{1.30}$$

similar to the corresponding equations (1.18) for transition probabilities. Applying (1.30) to (1.29), we get the equations

$$\sum_i p_i^* \lambda_{ij} = 0, \quad j = 0, 1, \ldots , \tag{1.31}$$

from which stationary probabilities can be found (if they exist).

Of course, it might happen that there is no stationary distribution, as in the case of the Poisson process, say.

We are going to show that a *unique stationary probability distribution* exists if there are j_0, $h = h_0 > 0$ and $\delta > 0$ such that

$$p_{ij_0}(h) \geqslant \delta > 0 \tag{1.32}$$

for any $i = 0, 1, \ldots .$

Note that (1.32) holds in the case of a *finite* number of states which can be reached one from another. Indeed, if $p_{ij}(s) > 0$ for some $s > 0$, then

$$p_{ij}(t) \geqslant p_{ij}(s)p_{jj}(t - s) > 0, \quad t \geqslant s,$$

since $p_{jj}(0) = 1$ and the continuity of $p_{jj}(t)$ at $t = 0$ imply

$$p_{jj}(t) \geqslant p_{jj}(h)p_{jj}(t-h) > 0, \quad t \geqslant 0.$$

THEOREM. *There exists a unique stationary probability distribution* p_j^*, $j = 0, 1, \ldots$, *and*

$$p_j(t) = \mathbf{P}\{\xi(t) = j\} \to p_j^*$$

as $t \to \infty$. *Moreover,*

$$|p_j(t) - p_j^*| \leqslant (1-\delta)^{(t/h)-1} \tag{1.33}$$

independently of j *and the initial probability distribution.*

Proof. Set

$$r_j(t) = \inf_i p_{ij}(t), \quad R_j(t) = \sup_i p_{ij}(t),$$

which give lower and upper bounds, respectively, for the probability

$$p_j(t) = \sum_i p_i^{(0)} p_{ij}(t) \begin{cases} \geqslant \sum_i p_i^{(0)} r_j(t) = r_j(t), \\ \leqslant \sum_i p_i^{(0)} R_j(t) = R_j(t). \end{cases}$$

Let us show that $r_j(t)$ monotonically increases and $R_j(t)$ monotonically decreases, as $t \to \infty$. In fact, for any $t \geqslant s$, we have

$$r_j(t) = \inf_i \left[\sum_k p_{ik}(t-s)p_{kj}(s)\right] \geqslant \inf_i \left[\sum_k p_{ik}(t-s)r_j(s)\right] = r_j(s),$$

$$R_j(t) = \sup_i \left[\sum_k p_{ik}(t-s)p_{kj}(s)\right] \leqslant \sup_i \left[\sum_k p_{ik}(t-s)R_j(s)\right] = R_j(s).$$

Furthermore,

$$R_j(t) - r_j(t) = \sup_{\alpha,\beta} [p_{\alpha i}(t) - p_{\beta j}(t)]$$

$$= \sup_{\alpha,\beta} \sum_k [p_{\alpha k}(h) - p_{\beta k}(h)]p_{kj}(t-h), \quad t \geqslant h.$$

Here,

$$\sum_k p_{\alpha k}(h) = \sum_k p_{\beta k}(h) = 1,$$

hence

$$0 = \sum_k [p_{\alpha k}(h) - p_{\beta k}(h)]$$

can be rewritten as

$$\sum_k{}^{+} [p_{\alpha k}(h) - p_{\beta k}(h)] = -\sum_k{}^{-} [p_{\alpha k}(h) - p_{\beta k}(h)],$$

where $\sum^+, \sum^-$ correspond to positive and negative summands, respectively. From condition (1.32), it is easy to show that

$$\sum_k{}^{+} [p_{\alpha k}(h) - p_{\beta k}(h)] = \frac{1}{2}\sum_k |p_{\alpha k}(h) - p_{\beta k}(h)| \leqslant \frac{1}{2}(2 - 2\delta) = 1 - \delta.$$

Hence,

$$\begin{aligned} R_j(t) - r_j(t) &\leqslant \sup_{\alpha,\beta} \left\{ \sum_k{}^{+} [p_{\alpha k}(h) - p_{\beta k}(h)] R_j(t-h) \right. \\ &\qquad \left. + \sum_k [p_{\alpha k}(h) - p_{\beta k}(h)] r_j(t-h) \right\} \\ &= \sup_{\alpha,\beta} \sum_k{}^{+} [p_{\alpha k}(h) - p_{\beta k}(h)] \big(R_j(t-h) - r_j(t-h)\big) \\ &\leqslant (1-\delta)\big(R_j(t-h) - r_j(t-h)\big). \end{aligned}$$

Consequently,

$$R_j(t) - r_j(t) \leqslant (1-\delta)^n \big(R_j(t-nh) - r_j(t-nh)\big) \leqslant (1-\delta)^{(t/h)-1},$$

where n is the entire part of t/h. Together with the monotonicity of $R_j(t)$, $r_j(t)$, this implies the existence of the limit

$$p_j^* = \lim_{t\to\infty} r_j(t) = \lim_{t\to\infty} p_j(t) = \lim_{t\to\infty} R_j(t)$$

as well as the uniform bound (1.33):

$$|p_j(t) - p_j^*| \leqslant R_j(t) - r_j(t) \leqslant (1 - \delta)^{(t/h)-1}.$$

To complete the proof of the theorem, we have to demonstrate that p_j^*, $j = 0, 1, \ldots$ define a stationary probability distribution.

Note first that $\sum_j p_j^* \leqslant 1$ as the last inequality is true for any *finite* number of summands:

$$\sum_j p_j^* = \lim_{t\to\infty} \sum_j p_j(t) \leqslant 1.$$

The sum $\sum_j p_j^* \neq 0$ which follows from (1.28) and

$$p_{j_0}^* \geqslant r_{j_0}(h) \geqslant \delta.$$

Moreover, from (1.28) with $t \to \infty$, $s = t - h$, $h \geqslant 0$ it follows that

$$p_j^* \geqslant \sum_i p_i^* p_{ij}(h).$$

In fact, the last inequality must be replaced by equality, since, by assuming that a strict inequality holds for some j, we obtain

$$\sum_j p_j^* > \sum_j \sum_i p_i^* p_{ij}(h) = \sum_i p_i^* \sum_j p_{ij}(h) = \sum_i p_i^*.$$

Therefore, the probability distribution

$$p_j^0 = \frac{p_j^*}{\sum_k p_k^*}, \quad j = 0, 1, \ldots,$$

is stationary:

$$p_j^0 = \sum_i p_i^0 p_{ij}(t), \quad t \geqslant 0.$$

By taking $p_j^0 = p_j(0)$, $j = 0, 1, \ldots$, as the initial distribution and using the first part of the theorem, we conclude that

$$p_j^* = \lim_{t\to\infty} p_j(t) = p_j^0, \quad j = 0, 1, \ldots.$$

The last relations being obviously valid for an arbitrary stationary distribution p_j^0, $j = 0, 1, \ldots$, shows that such distribution is unique, which completes the proof of the theorem. □

EXAMPLE (*Multi-server system*). Imagine a service system that is analogous to the system described on p. 93, but that has n, instead of one, lines of service. The service time of customers arriving at each service line, is random and is exponentially distributed with parameter λ. In particular, if j lines are occupied, the waiting time until one of them is free is

$$\tau = \min(\tau_1, \dots, \tau_j),$$

where the waiting times $\tau_1, \dots, \tau_j$ of the occupied lines are independent and have the same exponential distribution with the parameter λ. The variable τ is distributed exponentially with the parameter $j\lambda$. Let $\xi(t)$ be the number of occupied lines at time t. Then $\xi(t)$ is a homogeneous Markov process with $n+1$ states $j = 0, 1, \dots, n$ with the transition parameters (see (1.12)):

$$\lambda_{0j} = \begin{cases} -\mu, & j = 0, \\ \mu, & j = 1, \\ 0, & j \neq 0, 1, \end{cases} \qquad \lambda_{nj} = \begin{cases} n\lambda, & j = n-1, \\ -n\lambda, & j = n, \\ 0, & j \neq n-1, n, \end{cases}$$

$$\lambda_{ij} = \begin{cases} i\lambda, & j = i-1, \\ -i\lambda - \mu, & j = i, \\ \mu, & j = i+1, \\ 0, & j \neq i-1, i, i+1, \end{cases} \qquad 0 < i < n.$$

We recall that μ is the parameter of the exponential distribution of the interarrival time of customers. From (1.31) we obtain the system of equations

$$\begin{aligned} &-\mu p_0^* + \lambda p_1^* = 0, \\ &\mu p_{i-1}^* - (\mu + i\lambda) p_i^* + (i+1)\lambda p_{i+1}^* = 0, \quad 0 < i < n, \\ &\mu p_{n-1}^* - n p_n^* = 0, \end{aligned}$$

whose solution is given by *Erlang's formula*:

$$p_j^* = \frac{\frac{1}{j!}\left[\frac{\mu}{\lambda}\right]^j}{\sum_{k=0}^{n} \frac{1}{k!}\left[\frac{\mu}{\lambda}\right]^k}, \quad j = 0, 1, \dots, n.$$

For large n, one can apply the *Poisson approximation*

$$p_j^* \sim \frac{a^j}{j!} e^{-a}, \quad j = 0, 1, \dots,$$

with $a = \mu/\lambda$.

It is clear that, starting from an arbitrary state, the process can reach any other state, and that condition (1.32) holds. Therefore, the probability distribution of $\xi(t)$ converges as $t \to \infty$ to the (stationary) Erlang distribution (see (1.33)). The convergence is very explicit in the case $n = 1$, when the forward differential equations can be easily solved and, with 0 and 1 the only states, we obtain

$$p_{01}(t) = \frac{\lambda}{\lambda+\mu}\left[1 - e^{-(\lambda+\mu)t}\right] \longrightarrow \frac{\lambda}{\lambda+\mu} = p_1^*,$$

$$p_{10}(t) = \frac{\mu}{\lambda+\mu}\left[1 - e^{-(\lambda+\mu)t}\right] \longrightarrow \frac{\mu}{\lambda+\mu} = p_0^*, \quad t \to \infty.$$

EXAMPLE (*Energy supply*). Suppose there are n independent energy customers who use energy during random time intervals. With each customer, one can associate a homogeneous Markov process taking value 1 when energy is consumed and 0 when no energy is needed, with transition parameters $\lambda_{01} = \lambda$, $\lambda_{00} = -\lambda$, $\lambda_{10} = \mu$, $\lambda_{11} = -\mu$. Assuming that all these n processes are *independent*, their sum $\xi(t)$ forms a homogeneous Markov process with states $0, 1, \ldots, n$, $\xi(t) = k$ being the number of energy users at time t. Obviously, the transition parameters of $\xi(t)$ are

$$\begin{aligned}
\lambda_{01} &= n\lambda, \qquad \lambda_0 = -\lambda_{00} n\lambda, \\
\lambda_{k,k+1} &= (n-k)\lambda, \qquad \lambda_{k,k-1} = k\mu, \\
\lambda_k &= -\lambda_{kk} = (n-k)\lambda + k\mu, \quad 1 \leqslant k \leqslant n-1, \\
\lambda_{n,n-1} &= n\mu, \qquad \lambda_n = -\lambda_{nn} = n\mu.
\end{aligned}$$

One can easily verify that the limit stationary probabilities are

$$p_j^* = \frac{n!}{j!(n-j)!}\left(\frac{\lambda}{\lambda+\mu}\right)^j \left(\frac{\mu}{\lambda+\mu}\right)^j, \quad j = 0, 1, \ldots, n,$$

giving the binomial (Bernoulli) distribution with the parameter $p = \lambda/(\lambda+\mu)$.

2. Random Processes with Continuous States

2.1. THE BROWNIAN MOTION

Imagine a particle moving in a homogeneous fluid, in the result of chaotic collisions with the molecules of the fluid. The corresponding continuous chaotic motion of the particle is called the Brownian motion.

Let $\xi_1(t)$, $\xi_2(t)$ be the particle's plane coordinates at time $t \geqslant 0$, where $\xi_1(0) = 0$, $\xi_2(0) = 0$, say. From physical argument, $\xi_1(t)$, $\xi_2(t)$ can be assumed to be independent random variables with a probability density which is central symmetric with respect to the origin. Choose any axis on the plane going through the origin, and let $\xi(t)$, $\xi(0) = 0$ be the projection of $(\xi_1(t), \xi_2(t))$ onto it. Then, as we already know, $\xi(t)$ is a *normal* variable with the *normal* probability density

$$p(0, t, x) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} e^{-x^2/2\sigma^2(t)}, \quad -\infty < x < \infty,$$

with zero mean $\mathbf{E}\xi(t) = 0$ and the variance

$$\sigma^2(t) = \mathbf{D}\xi(t).$$

Of course, we can consider the Brownian motion process starting from $\xi(0) = x$; then the corresponding probability density is

$$p(x, t, y) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} e^{-(y-x)^2/2\sigma^2(t)}, \quad -\infty < y < \infty. \tag{2.1}$$

Consider the particle's diffusion at disjoint time intervals $(0, s)$ and (s, t), $0 < s < t$. The corresponding displacements $\xi(s) = \xi(s) - \xi(0)$ and $\xi(t) - \xi(s)$ arise in the result of physically independent collisions of molecules; i.e., we can assume that

$$\xi(t) - \xi(s), \quad \xi(s) - \xi(0)$$

are *independent* random variables. Moreover, because of homogeneity of the fluid, $\xi(t) - \xi(s)$ obeys the same probability law as $\xi(t - s) - \xi(0)$. To be more precise, the probability density of $\xi(t-s) - \xi(s)$ given any $\xi(s) = x$ is the same as the probability density of $\xi(t - s) - \xi(0)$ given $\xi(0) = 0$.

In particular,

$$\mathbf{D}[\xi(t) - \xi(0)] = \mathbf{D}[\xi(s) - \xi(0)] + \mathbf{D}[\xi(t) - \xi(s)],$$

which shows that the function

$$\sigma^2(t) = \sigma^2(s) + \sigma^2(t - s), \quad 0 \leqslant s \leqslant t < \infty,$$

is *linear*:

$$\sigma^2(t) = \sigma^2 t, \quad t \geqslant 0,$$

where the constant σ^2 is called the *diffusion coefficient.*

The corresponding conditional probability density of $\xi(t)$ given $\xi(s) = x$ equals

$$p(x, t-s, y) = \frac{1}{\sqrt{2\pi\sigma^2(t-s)}} e^{-(y-x)^2/2\sigma^2(t-s)}, \quad -\infty < y < \infty. \tag{2.2}$$

Roughly speaking, the Brownian motion has the property that, given a 'current' state $\xi(s) = x$, and *independently* of the 'past' $\xi(t)$, $t \leqslant s$, the 'future' $\xi(t)$, $t \geqslant s$, obeys the same probability law as the initial process starting at $s = 0$ from the point $\xi(0) = x$ (this is called the *time-homogeneous Markov property*).

2.2. TRAJECTORIES OF THE BROWNIAN MOTION

Having characterized the probability distribution of the random variables $\xi(t)$, $t \geqslant 0$ (see (2.1)–(2.2)), now we shall define (random) trajectory of the Brownian motion as a limit of discrete time piecewise linear approximations

$$\xi_n(t) = \frac{t - t_{k+1,n}}{t_{kn} - t_{k+1,n}} \xi(t_{kn}) - \frac{t - t_{kn}}{t_{k+1,n} - t_{kn}} \xi(t_{k+1,n}), \quad t_{kn} \leqslant t \leqslant t_{k+1,n}, \tag{2.3}$$

where

$$t_{kn} = \frac{k}{2^n}, \quad k = 0, 1, \ldots .$$

THEOREM. *The random functions* (2.3) *uniformly converge with probability* 1 *on each finite time interval.*

Proof. Consider the event

$$A_T^{m,n} = \left\{ \max_{0 \leqslant t \leqslant T} [\xi_n(t) - \xi_m(t)] > \varepsilon_m \right\},$$

where $n > m$, T are positive integers, $T > 0$. Note that, for $\xi_n(t)$ of (2.3), the maximum above is attained at a partition point $t_{kn} = k/2^n$, and that, in view of

monotonicity of the partitions, the maximum increases with n. For the union $A_T^m = \cup_{n>m} A_T^{m,n}$ of monotone increasing events $A_T^{m,n}$, $n = m+1, m+2, \ldots$, we have

$$\mathbf{P}(A_T^m) = \lim_{n\to\infty} \mathbf{P}(A_T^{m,n}).$$

In the following, we shall obtain an estimate of $\mathbf{P}(A_T^{m,n})$, uniformly in n, which applies also for $\mathbf{P}(A_T^m)$, or the probability of the event:

$$\max_{0\leqslant t\leqslant T} |\xi_n(t) - \xi_m(t)| > \varepsilon_m$$

for some $n > m$.

It is obvious that

$$A_T^{m,n} \subseteq \bigcup_i \left\{ \max_{t_{im}\leqslant t_{kn}\leqslant t_{i+1,m}} |\xi(t_{kn}) - \xi(t_{im})|, |\xi(t_{kn}) - \xi(t_{i+1,m})| > \varepsilon_m \right\}$$

and

$$\begin{aligned}
\mathbf{P}(A_T^{m,n}) &\leqslant \sum_i \mathbf{P}\left\{ \max_{t_{im}\leqslant t_{kn}\leqslant t_{i+1,m}} |\xi(t_{kn}) - \xi(t_{im})|, |\xi(t_{kn}) - \xi(t_{i+1,m})| > \varepsilon_m \right\} \\
&= 2^m T \cdot \mathbf{P}\left\{ \max_{0\leqslant t_{kn}\leqslant 2^{-m}} |\xi(t_{kn}) - \xi(0)|, |\xi(t_{kn}) - \xi(2^{-m})| > \varepsilon_m \right\} \\
&\leqslant 2^m T \cdot 4\mathbf{P}\left\{ \max_{0\leqslant k\leqslant 2^{n-m}} \xi(t_{kn}) > \varepsilon_m \right\};
\end{aligned}$$

in the last inequality we use the fact that the families

$$\pm[\xi(t_{kn}) - \xi(0)], \quad k = 0, 1, \ldots, 2^{n-m},$$

as well as

$$\pm[\xi(2^{-m}) - \xi(t_{kn})], \quad k = 2^{n-m}, \ldots, 1, 0,$$

of random variables obey the same probability law as the family $\xi(t_{kn})$, $k = 0, 1, \ldots, 2^{n-m}$. Next, we apply the following general lemma.

LEMMA. *Let $\xi_1, \dots, \xi_r$ be random variables such that, for any $k = 1, \dots, r-1$, the distribution of $\xi_r - \xi_k$ does not depend on $\xi_1, \dots, \xi_k$ and is symmetric with respect to 0. Then*

$$\mathbf{P}\left\{ \max_{1 \leqslant k \leqslant r} \xi_k > x \right\} \leqslant 2\mathbf{P}(\xi_r > x), \qquad x > 0. \tag{2.4}$$

Proof. Denote ξ_ν the first of the variables $\xi_1, \dots, \xi_r$ that exceeds x. As the event $\nu = k$ is determined by $\xi_1, \dots, \xi_k$ which do not depend on $\xi_r - \xi_k$, we obtain

$$\begin{aligned}
\mathbf{P}\left\{ \max_{1 \leqslant k < r} \xi_k > x,\ \xi_r \leqslant x \right\} &= \sum_{k=1}^{r-1} \mathbf{P}\{\nu = k,\ \xi_r \leqslant x\} \\
&\leqslant \sum_{k=1}^{r-1} \mathbf{P}\{\nu = k,\ \xi_r - \xi_k < 0\} \\
&= \sum_{k=1}^{r-1} \mathbf{P}\{\nu = k\}\mathbf{P}\{\xi_r - \xi_k < 0\} \\
&\leqslant \sum_{k=1}^{r-1} \mathbf{P}\{\nu = k\}\mathbf{P}\{\xi_r - \xi_k \geqslant 0\} \\
&= \sum_{k=1}^{r-1} \mathbf{P}\{\nu = k, \xi_r - \xi_k \geqslant 0\} \leqslant \mathbf{P}\{\xi_r > x\}.
\end{aligned}$$

Complementing the resulting the inequality by the following one:

$$\mathbf{P}\left\{ \max_{0 \leqslant k \leqslant r} \xi_k > x,\ \xi_r > x \right\} \leqslant \mathbf{P}\{\xi_r > x\},$$

we obtain the estimate (2.4).

Applying this estimate to the variables $\xi(t_{kn})$, $k = 0, 1, \dots, 2^{n-m}$, we get

$$\mathbf{P}\left\{ \max_{0 \leqslant k \leqslant 2^{n-m}} \xi(t_{kn}) > \varepsilon_m \right\} \geqslant 2\mathbf{P}\{\xi(2^{-m}) > \varepsilon_m\},$$

where

$$\begin{aligned}
\mathbf{P}\{\xi(2^{-m}) > \varepsilon_m\} &= \frac{1}{\sqrt{2\pi}} \int_{\varepsilon_m \sqrt{2^m}/\sigma} \mathrm{e}^{-x^2/2} \mathrm{d}x \\
&\leqslant \frac{1}{\sqrt{2\pi}} \frac{\sigma}{\varepsilon_m \sqrt{2^m}} \int_{\varepsilon_m \sqrt{2^m}/\sigma} x \mathrm{e}^{-x^2/2} \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \frac{\sigma}{\varepsilon_m \sqrt{2^m}} \mathrm{e}^{-\varepsilon_m^2 2^m / 2\sigma^2}.
\end{aligned}$$

Consequently, we obtain the following estimate:

$$\mathbf{P}\left(A_T^m\right) \leqslant 4T\sigma \frac{1}{\sqrt{2\pi}} \frac{\sqrt{2^m}}{\varepsilon_m} e^{-\varepsilon_m^2 2^m/2\sigma^2}.$$

Choose $\varepsilon_m \to 0$ so that the series

$$\sum_{m=1}^{\infty} \frac{\sqrt{2^m}}{\varepsilon_m} e^{-\varepsilon_m^2 2^m/2\sigma^2} < \infty$$

converges (for example, we may take $\varepsilon_m = 2^{-m/4}$). Then, as

$$\sum_{m=1}^{\infty} \mathbf{P}\left(A_T^m\right) < \infty,$$

using the Borel–Cantelli lemma, we obtain that, with probability 1, only a finite number of events A_T^m, $m = 1, 2, \ldots$, occur. In other words, with probability 1, for all sufficiently large m and any $n > m$,

$$\max_{0 \leqslant t \leqslant T} |\xi_n(t) - \xi_m(t)| \leqslant \varepsilon_m,$$

where $\varepsilon_m \to 0$ $(m \to \infty)$. We have that, with probability 1, the sequence (2.3) converges uniformly on every finite interval $0 \leqslant t \leqslant T$, or the statement of the theorem. □

We define now the Brownian motion process as the limit

$$\xi(t) = \lim_{n \to \infty} \xi_n(t), \quad t \geqslant 0, \tag{2.5}$$

which is a *random continuous* function, the distribution of the random variables $\xi(t)$, $t \geqslant 0$, being characterized in (2.1), (2.2).

More precisely, we assume that the basic probability space $(\Omega, \mathfrak{A}, \mathbf{P})$ is chosen so that the limit

$$\xi(t, \omega) = \lim_{n \to \infty} \xi_n(t, \omega), \quad t \geqslant 0, \tag{2.5$'$}$$

exists and satisfies the statement of the last theorem for every $\omega \in \Omega$. In particular, every *trajectory*

$$\xi(t) = \xi(\omega, t), \quad t \geqslant 0, \tag{2.6}$$

is a continuous function, which on any finite interval $0 \leqslant t \leqslant T$, is represented by the uniform limit (2.5)′ of our continuous piecewise linear functions $\xi_n(t)$, $t \geqslant 0$.

For the above continuous model, we can restate our probabilistic characterization of the Brownian motion process $\xi(t)$, $t \geqslant 0$ (which is also called the *Wiener process*), as follows: 1) $\xi(0) = 0$; 2) for any $0 < s < t$, the increment $\xi(t) - \xi(s)$ has normal distribution with expectation 0 and variance $\sigma^2(t-s)$; 3) for any $0 < t_1 < \ldots < t_n$, the increments $\xi(t_1) - \xi(0), \ldots, \xi(t_n) - \xi(t_{n-1})$ are independent. □

Some experimental trajectories of the Brownian motion can be seen in Figure 10*.

Visually, trajectories of the Brownian motion look as if they were chaotically drawn by a jittering pen (which reflects the character of the physical process of Brownian motion, where the particle is subject to infinitely frequent impulses from the molecules, and every impulse produces an infinitely small displacement). As we shall see below, with probability 1 the trajectory has infinite variation on any interval:

$$\sup_{s=t_0<t_1<\cdots<t_n=t} \sum_{k=1}^{n} |\xi(t_k) - \xi(t_{k-1})| = \infty. \tag{2.7}$$

THEOREM. *For any interval* $[s,t]$, *with probability* 1

$$\lim_{n\to\infty} \sum_{k=1}^{n} \left[\xi(t_k) - \xi(t_{k-1})\right]^2 = \sigma^2(t-s), \tag{2.8}$$

where the limit is taken over a sequence $\{t_{kn}\}$, $n \geqslant 1$ *of partitions* $s = t_0 < t_1 < \cdots < t_n = t$, $t_k \equiv t_{kn}$ *such that*

$$h_n = \max_{1\leqslant k\leqslant n} |t_{kn} - t_{k-1,n}| \leqslant 2^{-n}.$$

Proof. Let show first that (2.8) holds for any sequence of partitions with $h_n \to 0$, if we replace the convergence with probability 1 by the convergence in the square mean. In fact, we have

$$\mathbf{E}\left[\xi(t_k) - \xi(t_{k-1})\right]^2 = \sigma^2(t_k - t_{k-1}).$$

Set

$$\Delta_k = \left[\xi(t_k) - \xi(t_{k-1})\right]^2 - \sigma^2(t_k - t_{k-1}).$$

* Wold, H.O. (ed.): *Bibliography of Time Series and Stochastic Processes,* pp. 10–11, Edinburgh, London, 1965.

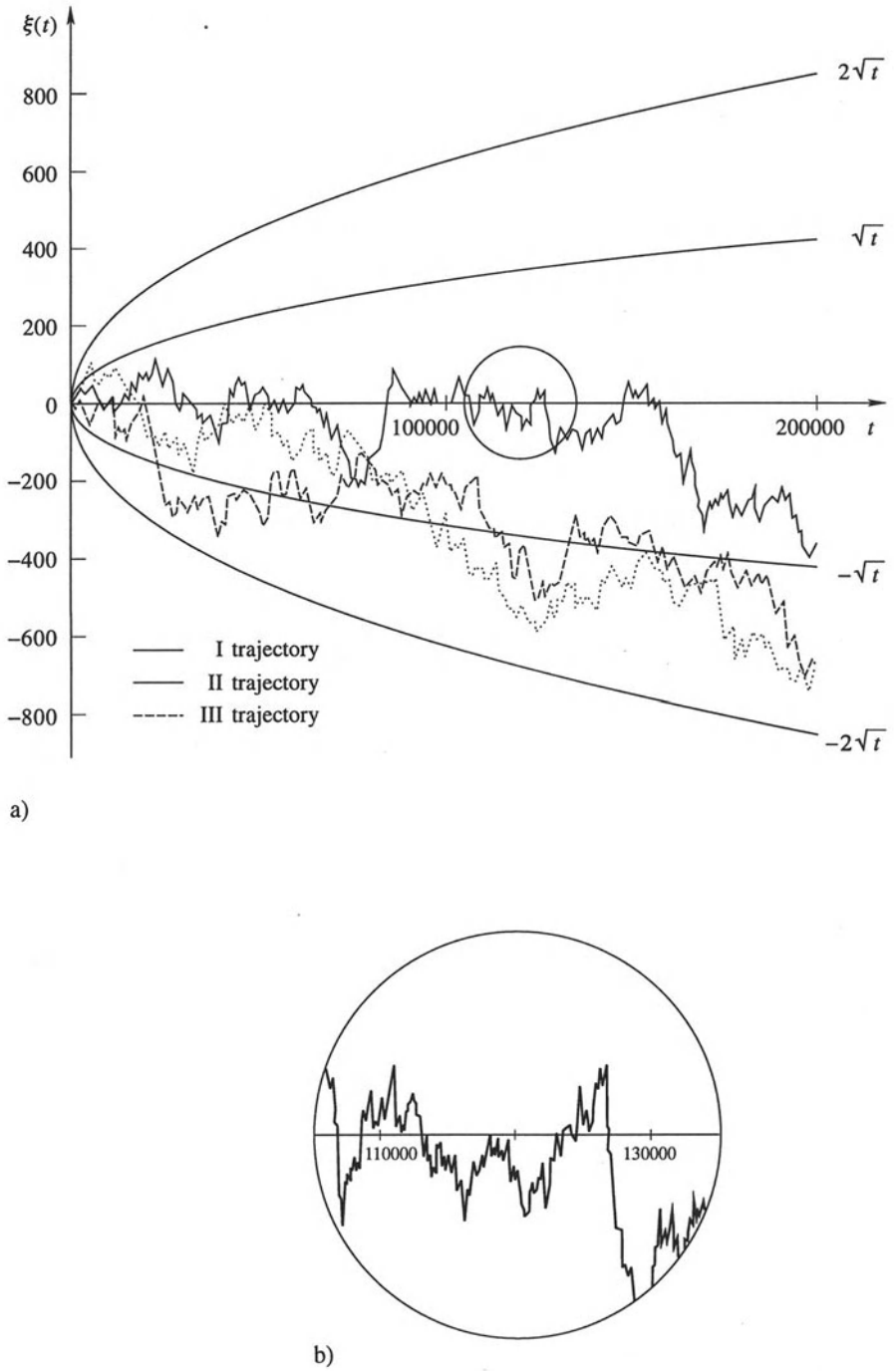

Fig 10 a) Experimental trajectories of the Brownian motion with the diffusion coefficient $\sigma^2 = 1$, b) a magnified part of the trajectory

Consider the sum

$$\sum_{k=1}^{n} \left[\xi(t_k) - \xi(t_{k-1})\right]^2 - \sigma^2(t-s) = \sum_{k=1}^{n} \Delta_k$$

of the independent variables Δ_k, with mean 0 and the variance

$$\mathbf{E}\Delta_k^2 = \mathbf{E}\left[\xi(t_k) - \xi(t_{k-1})\right]^4 - \sigma^4(t_k - t_{k-1})^2 = 2\sigma^2(t_k - t_{k-1})^2$$

(see p. 79 for moments of the normal distribution). We obtain

$$\mathbf{E}\left[\sum_{k=1}^{n} \Delta_k\right]^2 = \sum_{k=1}^{n} \mathbf{E}\Delta_k^2 = 2\sigma^4 \sum_{k=1}^{n} (t_k - t_{k-1})^2$$

$$\leqslant 2\sigma^4 \max_{1 \leqslant k \leqslant n} (t_k - t_{k-1}) \sum_{k=1}^{n} (t_k - t_{k-1}) = 2\sigma^4 h_n (t-s) \longrightarrow 0$$

as $n \to \infty$. Next, from the Chebyshev inequality,

$$\mathbf{P}\left\{\left|\sum_{k=1}^{n} \Delta_k\right| > \varepsilon_n\right\} \leqslant 2\sigma^4 (t-s) \frac{h_n}{\varepsilon_n^2}.$$

As $h_n \leqslant 2^{-n}$, we can choose $\varepsilon_n \to 0$ so that

$$\sum_{n=1}^{\infty} \mathbf{P}\left\{\left|\sum_{k=1}^{n} \Delta_k\right| > \varepsilon_n\right\} < \infty.$$

Using the Borel–Cantelli lemma, we infer that, with probability 1, only finitely many of events

$$\left|\sum_{k=1}^{n} \Delta_k\right| > \varepsilon_n$$

occur, i.e., for all sufficiently large n we have

$$\left|\sum_{k=1}^{n} \Delta_k\right| \leqslant \varepsilon_n, \quad \text{where} \quad \varepsilon_n \longrightarrow 0.$$

The theorem is proved. □

In particular, *a.e. trajectory of the Brownian motion has infinite variation on any interval.* Indeed,

$$\sum_{k=1}^{n} |\xi(t_k) - \xi(t_{k-1})| \geqslant \frac{1}{\max_k |\xi(t_k) - \xi(t_{k-1})|} \sum_{k=1}^{n} |\xi(t_k) - \xi(t_{k-1})|^2 \longrightarrow \infty$$

since $\max_k |\xi(t_k) - \xi(t_{k-1})| \to 0$, due to continuity of the trajectory on the interval $[s, t]$.

2.3. MAXIMA AND HITTING TIMES

Let us consider the standard Brownian motion $\xi(t)$, $t \geqslant 0$, with the diffusion coefficient $\sigma^2 = 1$, starting at $\xi(0) = 0$. We are interested in the probability distribution of the *maximal displacement* (or *maximum*)

$$\xi_t^+ = \max_{0 \leqslant s \leqslant t} \xi(s),$$

and of the *hitting time*

$$\tau_x = \min\{t\colon\ \xi(t) \geqslant x\}$$

of a point $x > 0$. Because of continuity of the Brownian motion, the random variables ξ_t^+ and τ_x are clearly related between themselves, namely

$$\{\xi_t^+ \geqslant x\} = \{\tau_x \leqslant t\}.$$

Using the symmetry of the Brownian motion with respect to any starting point, one can see that, under the condition $t > \tau_x$, both events $\xi(t) \geqslant x$ and $\xi(t) \leqslant x$ are equiprobable. Hence

$$\mathbf{P}\{\xi(t) \geqslant x \mid \tau_x \leqslant t\} = \frac{\mathbf{P}\{\xi(t) \geqslant x\}}{\mathbf{P}\{\tau_x \leqslant t\}} = \frac{1}{2}$$

and therefore

$$\mathbf{P}\{\tau_x \leqslant t\} = 2\mathbf{P}\{\xi(t) \geqslant x\} = 2\left[1 - \Phi\left(\frac{x}{\sqrt{t}}\right)\right], \quad t > 0, \tag{2.9}$$

where

$$\Phi\Big(\frac{x}{\sqrt{t}}\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x/\sqrt{t}} \mathrm{e}^{-u^2/2} \mathrm{d}u.$$

is the normal distribution function.

Of course, we also have

$$\mathbf{P}\{\xi_t^+ \geqslant x\} = 2\mathbf{P}\{\xi(t) \geqslant x\} = 2\Big[1 - \Phi\Big(\frac{x}{\sqrt{t}}\Big)\Big], \quad x > 0. \tag{2.10}$$

Differentiating (2.9) with respect to $t > 0$ and (2.10) with respect to $x > 0$, we obtain the corresponding probability densities of the random variables τ_x and ξ_t^+, namely

$$p_{\tau_x}(t) = \begin{cases} \dfrac{x}{\sqrt{2\pi}} t^{-3/2} \mathrm{e}^{-x^2/2t}, & t > 0, \\ 0, & t \leqslant 0 \end{cases} \tag{2.11}$$

and

$$p_{\xi_t^+}(x) = \begin{cases} \sqrt{\dfrac{2}{\pi t}}\, \mathrm{e}^{-x^2/2t}, & x \geqslant 0, \\ 0, & x < 0. \end{cases} \tag{2.12}$$

By the symmetry of the Brownian motion with respect to the origin $\xi(0) = 0$, the hitting time

$$\tau_x = \min\{t\colon\ \xi(t) < x\}$$

of a point $x < 0$, and the minimum

$$\xi_t^- = \min_{0 \leqslant s \leqslant t} \xi(s) = -\max_{0 \leqslant s \leqslant t} [-\xi(s)]$$

obey the same probability laws (2.11), (2.12), respectively, with x replaced by $-x$. In particular, the Brownian particle hits any point $x = a$ with probability 1. Moreover, before hitting a point $a < 0$, say, the particle travels for some time in the opposite direction $x > 0$, with τ_x, $\xi_t^{\pm}$ asymptotically behaving like x^2, $\pm\sqrt{t}$, respectively (as x and t increase). □

Let $0 \leqslant \tau \leqslant t$ be the time when the Brownian trajectory attains its maximum ξ_t^+. To find the distribution of this random variable, one can proceed as follows.

Let us consider the maximum ξ_t^+ on an interval $0 \leqslant s \leqslant t$ and the hitting time τ_a of $a > 0$. Obviously, $\xi_t^+ = x \geqslant a$ implies $\tau_a \leqslant t$ and, given $\tau_a = s < t$,

$$\xi_t^+ = a + \max_{s \leqslant u \leqslant t} [\xi(u) - a].$$

The Brownian motion $\xi(u)$, $u \geqslant s$, starting at $\xi(s) = \xi(\tau_a) = a$ obeys the same probability law as $\xi(u - s) + a$, $u - s \geqslant 0$, $\xi(0) = 0$. Therefore, the conditional probability density of ξ_t^+ given $\tau_a = s \leqslant t$ coincides with the probability density of the random variable

$$a + \max_{0 \leqslant h \leqslant t-s} \xi(h) = a + \xi_{t-s}^+$$

and can be found from (2.12), to be equal to

$$p_{\xi_t^+}(x \mid s) = \sqrt{\frac{2}{\pi(t-s)}}\, e^{-(x-a)^2/2(t-s)}, \quad a \leqslant x < \infty.$$

Hence, using the Bayes formula and the hitting time distribution (2.11), we find the conditional probability density of τ_a given $\xi_t^+ = x \geqslant a$:

$$p_{\tau_a}(s \mid x) \equiv p_{\tau_a}(s) \frac{p_{\xi_t^+}(x \mid s)}{p_{\xi_t^+}(x)}$$

$$\equiv \frac{a}{\sqrt{2\pi}}\, s^{-3/2} e^{-a^2/2s} \sqrt{\frac{2}{\pi(t-s)}}\, e^{-(x-a)^2/2(t-s)} \frac{1}{p_{\xi_t^+}(x)},$$

$$0 < s < t, \ a \leqslant x < \infty.$$

The above identity remains valid for $x = a$ as well:

$$p_{\tau_a}(s \mid a) = \frac{1}{\pi\sqrt{s(t-s)}} \frac{a}{s} e^{-a^2/2s} \frac{1}{p_{\xi_t^+}(a)}, \quad 0 < s < t.$$

Conditioned at $\xi_t^+ = a$, $0 < a < \infty$, we have $\tau_a = \tau$ as the maximum point. Therefore, the conditional density of τ given $\xi_t^+ = x$, $0 < x < \infty$ is equal to

$$p_\tau(s \mid x) = \frac{1}{\pi\sqrt{s(t-s)}} \frac{x}{s} e^{-x^2/2s} \frac{1}{p_{\xi_t^+}(x)}, \quad 0 < s < t.$$

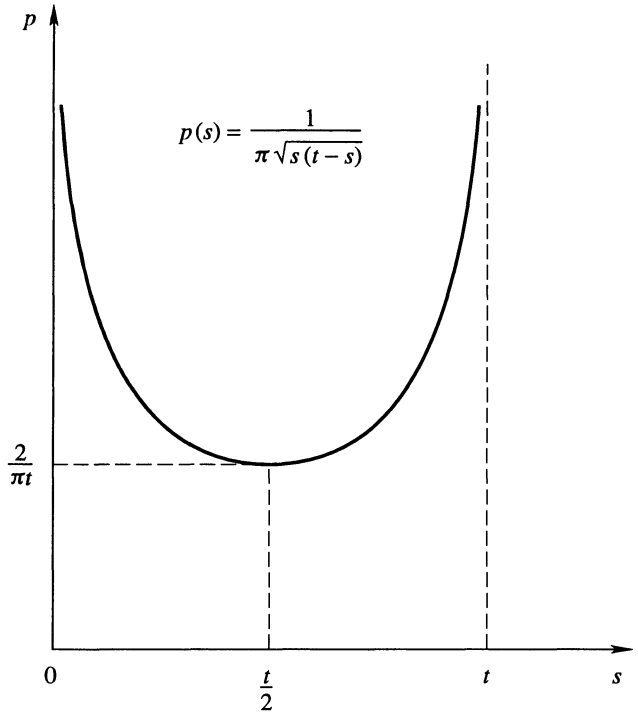

Fig. 11.

Hence, the joint probability density of τ and ξ_t^+ is

$$p_{\tau,\xi_t^+}(s,x) = \frac{1}{\pi\sqrt{s(t-s)}} \frac{x}{s} e^{-x^2/2s}, \quad 0 < s < t, \ 0 \leqslant x < \infty, \tag{2.13}$$

which leads to the following formula for the probability density of the maximum point τ:

$$p_\tau(s) = \int_0^\infty p_{\tau,\xi_t^+}(s,x)\,dx = \frac{1}{\pi\sqrt{s(t-s)}}, \quad 0 < s < t.$$

This probability distribution is known as the *arc sine law*, since

$$\mathbf{P}\{\tau \leqslant s\} = \int_0^s \frac{1}{\pi\sqrt{u(t-u)}}\,du = \frac{2}{\pi}\arcsin\sqrt{\frac{s}{t}}, \quad 0 \leqslant s \leqslant t \tag{2.14}$$

(see Figure 11). Note that the *maximum* point $\tau = s$ is much more likely to occur towards the ends $s = 0$ and $s = t$ of the interval $(0, t)$ than somewhere in the middle (this happens, for example, if the Brownian particle drifts from the origin in the direction $x < 0$).

2.4. DIFFUSION PROCESSES

Consider a particle moving in a non-homogeneous medium, whose movement $\xi(t)$, $t \geqslant 0$, *locally* resembles the Brownian motion, but, given the position $\xi(s) = x$ of the particle at a time s, the increment $\xi(s+h) - \xi(s)$ *depends* on $x = \xi(s)$.

Of course, the probabilistic characterization of such a movement cannot be as simple as of the Brownian motion in general. However, we shall assume the following *Markov property*: given any 'current' state $\xi(s) = x$, the 'future' $\xi(t)$, $t \geqslant s$, of the random process is conditionally independent of its 'past' $\xi(t)$, $t \leqslant s$.

Suppose that, for any $t > s$ and $-\infty < x < \infty$, there exists the *transition probability density*

$$p(x, t-s, y), \quad -\infty < y < \infty,$$

which is the conditional probability density of the random variable $\xi(t)$ given $\xi(s) = x$. □

Let us consider the joint probability distribution of $\xi(s)$, $\xi(t)$, $0 < s < t$, given $\xi(0) = x$. The corresponding probability density can be written as

$$p(x, s, z)\, p(z, t-s, y), \quad -\infty < z, y < \infty,$$

with $p(z, t-s, y)$, $-\infty < y < \infty$, being the conditional density of $\xi(t)$ given $\xi(0) = x$, $\xi(s) = z$ (the latter density does not depend on $\xi(0) = x$, thanks to the Markov property). Integrating over $-\infty < z < \infty$, we obtain the probability density of $\xi(t)$:

$$p(x, t, y) = \int_{-\infty}^{\infty} p(x, s, z) p(z, t-s, y)\, \mathrm{d}z, \quad -\infty < y < \infty. \tag{2.15}$$

This is the so-called *Kolmogorov–Chapman equation* (which is quite similar to Equation (1.11) and reflects the total probability formula).

We assume that, for any fixed $\varepsilon > 0$,

$$\begin{aligned}
&\int_{|y-x|>\varepsilon} p(x, h, y)\, \mathrm{d}y = o(h), \\
&\int_{|y-x|\leqslant\varepsilon} (y-x) p(x, h, y)\, \mathrm{d}y = a(x) \cdot h + o(h), \\
&\int_{|y-x|\leqslant\varepsilon} (y-x)^2 p(x, h, y)\, \mathrm{d}y = b(x) \cdot h + o(h),
\end{aligned} \tag{2.16}$$

where $o(h)/h \to 0$ as $h \to 0$. A random process $\xi(t)$, $t \geqslant 0$, which satisfies the above mentioned properties, is usually called a *diffusion process*. The functions $a(x)$ and $b(x)$, appearing in (2.16), are called the drift coefficient and the diffusion coefficient, respectively.

The Brownian motion process considered above is an example of a diffusion process, with the corresponding coefficients $a(x) = 0$, $b(x) = \sigma^2$.

THEOREM. *Suppose that the derivatives $\partial p/\partial t$, $\partial p/\partial x$, $\partial^2 p/\partial x^2$ of the transition density $p(x, t, y)$ exist and are continuous with respect to x, uniformly in y from each finite interval $y_0 \leqslant y \leqslant y_1$. Then $p(x, t, y)$ satisfies the diffusion equation*

$$\frac{\partial p}{\partial t} = a(x)\frac{\partial p}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2 p}{\partial x^2}. \tag{2.17}$$

(Equation (2.17) is known as the *Kolmogorov backward equation.*)

Proof. Take a continuous function $\varphi(x)$ vanishing outside a finite interval, and set

$$\varphi(t, x) = \int_{-\infty}^{\infty} \varphi(y)p(x, t, y)\,\mathrm{d}y.$$

From the Kolmogorov–Chapman equation it follows that

$$\begin{aligned}\varphi(t, x) &= \int_{-\infty}^{\infty} \varphi(y) \int_{-\infty}^{\infty} p(x, s, z)p(z, t - s, y)\,\mathrm{d}z\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \varphi(s, z)p(x, t - s, z)\,\mathrm{d}z.\end{aligned}$$

Obviously, the function $\varphi(t, x)$ has continuous derivatives $\partial\varphi/\partial t$, $\partial\varphi/\partial x$, $\partial^2\varphi/\partial x^2$. Using the Taylor formula in the neighbourhood of x (with s fixed), we obtain

$$\varphi(s, z) - \varphi(s, x) = \frac{\partial\varphi(s, x)}{\partial x}(z - x) + \frac{1}{2}\left[\frac{\partial^2\varphi(s, x)}{\partial x^2} + O(\delta_\varepsilon)\right](z - x)^2,$$

where

$$\delta_\varepsilon = \sup_{|z-x|\leqslant\varepsilon} \left|\frac{\partial^2\varphi(s, z)}{\partial x^2} - \frac{\partial^2\varphi(s, x)}{\partial x^2}\right|,$$

so that $O(\delta_\varepsilon) \to 0$ as $\varepsilon \to 0$. From the relations (2.16), with $t - s = h \to 0$, we get

$$\begin{aligned}
\varphi(t,x) - \varphi(s,x) &= \int_{-\infty}^{\infty} \big[\varphi(s,z) - \varphi(s,x)\big] p(x,h,z)\,\mathrm{d}z \\
&= \int_{|z-x|\leqslant\varepsilon} \big[\varphi(s,z) - \varphi(s,x)\big] p(x,h,z)\mathrm{d}z + o(h) \\
&= \frac{\partial\varphi(s,x)}{\partial x} \int_{|z-x|\leqslant\varepsilon} (z-x)p(x,h,z)\,\mathrm{d}z \\
&\quad + \frac{1}{2}\left[\frac{\partial^2\varphi(s,x)}{\partial x^2} + O(\delta_\varepsilon)\right] \int_{|z-x|\leqslant\varepsilon} (z-x)^2 p(x,h,z)\,\mathrm{d}z + o(h) \\
&= \left\{a(x)\frac{\partial\varphi(s,x)}{\partial x} + \frac{1}{2}b(x)\left[\frac{\partial^2\varphi(s,x)}{\partial x^2} + O(\delta_\varepsilon)\right]\right\} h + o(h).
\end{aligned}$$

Hence

$$\lim_{h=t-s\to 0} \frac{\varphi(t,x) - \varphi(s,x)}{h} = a(x)\frac{\partial\varphi(s,x)}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2\varphi(s,x)}{\partial x^2},$$

or

$$\frac{\partial\varphi}{\partial t} = a(x)\frac{\partial\varphi}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2\varphi}{\partial x^2}, \quad t > 0.$$

Using the definition of $\varphi(t,x)$, we can rewrite the above equation as

$$\int_{-\infty}^{\infty} \varphi(y)\left[-\frac{\partial p}{\partial t} + a(x)\frac{\partial p}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2\varphi}{\partial x^2}\right]\mathrm{d}y = 0,$$

where (recall) $\varphi(y)$ is an arbitrary continuous function vanishing outside some finite interval. Hence, the equation

$$-\frac{\partial p}{\partial t} + a(x)\frac{\partial p}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2\varphi}{\partial x^2} = 0$$

is satisfied. The theorem is proved.

THEOREM. *Suppose that the derivatives*

$$\frac{\partial p(x,t,y)}{\partial t}, \quad \frac{\partial[a(y)p(x,t,y)]}{\partial y}, \quad \frac{\partial^2[b(y)p(x,t,y)]}{\partial y^2}$$

exist and are continuous. Then the transition density $p(x,t,y)$ satisfies the differential equation

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial y}\big[a(y)p(x,t,y)\big] + \frac{1}{2}\frac{\partial^2}{\partial y^2}\big[b(y)p(x,t,y)\big]. \tag{2.18}$$

Proof. Similarly to the proof of the previous theorem, we can show that, for any twice continuously differentiable function $\varphi(x)$ vanishing outside some finite interval, the limit

$$\lim_{h\to 0}\frac{1}{h}\left[\int_{-\infty}^{\infty}\varphi(y)p(x,h,y)\,\mathrm{d}y - \varphi(x)\right] = a(x)\varphi'(x) + \frac{1}{2}b(x)\varphi''(x)$$

exists. We obtain

$$\begin{aligned}
&\frac{\partial}{\partial t}\int_{-\infty}^{\infty} p(x,t,y)\varphi(y)\,\mathrm{d}y \\
&\quad = \lim_{h\to 0}\frac{1}{h}\left[\int_{-\infty}^{\infty} p(x,t+h,y)\varphi(y)\,\mathrm{d}y - \int_{-\infty}^{\infty} p(x,t,z)\varphi(z)\,\mathrm{d}z\right] \\
&\quad = \int_{-\infty}^{\infty} p(x,t,z)\lim_{h\to 0}\frac{1}{h}\left[\int_{-\infty}^{\infty} p(z,h,y)\varphi(y)\,\mathrm{d}y - \varphi(z)\right]\mathrm{d}z \\
&\quad = \int_{-\infty}^{\infty} p(x,t,z)\Big[a(z)\varphi'(z) + \frac{1}{2}b(z)\varphi''(z)\Big]\mathrm{d}z.
\end{aligned}$$

Integrating the last expression by parts, we get

$$\begin{aligned}
\frac{\partial}{\partial t}\int_{-\infty}^{\infty} p(x,t,y)\varphi(y)\,\mathrm{d}y &= \int_{-\infty}^{\infty}\Big[\frac{\partial}{\partial t}p(x,t,y)\Big]\varphi(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left\{-\frac{\partial}{\partial y}\big[a(y)p(x,t,y)\big] + \frac{1}{2}\frac{\partial^2}{\partial y^2}\big[b(y)p(x,t,y)\big]\right\}\varphi(y)\,\mathrm{d}y.
\end{aligned}$$

Hence, equation (2.18) follows, as $\varphi(y)$ is arbitrary. The theorem is proved.

CHAPTER 3

An Introduction to Mathematical Statistics

1. Some Examples of Statistical Problems and Methods

1.1. ESTIMATION OF THE SUCCESS PROBABILITY IN BERNOULLI TRIALS

According to a common belief, the birth of a girl or of a boy are equiprobable events. Let us adopt this as the initial *hypothesis*, and check how it fits some available data. For example, in the period 1871–1900 there were $n = 2,644,757$ babies born in Switzerland including $m = 1,359,671$ boys and $n - m = 1,285,086$ girls.* How well does this data agree with our hypothesis that the probability of a boy's birth is 0.5? By calling the last event a 'success', let us discuss the data in the framework of $n = 2,644,757$ Bernoulli trials, with unknown success probability p; the corresponding frequency is

$$\frac{m}{n} = \frac{1,359,671}{2,644,757} = 0.5141.$$

No doubt that everybody would reject a hypothesis like $p = 0.1$, say. To give a rigorous answer for any hypothesis about the probability p, consider *apriori* the frequency as a random variable whose probability distribution is well-known. Namely, as n is very large, one can apply to this random variable, denoted by ξ, the normal approximation. The normalized random variable

$$\frac{\xi - \mathbf{E}\xi}{\sqrt{\mathbf{D}\xi}} = \sqrt{\frac{n}{p(1-p)}}\,(\xi - p)$$

satisfies the inequality

$$\sqrt{\frac{n}{p(1-p)}}\,(\xi - p) \leqslant x_\alpha \tag{1.1}$$

* See Van der Waerden: *Mathematische Statistik*, Springer-Verlag, Berlin, 1957.

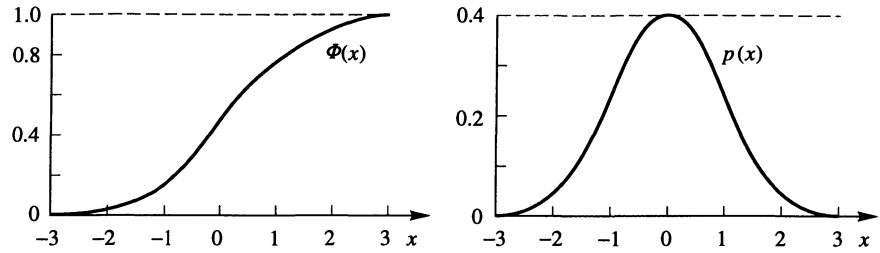

Fig. 12. a) Normal distribution function with $\sigma = 1$; b) normal density with $\sigma = 1$.

with the probability

$$1 - \alpha = \Phi(x_\alpha),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2}\, dx, \quad -\infty < x < \infty,$$

is the normal distribution function (see Figure 12 and Table II). Here, α is a *significance level* and x_α is the corresponding *quantile*.

Let us return to the above data which gives the value

$$\sqrt{\frac{n}{p(1-p)}}\,(\xi - p) = 37$$

Table II. The normal distribution function $\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-u^2/2}\, du$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.0	0.500 000	1.5	0.933 193	3.0	0.998 650
0.1	0.539 828	1.6	0.945 201	3.1	0.999 032
0.2	0.579 260	1.7	0.955 435	3.2	0.999 313
0.3	0.617 911	1.8	0.964 070	3.0	0.999 517
0.4	0.655 422	1.9	0.971 283	3.0	0.999 663
0.5	0.691 462	2.0	0.977 250	3.0	0.999 767
0.6	0.725 747	2.1	0.982 136	3.0	0.999 841
0.7	0.758 036	2.2	0.986 097	3.0	0.999 892
0.8	0.788 145	2.3	0.989 276	3.0	0.999 928
0.9	0.815 940	2.4	0.991 802	3.0	0.999 952
1.0	0.841 345	2.5	0.993 790	4.0	0.999 968
1.1	0.864 334	2.6	0.995 339	4.1	0.999 979
1.2	0.884 930	2.7	0.996 533	4.2	0.999 987
1.3	0.903 200	2.8	0.997 445	4.3	0.999 991
1.4	0.919 243	2.9	0.998 134	4.4	0.999 995
				4.5	0.999 997

in the case of $p = 0.5$. This is far beyond the extreme point $x_\alpha = 4.5$ of Table II corresponding to $\alpha = 0.000,003$. In this particular case, we can either accept the incredible event of the probability less than 0.000,003, or just reject the hypothesis $p = 0.5$. Of course, the hypothesis has to be rejected. Using the above data, one can take

$$\hat{p} = 0.5141 \quad \left(= \frac{m}{n}\right)$$

as the corresponding statistical estimate of the unknown probability p. How is it reliable?

The question concerning the reliability of our knowledge about p can be approached as follows. According to the normal approximation, we have

$$-x_\alpha \leqslant \sqrt{\frac{n}{p(1-p)}}\,(\xi - p) \leqslant x_\alpha$$

or, equivalently,

$$\xi - \sqrt{p(1-p)}\,\frac{x_\alpha}{\sqrt{n}} \leqslant p \leqslant \xi + \sqrt{p(1-p)}\,\frac{x_\alpha}{\sqrt{n}}$$

with the probability $1 - 2\alpha$, where $\alpha = 1 - \Phi(x_\alpha)$. Hence we can be a priori sure that

$$\xi - 2\frac{x_\alpha}{\sqrt{n}} \leqslant p \leqslant \xi + 2\frac{x_\alpha}{\sqrt{n}}$$

with the probability $1 - 2\alpha$ at least, since $p(1-p) \leqslant 1/4$. For the presented data, $\xi = 0.5141$ a posteriori, and we can trust the corresponding estimate (called the confidence interval)

$$0.5141 - 0.0003\,x_\alpha \leqslant p \leqslant 0.5141 + 0.0003\,x_\alpha$$

with probability $1 - 2\alpha$.

1.2. ESTIMATION OF PARAMETERS IN A NORMAL SAMPLE

Suppose we observe independent identically distributed random variables $\xi_1, \ldots, \xi_n$; the corresponding data is usually called a *statistical sample* of size n. What can we

say about its unknown probability distribution? Suppose we know that this distribution is *normal*, for example, and then one can ask about the unknown parameters

$$a = \mathbf{E}\xi_k, \quad \sigma^2 = \mathbf{D}\xi_k.$$

One can apply the sample mean

$$\hat{a} = \frac{1}{n}\sum_{k=1}^{n}\xi_k \tag{1.2}$$

as an estimate of the unknown parameter a; how is it reliable? The mean square error of this estimate is

$$\mathbf{E}(\hat{a} - a)^2 = \frac{\sigma^2}{n},$$

which is not very useful if σ^2 is unknown. One can take

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(\xi_k - \hat{a})^2 \tag{1.3}$$

as an *unbiased* estimate of σ^2, i.e.

$$\mathbf{E}\hat{\sigma}^2 \equiv \sigma^2,$$

which easily follows from

$$\sum_{k=1}^{n}(\xi_k - \hat{a})^2 = \sum_{k=1}^{n}\xi_k^2 - \frac{1}{n}\left(\sum_{k=1}^{n}\xi_k\right)^2.$$

Set

$$\tau = \frac{\sqrt{n}(\hat{a} - a)}{\hat{\sigma}}. \tag{1.4}$$

The probability distribution of the random variable τ does not depend on the parameters (a, σ^2), or the substitution of ξ_k by $(\xi_k - a)/\sigma$, $k = 1, \ldots, n$. Assume for a while that $a = 0$, $\sigma^2 = 1$. A linear *orthogonal* transformation

$$\eta_j = \sum_{k=1}^{n} c_{jk}\xi_k, \quad j = 1, \ldots, n,$$

with

$$\eta_1 = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \xi_k$$

in particular, results in *independent* normal random variables η_j, with

$$\mathbf{E}\eta_j = 0, \qquad \mathbf{D}\eta_j = 1, \quad j = 1, \ldots, n.$$

We have

$$\sum_{k=1}^{n} \xi_k^2 = \sum_{j=1}^{n} \eta_j^2$$

and

$$\sum_{k=1}^{n} (\xi_k - \hat{a})^2 = \sum_{k=1}^{n} \xi_k^2 - \frac{1}{n}\left(\sum_{k=1}^{n} \xi_k\right)^2 =$$

$$= \sum_{j=1}^{n} \eta_j^2 - \eta_1^2 = \sum_{j=2}^{n} \eta_j^2.$$

Hence

$$\tau = \sqrt{n-1}\,\frac{\eta_1}{\chi},$$

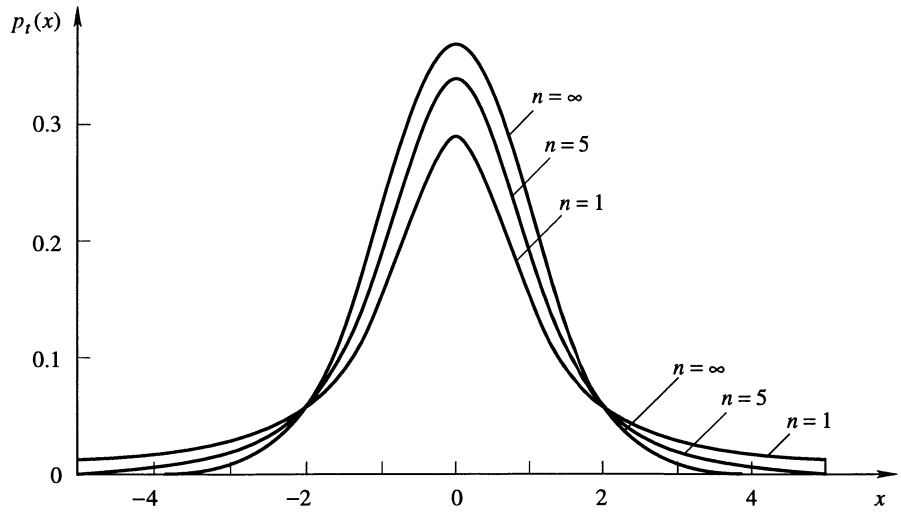

Fig. 13.

Table III. Values of x_α in (bilateral) confidence bounds for $2\alpha = 0.05; 0.02; 0.001$ for Student's distribution with $n-1$ degrees of freedom

$n-1$	0.05	0.02	0.01	0.001	$n-1$	0.05	0.02	0.01	0.001
1	12.71	31.82	63.66	636.6	20	2.086	2.528	2.845	3.850
2	4.303	6.965	9.925	31.60	21	2.080	2.518	2.831	3.819
3	3.182	4.541	5.841	12.92	22	2.074	2.508	2.819	3.792
4	2.776	3.747	4.604	8.610	23	2.069	2.500	2.807	3.767
5	2.571	3.365	4.032	6.869	24	2.064	2.492	2.797	3.745
6	2.447	3.143	3.707	5.959	25	2.060	2.485	2.787	3.725
7	2.365	2.998	3.499	5.408	26	2.056	2.479	2.779	3.707
8	2.306	2.896	3.355	5.041	27	2.052	2.473	2.771	3.690
9	2.262	2.821	3.250	4.781	28	2.048	2.467	2.763	3.674
10	2.228	2.764	3.169	4.587	29	2.045	2.462	2.756	3.659
11	2.201	2.718	3.106	4.437	30	2.042	2.457	2.750	3.646
12	2.179	2.681	3.055	4.318	40	2.0221	2.423	2.704	3.551
13	2.160	2.650	3.012	4.221	50	2.009	2.403	2.678	3.495
14	2.145	2.624	2.977	4.140	60	2.000	2.390	2.660	3.460
15	2.131	2.602	2.947	4.073	70	1.990	2.374	2.639	3.415
16	2.120	2.583	2.921	4.015	80	1.984	2.365	2.626	3.389
17	2.110	2.567	2.898	3.965	90	1.972	2.345	2.601	3.339
18	2.101	2.552	2.878	3.922	100	1.965	2.334	2.586	3.310
19	2.093	2.539	2.861	3.883	∞	1.960	2.326	2.576	3.291

where $\chi^2 = \sum_{j=2}^{n} \eta_j^2$ is a random variable with the chi-square distribution (see p. 76). The joint probability density of *independent* random variables $\eta = \eta_1$ and $\zeta = \chi^2$ is given by

$$p_{\eta,\zeta}(y, z) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \times \frac{1}{2^{(n-1)/2}\Gamma\left(\frac{n-1}{2}\right)} z^{(n-1)/2-1} e^{-z/2},$$

$$-\infty < y < \infty, \quad 0 < z < \infty.$$

The distribution function of the very τ can be obtained in the form

$$F(x) = \int\int_{\sqrt{n-1}\, y \leqslant x\sqrt{z}} p_{\eta,\zeta}(y, z)\, dy\, dz, \quad -\infty < x < \infty,$$

which leads to the probability density of τ:

$$p(x) = \frac{1}{\sqrt{2\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{\sqrt{n-1}} \left(1 + \frac{x^2}{n-1}\right)^{-n/2}, \quad -\infty < x < \infty. \tag{1.5}$$

This is the so-called *Student's distribution* with $(n-1)$ degrees of freedom. Similarly to the normal distribution with parameters $a = 0$, $\sigma^2 = 1$, it is symmetric and bell-shaped (see Figure 13); one can easily verify that it tends to the normal distribution

when $n \to \infty$. Table III gives the corresponding *quantiles* x_α,

$$F(x_\alpha) = 1 - \alpha,$$

for various n and the significance levels $2\alpha = 0.05, 0.02, 0.01, 0.001$. In particular, for any n we have

$$\mathbf{P}\{-x_\alpha \leqslant \tau \leqslant x_\alpha\} = 1 - 2\alpha.$$

Hence, according to (1.4),

$$\hat{a} - \hat{\sigma}x_\alpha/\sqrt{n} \leqslant a \leqslant \hat{a} + \hat{\sigma}x_\alpha/\sqrt{n} \tag{1.6}$$

with the probability $1 - 2\alpha$, where $\hat{a}, \hat{\sigma}$ are the statistics suggested in (1.2), (1.3), respectively. Thus, we get an estimate for the unknown mean value a, in the form of the corresponding *confidence interval* (1.6).

1.3. CHI-SQUARE CRITERION FOR PROBABILITY TESTING

Let us consider the scheme with *disjoint* events A_i, $i = 1, \ldots, r$, formally representing all possible outcomes of an 'experiment'. The problem is to verify how given probabilities

$$p_i = \mathbf{P}(A_i), \quad i = 1, \ldots, r,$$

fit into the real data obtained from n *independent* trials (experiments). Let ξ_{ik} be the indicator of the event A_i in the kth trial, $\xi_{ik} = 1$ if a_i occurs, $\xi_{ik} = 0$ otherwise. Set

$$\nu_i = \sum_{k=1}^{n} \xi_{ik} \quad (i = 1, \ldots, r).$$

Of course, ν_i/n is the *frequency* of the occurrence of A_i in the trial series. Consider

$$\Delta_i = \frac{\nu_i - np_i}{\sqrt{np_i}} = \sum_{k=1}^{n} \frac{\xi_{ik} - p_i}{\sqrt{np_i}}, \quad i = 1, \ldots, r.$$

According to

$$\mathbf{E}(\nu_i - np_i) = 0,$$

$$\mathbf{E}(\nu_i - np_i)(\nu_j - np_j) = \sum_{k=1}^{n} \mathbf{E}(\xi_{ik} - p_i)(\xi_{jk} - p_j)$$

$$= n \begin{cases} -p_i p_j, & i \neq j, \\ p_i(1 - p_i), & i = j, \end{cases}$$

the correlation matrix $B = \{B_{ij}\}$,

$$B_{ij} = \mathbf{E}\Delta_i \Delta_j, \quad i, j = 1, \ldots, r,$$

can be represented as

$$B = I - \begin{pmatrix} \sqrt{p}_1 \\ \vdots \\ \sqrt{p}_r \end{pmatrix} (\sqrt{p}_1, \ldots, \sqrt{p}_r),$$

with the unit matrix I. By applying a linear orthogonal transformation

$$\eta_j = \sum_{i=1}^{r} c_{ij} \Delta_j, \quad i = 1, \ldots, r,$$

with $c_{1j} = \sqrt{p}_j$, $j = 1, \ldots, r$, we obtain random variables η_i, $\mathbf{E}\eta_i = 0$, with the correlation matrix

$$CBC' = I - C \begin{pmatrix} \sqrt{p}_1 \\ \vdots \\ \sqrt{p}_r \end{pmatrix} (\sqrt{p}_1, \ldots, \sqrt{p}_r) C'$$

$$= I - \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} (1, 0, \ldots, 0),$$

Table IV. Values of x_α in confidence bounds for $\alpha = 0.05, 0.01; 0.001$ for chi-square distributions with n degrees of freedom

n	0.05	0.01	0.001	n	0.05	0.01	0.001
1	3.84	6.63	10.8	26	38.9	45.6	54.1
2	5.99	9.21	13.8	27	40.1	47.0	55.5
3	7.81	11.3	16.3	28	41 3	48.3	56.9
4	9.49	13.3	18.5	29	42.6	49.6	58.3
5	11.1	15.1	20.5	30	43.8	50.9	59.7
6	12.6	16.8	22.5	31	45.0	52.2	61.1
7	14.1	18.5	24.3	32	46.2	53.5	62.5
8	15.5	20.1	26.1	33	47.4	54.8	63.9
9	16.9	21.7	27.9	34	48.6	56.1	65.2
10	18.3	23.2	29.6	35	49.8	57.3	66.6
11	19.7	24.7	31.3	36	51.0	58.6	68.0
12	21.0	26.2	32.9	37	52.2	59.9	69.3
13	22.4	27.7	34.5	38	53.4	61.2	70.7
14	23.7	29.1	36.1	39	54.6	62.4	72.1
15	25.0	30.6	37.7	40	55.8	63.7	73.4
16	26.3	32.0	39.3	41	56.8	65.0	74.7
17	27.6	33.4	40.8	42	58.1	66.2	76.1
18	28.9	34.8	42.3	43	59.3	67.5	77.4
19	30.1	36.2	43.8	44	60.5	68.7	78.7
20	31.4	37.6	45.3	45	61.7	70.0	80.1
21	32.7	38.9	46.8	46	62.8	71.2	81.4
22	33.9	40.3	48.3	47	64.0	72.4	82.7
23	35.2	41.6	49.7	48	65.2	73.7	84.0
24	36.4	43.0	51.2	49	66.3	74.9	85.4
25	37.7	44.3	52.6	50	67.5	76.2	86.7

where $C' = \{c_{kj}\}$ is the conjugate matrix to $C = \{c_{jk}\}$, $CC' = I$, i.e.

$$\mathbf{E}\eta_i\eta_j = \begin{cases} 0, & i \neq j \text{ or } i = j = 1, \\ 1, & i = j \ (= 2, \dots, r). \end{cases}$$

In particular,

$$\eta_1 = \sum_{j=1}^{r} \frac{\sqrt{p_j}(\nu_j - np_j)}{\sqrt{np_j}} = \frac{1}{\sqrt{n}}\left(\sum_{j=1}^{r} \nu_j - n\right) = 0.$$

The orthogonal transformation preserves

$$\sum_{j=1}^{r} \Delta_j^2 = \sum_{i=1}^{r} \eta_i^2 \left(= \sum_{i=2}^{r} \eta_i^2 \right),$$

hence

$$\chi^2 = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{\sqrt{np_j}} = \sum_{i=2}^{r} \eta_i^2 \tag{1.7}$$

is the sum of squares of $k-1$ *uncorrelated* random variables η_i with zero mean $\mathbf{E}\eta_i = 0$ and variance $\mathbf{D}\eta_i = 1$.

One can apply to η_i the normal approximation, since

$$\eta_i = \sum_{k=1}^{n} \eta_{ik}$$

are sums of independent identically distributed random variables

$$\eta_{ik} = \sum_{j=1}^{r} c_{ij}\left(\xi_{jk} - p_j\right)/\sqrt{np_j}.$$

Hence the random variable χ^2 of (1.7) has approximately the chi-square distribution with $m = r - 1$ degrees of freedom (see p. 76 and Figure 14). Table IV gives the corresponding quantiles x_α,

$$\mathbf{P}\{\chi^2 \geqslant x_\alpha\} = \alpha,$$

for the significance levels $\alpha = 0.05, 0.01, 0.001$, and $n = 1, \ldots, 50$.

Suppose our data give

$$\chi^2 \geqslant x_\alpha; \tag{1.8}$$

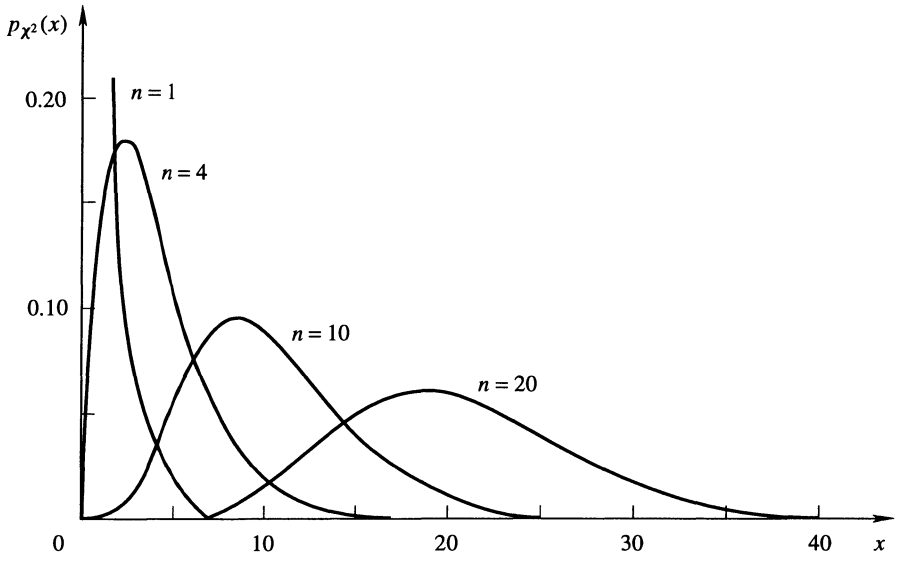

Fig. 14.

then we can either accept the occurrence of the very unlikely event (of a small probability $\leqslant \alpha$), or reject the given hypothesis

$$p_i = \mathbf{P}(A_i), \quad i = 1, \ldots, r,$$

and the chi-squares criterion suggests to make the latter choice.

1.4. SEQUENTIAL ANALYSIS OF ALTERNATIVE HYPOTHESES

Suppose we deal with a sequence of Bernoulli trials having in mind two hypotheses concerning the 'success' probability, namely, H_0: $p = p_0$ and H_1: $p = p_1$, say. We have to decide between H_0 or H_1, using the available data. Of course, one can make a wrong decision, by rejecting a hypothesis when it is true.

Suppose that our preference lies with the hypothesis H_0, in the sense that the probability α_0 of rejecting H_0 when it is true should not exceed a given α_0^*, say; then we are to find a decision rule which at the same time rejects H_1 (when it is true) with a possibly small probability α_1. It is natural to look for a decision rule which satisfies

$$\alpha_0 \leqslant \alpha_0^*, \qquad \alpha_1 \leqslant \alpha_1^*,$$

with given limits α_0^*, α_1^* for our possible errors; of course, such a decision rule requires sufficiently large amount of data (a sufficiently large series of the Bernoulli trials, say).

Let ξ_k be the indicator of 'success' in the kth trial ($k = 1, \ldots, n$), taking values $x = 1, 0$, with the corresponding probability

$$\mathbf{P}(x \mid H_0) = \begin{cases} p_0, & x = 1, \\ 1 - p_0, & x = 0, \end{cases} \qquad \mathbf{P}(x \mid H_1) = \begin{cases} p_1, & x = 1, \\ 1 - p_1, & x = 0, \end{cases}$$

under the hypothesis H_0, H_1, respectively. Introduce the so-called *likelihood ratio*

$$L(x_1, \ldots, x_n) = \frac{\mathbf{P}(x_1, \ldots, x_n \mid H_1)}{\mathbf{P}(x_1, \ldots, x_n \mid H_0)}, \tag{1.9}$$

defined by means of the joint distributions of $\xi_1, \ldots, \xi_n$ under H_0 and H_1, on which the decision rule will be based.

Consider the sequence

$$\log L(\xi_1, \ldots, \xi_n) = \sum_{k=1}^{n} \log \frac{\mathbf{P}(\xi_k \mid H_1)}{\mathbf{P}(\xi_k \mid H_0)}, \quad n = 1, 2, \ldots, \tag{1.10}$$

consisting of sums of independent identically distributed random variables

$$\log \frac{\mathbf{P}(\xi_k \mid H_1)}{\mathbf{P}(\xi_k \mid H_0)}, \quad k = 1, 2, \ldots, n.$$

The elementary inequality

$$p \log t_1 + q \log t_2 < \log (pt_1 + qt_2),$$

$p, q \geqslant 0$, $p + q = 1$, for the concave function $\log t$, $0 < t_1 \leqslant t \leqslant t_2 < \infty$, shows that

$$\begin{aligned} &\mathbf{E} \log \frac{\mathbf{P}(\xi_k \mid H_1)}{\mathbf{P}(\xi_k \mid H_0)} \\ &\quad = p_0 \log \frac{p_1}{p_0} + q_0 \log \frac{q_1}{q_0} < \log (p_1 + q_1) = 0 \end{aligned}$$

under the hypothesis H_0, while

$$\begin{aligned} \mathbf{E} \log \frac{\mathbf{P}(\xi_k \mid H_1)}{\mathbf{P}(\xi_k \mid H_0)} &= -\mathbf{E} \log \frac{\mathbf{P}(\xi_k \mid H_0)}{\mathbf{P}(\xi_k \mid H_1)} \\ &= -\left(p_1 \log \frac{p_0}{p_1} + q_1 \log \frac{q_0}{q_1} \right) \\ &> -\log (p_0 + q_0) = 0 \end{aligned}$$

under the hypothesis H_1. Hence, according to the law of large numbers, with probability 1

$$\log L(\xi_1, \ldots, \xi_n) \longrightarrow -\infty, \tag{1.11}$$

when H_0 is true, and

$$\log L(\xi_1, \ldots, \xi_n) \longrightarrow +\infty \tag{1.11$'$}$$

when H_1 is true. In the case (1.11), the sequence (1.10) is *bounded from above*, and the probability to exceed a high level l_0 tends to zero when $l_0 \to \infty$. Thus, *there is*

an upper level $l_0 > 0$ such that our sequence (1.10) crosses l_0 with the probability less than the given α_0^*. Similarly, in the case (1.11)′ *there is a lower level* $l_1 < 0$ such that our sequence (1.10) crosses l_1 with the probability less than the given α_1^*. At the same time, with the probability 1, our sequence (1.10) crosses the *upper level* $l_0 > 0$ in the case (1.11)′, and the *lower level* $l_1 < 0$ in the case (1.11). By observing the sequence (1.10) for $n = 1, 2, \ldots$, one can accept H_1 when the sequence (1.10) *first exceeds the upper level* $l_0 > 0$, and accept H_1, when it *first exceeds the lower level* $l_1 < 0$. Obviously, this decision rule satisfies $\alpha_0 \leqslant \alpha_0^*$, $\alpha_1 \leqslant \alpha_1^*$, with arbitrarily chosen bounds $\alpha_0^* > 0$, $\alpha_1^* > 0$ of possible errors (recall that α_i is the probability to reject the true hypothesis H_i, $i = 0, 1$).

Everything is fine here except that we don't know how to find the levels l_0, l_1.

Suppose, a decision rule of the above type with l_0, l_1 fits our demands for the error probabilities $\alpha_0 = \alpha_0^*$ and $\alpha_1 = \alpha_1^*$, say. We show that there is another decision rule of the same type corresponding to some other levels l_0, l_1 which can be easily determined by the given α_0^*, α_1^*. Namely, let us consider all $x_1, \ldots, x_n$ $(n = 1, 2, \ldots)$ such that

$$\log L(x_1, \ldots, x_k) < l_0, \quad k = 1, \ldots, n-1, \qquad \log L(x_1, \ldots, x_n) \geqslant l_0;$$

in particular,

$$\mathbf{P}(x_1, \ldots, x_n \mid H_1) \geqslant c\mathbf{P}(x_1, \ldots, x_n \mid H_0)$$

with $\log c = l_0$. Summing up over all such $x_1, \ldots, x_n$ $(n = 1, 2, \ldots)$ the left hand side of the last inequality gives the probability $1 - \alpha_1$ to accept the true hypothesis H_1, while the corresponding sum on the right hand side gives the probability α_0 to reject the true hypothesis H_0. Hence

$$1 - \alpha_1 \geqslant c\alpha_0,$$

or

$$l_0 \leqslant \log \frac{1 - \alpha_1}{\alpha_0},$$

which implies

$$l_0 \leqslant -\log \alpha_0 \quad (\alpha = \alpha_0^*) \tag{1.12}$$

for any $0 \leqslant \alpha_1 \leqslant 1$. In a similar way,

$$l_1 \geqslant \log \frac{\alpha_1}{1-\alpha_0}$$

and

$$l_1 \geqslant \log \alpha_1 \quad (\alpha_1 = \alpha_1^*) \tag{1.12}'$$

for any $0 \leqslant \alpha_0 \leqslant 1$. Choosing

$$l_0 = -\log \alpha_0^*$$

as the new *upper* level (with the given α_0^*) obviously can only diminish the probability α_0 to reject H_0 when it holds true; thus

$$\alpha_0 \leqslant \alpha_0^*.$$

Similarly, if

$$l_1 = \log \alpha_1^*$$

is chosen as the new *lower* level (with the given α_1^*), we get for the corresponding probability α_1 (to reject H_1 when it holds true), according to the general inequality (1.12), that

$$\log \alpha_1 \leqslant l_1 = \log \alpha_1^*,$$

or

$$\alpha_1 \leqslant \alpha_1^*.$$

1.5. BAYESIAN APPROACH TO HYPOTHESES TESTING AND PARAMETERS ESTIMATION

Let us imagine that the 'success' probability p in the observed Bernoulli trials is random (depends on some external random factors). For example, we are given an urn with a random number θ of white balls and $r - \theta$ black balls. A trial consists

of drawing a ball at random (with immediate replacement), with a white ball drawn considered as 'success'.

Suppose, the first n Bernoulli trials resulted in successes. What can we say about the probability of success at the next $(n+1)$th trial? Assuming that all possible numbers $\theta = i,\ i = 0, 1, \dots, r$, of white balls in the urn are *equiprobable*, we obtain the joint probability distribution of θ and $\xi_1, \dots, \xi_n$ (ξ_k is the indicator of the success at kth step, taking values $x_k = 1, 0$):

$$\mathbf{P}\{\theta = i, \xi_1 = x_1, \dots, \xi_n = x_n\} = \frac{1}{r+1}\binom{n}{m} p^m (1-p)^{n-m},$$

with $m = \sum_{k=1}^{n} x_k$ representing the total number of successes and $p = i/r$. Therefore

$$\begin{aligned}\mathbf{P}\{\xi_1 = x_1, \dots, \xi_n = x_n\} &= \sum_{i=0}^{r} \mathbf{P}\{\theta = i, \xi_1 = x_1, \dots, \xi_n = x_n\} \\ &= \frac{1}{r+1} \sum_{i=1}^{r} \binom{n}{m} \left(\frac{i}{r}\right)^m \left(1 - \frac{i}{r}\right)^{n-m}.\end{aligned}$$

The corresponding a posteriori probability at the $(n+1)$th step, given $\xi_1 = 1, \dots, \xi_n = 1$, equals

$$\begin{aligned}&\mathbf{P}\{\xi_{n+1} = 1 \mid \xi_1 = 1, \dots, \xi_n = 1\} \\ &= \frac{\frac{1}{r+1}\sum_{i=1}^{r} \left(\frac{i}{r}\right)^{n+1}}{\frac{1}{r+1}\sum_{i=1}^{r} \left(\frac{i}{r}\right)^{n}} \approx \frac{\int_0^1 t^{n+1}\,dt}{\int_0^1 t^n\,dt} = \frac{n+1}{n+2}.\end{aligned}$$

Of course, $\xi_k = 1,\ k = 1, \dots, n$, for large n suggests that nearly all balls in the urn are white. What is the best estimate of their number $\theta = 0, 1, \dots, r$? As we know, the best estimate $\hat{\theta}$, as a function of the observations $\xi_1, \dots, \xi_n$, is the *maximum point* of the corresponding *a posteriori* probabilities:

$$\pi(\hat{\theta} \mid \xi_1, \dots, \xi_n) = \max_{0 \leqslant i \leqslant r} \pi(i \mid \xi_1, \dots, \xi_n), \tag{1.13}$$

where, for any $\xi_1 = x_1, \dots, \xi_n = x_n$,

$$\begin{aligned}&\pi(i \mid x_1, \dots, x_n) = \mathbf{P}\{\theta = i \mid \xi_1 = x_1, \dots, \xi_n = x_n\} \\ &= \frac{\left(\frac{i}{r}\right)^m \left(1 - \frac{i}{r}\right)^{n-m}}{\sum_{j=0}^{r} \left(\frac{j}{r}\right)^m \left(1 - \frac{j}{r}\right)^{n-m}}, \quad m = \sum_{k=1}^{n} x_k\end{aligned}$$

(see p. 11). In particular, in the extreme case $m = n$ from (1.13) we obtain

$$\hat{\theta} = r.$$

In the general case, consider the best estimate $\hat{\theta}$ of the unknown parameter θ defined in (1.13), where

$$\pi(i \mid \xi_1, \ldots, \xi_n) = \frac{\pi(i)\,\mathbf{P}(\xi_1, \ldots, \xi_n \mid \theta = i)}{\sum_\theta \pi(\theta)\,\mathbf{P}(\xi_1, \ldots, \xi_n \mid \theta)}, \quad i = 0, 1, \ldots, r \tag{1.14}$$

correspond to arbitrary *a priori* probabilities

$$\pi(i) = \mathbf{P}\{\theta = i\}, \quad i = 0, 1, \ldots, r.$$

Consider the *likelihood ratio*

$$L(x_1, \ldots, x_n \mid \theta) = \frac{\mathbf{P}(x_1, \ldots, x_n \mid \theta)}{\mathbf{P}(x_1, \ldots, x_n \mid \theta_0)}$$

of the joint distributions of $\xi_1, \ldots, \xi_n$, for $\theta = 0, 1, \ldots, r$ and some θ_0 (see p. 141). Namely, as it was actually shown (see (1.10), (1.11)), for $\theta \neq \theta_0$

$$\log L(\xi_1, \ldots, \xi_n \mid \theta) \to -\infty, \qquad L(\xi_1, \ldots, \xi_n \mid \theta) \to 0$$

with probability 1, provided θ_0 is the true value of θ. Together with (1.14) this implies

$$\pi(\theta_0 \mid \xi_1, \ldots, \xi_n) = \frac{\pi(\theta_0)}{\pi(\theta_0) + \sum_{\theta \neq \theta_0} \pi(\theta) L(\xi_1, \ldots, \xi_n \mid \theta)} \longrightarrow 1, \quad n \to \infty,$$

with probability 1 provided $\pi(\theta_0) \neq 0$. Therefore,

$$\pi(\theta_0 \mid \xi_1, \ldots, \xi_n) = \max_{1 \leqslant i \leqslant r} (i \mid \xi_1, \ldots, \xi_n)$$

for sufficiently large n, which implies

$$\hat{\theta} = \theta_0$$

according to (1.13). Of course, for a *given* n one cannot be sure that $\hat{\theta} = \theta_0$, although in any case

$$\hat{\theta} \to \theta_0 \quad (n \to \infty) \tag{1.15}$$

with probability 1 (this is called the *consistency* property of the estimate $\hat{\theta}$).

For a priori equiprobable $\{\theta = i\}$ with $\pi(i) = \mathbf{P}\{\theta = i\} = 1/(r+1)$, $i = 0, 1, \ldots, r$, the best estimate $\hat{\theta}$, obtained from (1.13), (1.14), is the *most likely* one, in the sense that it maximizes the conditional probability

$$\mathbf{P}(\xi_1, \ldots, \xi_n \mid \theta), \quad \theta = 0, 1, , \ldots, r,$$

for given observations $\xi_1, \ldots, \xi_n$.

1.6. MAXIMUM LIKELIHOOD METHOD

Suppose we observe discrete random variables $\xi_1, \ldots, \xi_n$, whose joint probability distribution $\mathbf{P}(x_1, \ldots, x_n \mid \theta)$ depends on an unknown parameter $\theta \in \Theta$, which we wish to estimate. The *maximum likelihood* method suggests an estimate $\hat{\theta}$ which is the most likely one in the sense that it is the maximum point $\theta = \hat{\theta}$ of the probability $\mathbf{P}(x_1, \ldots, x_n \mid \theta)$, $\theta \in \Theta$, namely

$$\hat{\theta}\colon\ \mathbf{P}(x_1, \ldots, x_n \mid \hat{\theta}) = \max_{\theta} \mathbf{P}(x_1, \ldots, x_n \mid \theta), \tag{1.16}$$

given observations $\xi_1 = x_1, \ldots, \xi_n = x_n$.

EXAMPLE. Suppose, $\xi_k = x_k$, $k = 1, \ldots, n$, are the indicators of the success in Bernoulli trials with unknown success probability $p = \theta$, $0 < \theta < 1$. Then the maximum likelihood estimate is the frequency:

$$\hat{p} = \frac{1}{n} \sum_{k=1}^{n} x_k,$$

which can be found as the maximum point $\theta = \hat{p}$ of

$$\log \mathbf{P}(x_1, \ldots, x_n \mid \theta)$$

$$= \log \binom{n}{m} + m \log \theta + (n - m) \log (1 - \theta), \quad m = \sum_{k=1}^{n} x_k.$$

EXAMPLE. Suppose, $\xi_k = x_k$ $(k = 1, \dots, n)$ represent a statistical sample of independent random variables having a Poisson distribution with parameter $\lambda = \theta$, $\theta > 0$. Then the maximum likelihood estimate is the *sample mean*

$$\hat{\lambda} = \frac{1}{n} \sum_{k=1}^{n} x_k,$$

which can be found as the maximum point $\theta = \hat{\lambda}$ of

$$\log \mathbf{P}(x_1, \dots, x_n \mid \theta) = \log \frac{1}{x_1! \dots x_n!} + \sum_{k=1}^{n} \log \theta x_k - n\theta.$$

In the case of continuous random variables $\xi_1, \dots, \xi_n$ with the joint probability density $p(x_1, \dots, x_n \mid \theta)$ depending on unknown parameter $\theta \in \Theta$, the *maximum likelihood* estimate $\hat{\theta}$ is defined as the maximum point of $p(x_1, \dots, x_n \mid \theta)$, $\theta \in \Theta$, i.e.

$$\hat{\theta}\colon\ p(x_1, \dots, x_n \mid \hat{\theta}) = \max_{\theta} p(x_1, \dots, x_n \mid \theta) \tag{1.16}'$$

for observed values $\xi_1 = x_1, \dots, \xi_n = x_n$.

EXAMPLE. Suppose, $\xi_k = x_k$ $(k = 1, \dots, n)$ represent a statistical sample of independent random variables having a normal distribution with parameter $\theta = (a, \sigma^2)$, $-\infty < a < \infty$, $\sigma^2 > 0$, where

$$a = \mathbf{E}\xi_k, \qquad \sigma^2 = \mathbf{D}\xi_k \quad (k = 1, \dots, n).$$

One can easily verify that the maximum likelihood estimate $\hat{\theta} = (\hat{a}, \hat{\sigma}^2)$ is given by

$$\hat{a} = \frac{1}{n} \sum_{k=1}^{n} x_k, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{a})^2.$$

EXAMPLE. Consider a statistical sample $\xi_k = x_k$ $(k = 1, \dots, n)$ of independent random variables having a *Laplace distribution* with the probability density

$$p(x \mid \theta) = \frac{\lambda}{2} \mathrm{e}^{-\lambda |x - \theta|}, \quad -\infty < x < \infty,$$

where θ, $-\infty < \theta < \infty$, is the unknown shift parameter,

$$\theta = \mathbf{E}\xi_k \quad (k = 1, \ldots, n).$$

The maximum likelihood estimate can be found as the maximum point $\theta = \hat{\theta}$ of

$$\log p(x_1, \ldots, x_n \mid \theta) = n \log \frac{\lambda}{2} - \lambda \sum_{k=1}^{n} |x_k - \theta|,$$

which is a piecewise linear function in θ, attaining its maximum at one of the points x_k, $k = 1, \ldots, n$. Considering these points in their natural order:

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$$

on the real line, we get

$$\sum_{k=1}^{n} |x_{(j+1)} - x_{(k)}| - \sum_{k=1}^{n} |x_{(j)} - x_{(k)}|$$
$$= -(x_{(j+1)} - x_{(j)})(n - 2j).$$

We see that the maximum point $\hat{\theta} = x_{(j)}$ is with $j = m$ if $n = 2m$, or with $j = m+1$ if $n = 2m + 1$; the estimate

$$\hat{\theta} = x_{(m)}$$

is called the *sample median.*

1.7. SAMPLE DISTRIBUTION FUNCTION AND THE METHOD OF MOMENTS

Suppose we observe a statistical sample $\xi_k = x_k$ $(k = 1, \ldots, n)$ of independent identically distributed random variables, whose probability distribution function $F(x)$, $-\infty < x < \infty$, is unknown. By introducing the ordered sequence

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$$

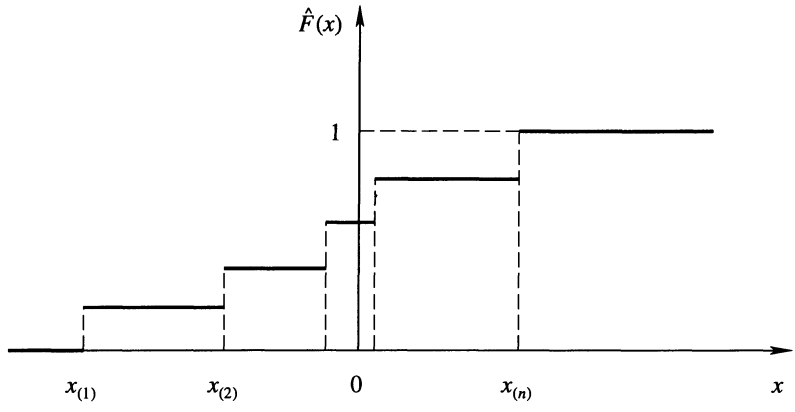

Fig. 15.

(called the *variation series*), one can define the *sample distribution* function

$$\hat{F}(x) = \begin{cases} 0, & x < x_{(1)}, \\ \frac{m}{n}, & x_{(m)} \leqslant x < x_{(m+1)}, \\ 1, & x \geqslant x_{(n)} \end{cases} \qquad m = 1, \ldots, n-1, \tag{1.17}$$

(see Figure 15). According to the law of large numbers, for any x, $-\infty < x < \infty$,

$$\hat{F}(x) \to F(x)$$

as $n \to \infty$.

Suppose, the unknown probability distribution depends on a r-dimensional parameter $\theta = (\theta_1, \ldots, \theta_r)$ which is determined by the moments

$$a_m(\theta) = \mathbf{E}\xi_k^m, \quad m = 1, \ldots, r,$$

by a one-to-one continuous mapping

$$\theta = (\theta_1, \ldots, \theta_r) \longleftrightarrow (a_1, \ldots, a_m) = a.$$

Then, we can apply the so-called *sample moments*

$$\hat{a}_m = \frac{1}{n}\sum_{k=1}^{n} x_k^m, \quad m = 1, \ldots, r,$$

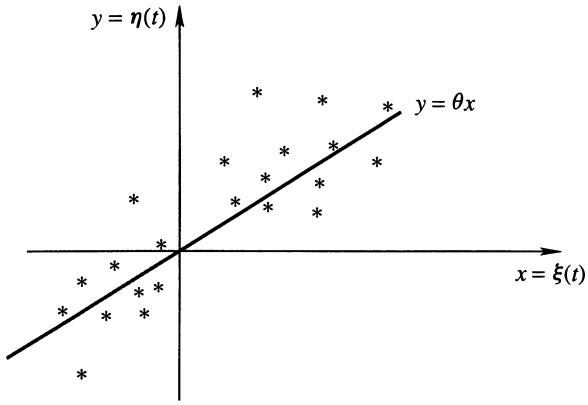

Fig. 16.

which represent the corresponding moments with respect to the *sample probability* distribution (1.17), and estimate the unknown θ by solving the system of equations

$$a_m(\theta) = \hat{a}_m, \quad m = 1, \dots, r. \tag{1.18}$$

The solution $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_r)$ continuously depends on $\hat{a} = (\hat{a}_1, \dots, \hat{a}_r)$ and

$$\hat{\theta} \to \theta \quad (n \to \infty) \tag{1.19}$$

with probability 1. Indeed,

$$\hat{a} = (\hat{a}_1, \dots, \hat{a}_r) \to a = (a_1, \dots, a_r)$$

according to the law of large numbers:

$$\hat{a}_m = \frac{1}{n} \sum_{k=1}^{n} \xi_k^m \to \mathbf{E}\xi_k^m = a_m(\theta), \quad m = 1, \dots, r,$$

and θ is the solution of the limit equation:

$$a_m(\theta) = \hat{a}_m (= a_m), \quad m = 1, \dots, r.$$

1.8. THE METHOD OF LEAST SQUARES

Suppose, $\xi(t)$ and $\eta(t)$, $t = 1, 2, \dots$, are related by

$$\eta(t) = \theta\xi(t) + \Delta(t) \tag{1.20}$$

where θ is an unknown *constant*, and $\Delta(t)$ are 'small perturbations', represented by independent random variables with mean value $\mathbf{E}\Delta(t) = 0$ and variance $\mathbf{D}\Delta(t) \leqslant \sigma^2$, which are uncorrelated with $\xi(t)$, say. Let be given observations of $\xi(t)$, $\eta(t)$ at 'times' $t = 1, \ldots, n$ (e.g., the data in Figure 16); how can we estimate the unknown parameter θ which characterizes the linear dependence between $x = \xi(t)$ and $y = \eta(t)$? The *method of least squares* suggests the following algorithm. Consider $\xi = \xi(t)$, $t = 1, \ldots, n$, and $\eta = \eta(t)$, $t = 1, \ldots, n$, as vectors of the $\mathbb{R}^n$-space. The best approximation

$$\hat{\eta} = \hat{\theta}\xi$$

of $\eta \in \mathbb{R}^n$, by means of all vectors $\theta\xi$, $-\infty < \theta < \infty$, is given by the orthogonal projection of the vector $\eta \in \mathbb{R}^n$ on the linear subspace of all $\theta\xi$, $-\infty < \theta < \infty$, and is determined by the orthogonality condition

$$(\eta - \hat{\eta}, \xi) = \sum_{t=1}^{n} \big[\eta(t) - \hat{\eta}(t)\big]\xi(t) = 0$$

in the $\mathbb{R}^n$-space. It gives the equation

$$\hat{\theta} \sum_{t=1}^{n} \xi(t)^2 = \sum_{t=1}^{n} \xi(t)\eta(t)$$

and the corresponding *least squares estimate*

$$\hat{\theta} = \frac{\sum_{t=1}^{n} \xi(t)\eta(t)}{\sum_{t=1}^{n} \xi(t)^2}. \tag{1.21}$$

How close is $\hat{\theta}$ to the true value θ?

According to (1.20),

$$\sum_{t=1}^{n} \xi(t)\eta(t) = \theta \sum_{t=1}^{n} \xi(t)^2 + \sum_{t=1}^{n} \xi(t)\Delta(t),$$

hence

$$(\hat{\theta} - \theta) \sum_{t=1}^{n} \xi(t)^2 = \sum_{t=1}^{n} \xi(t)\Delta(t)$$

and one can expect in general that

$$\hat{\theta} - \theta = \frac{\sum_{t=1}^{n} \xi(t)\Delta(t)}{\sum_{t=1}^{n} \xi(t)^2} \longrightarrow 0 \tag{1.22}$$

when

$$\sum_{t=1}^{n} \xi(t)^2 \longrightarrow \infty \quad (n \to \infty).$$

EXAMPLE (*Correlation estimate*). Suppose, in the scheme (1.20) we have random variables $\xi(t)$ with $\mathbf{E}\xi(t) = 0$, $\mathbf{D}\xi(t) = 1$, so that the parameter θ represents the *correlation*

$$\theta = \mathbf{E}\xi(t)\eta(t).$$

Moreover, assume that we deal with independent trials at 'times' $t = 1, \ldots, n$, and that the random variables $\xi(t)$ and $\eta(t)$ have the same joint distribution for each t. Then, according to the law of large numbers, with probability 1

$$\frac{1}{n} \sum_{t=1}^{n} \xi(t)^2 \longrightarrow 1$$

and

$$\frac{1}{n} \sum_{t=1}^{n} \xi(t)\Delta(t) \longrightarrow 0.$$

Thus, (1.22) holds with probability 1.

EXAMPLE (*Trend estimation*). Suppose, $\xi(t) = x(t)$, $t = 1, 2, \ldots$, in (1.20) represent a deterministic function of t, which characterizes how the mean value

$$y(t) = \mathbf{E}\eta(t) = \theta x(t)$$

varies in 'time', and the random variables, $\Delta(t)$, $t = 1, 2, \ldots$, are uncorrelated between themselves. Then, with

$$\sum_{t=1}^{n} x(t)^2 \longrightarrow \infty \quad (n \to \infty),$$

we have

$$\mathbf{E}\left[\sum_{t=1}^{n} x(t)\,\Delta(t)\right]^2 = \sum_{t=1}^{n} x(t)^2\,\mathbf{E}\Delta(t)^2 \leqslant \sigma^2 \sum_{t=1}^{n} x(t)^2$$

and

$$\mathbf{E}\,(\hat{\theta}-\theta)^2 = \frac{\mathbf{E}\left[\sum_{t=1}^{n} x(t)\,\Delta(t)\right]^2}{\left[\sum_{t=1}^{n} x(t)^2\right]^2} \leqslant \frac{\sigma^2}{\sum_{t=1}^{n} x(t)^2} \longrightarrow 0.$$

Thus, (1.22) holds in the square mean.

2. Optimality of Statistical Decisions

2.1. THE MOST POWERFUL CRITERION

Let be given a statistical sample of random variables $\xi_1, \ldots, \xi_n$, and their joint probability distribution depending on some unknown parameter θ. We are to make certain decision about $\theta \in \Theta$. The decision rule will be based on the corresponding *likelihood ratio*

$$L(x \mid \theta), \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad \theta \in \Theta$$

as a function of the parameter θ. In the case of discrete probability distribution $\mathbf{P}(x \mid \theta)$ of $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$, the likelihood ratio is defined by

$$L(x \mid \theta) = \frac{\mathbf{P}(x \mid \theta)}{\mathbf{P}(x \mid \theta_0)},$$

while, in the case when the probability density $p(x \mid \theta)$ of $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$ exists,

$$L(x \mid \theta) = \frac{p(x \mid \theta)}{p(x \mid \theta_0)}.$$

In any case, $L(x \mid \theta)$ is assumed to have the property that for any function $\varphi(\xi)$ of $\xi = (\xi_1, \ldots, \xi_n)$, $\mathbf{E}\varphi(\xi)$ depends on $\theta \in \Theta$ in such a way that

$$\mathbf{E}\varphi(\xi) = \mathbf{E}_0 \varphi(\xi) L(\xi \mid \theta), \tag{2.1}$$

where

$$\mathbf{E}\varphi(\xi) = \mathbf{E}_\theta\varphi(\xi), \qquad \mathbf{E}_0\varphi(\xi) = \mathbf{E}_{\theta_0}\varphi(\xi)$$

and $\theta_0 \in \Theta$ is a fixed point.

Let two hypotheses H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$ be given; we have to choose one of them according to a given statistical sample $\xi = (\xi_1, \ldots, \xi_n) \subseteq \mathbb{R}^n$. We consider criteria of the following type. Namely, we choose a *critical region* $S \subseteq \mathbb{R}^n$, and reject or accept the hypothesis H_0, depending on whether $\xi \in S$ or $\xi \notin S$.

By rejecting H_0, we accept H_1; suppose, our preference lies with H_0, and we have to choose the critical region $S \subseteq \mathbb{R}^n$ in such a way that the probability to reject the true H_0 is

$$\alpha_0 = \mathbf{P}\{\xi \in S \mid \theta_0\} \leqslant \alpha_0^*, \tag{2.2}$$

where α_0^* is a given bound for the error probability with respect to H_0: $\theta = \theta_0$. A *criterion* S which satisfies condition (2.2), is called the *most powerful*, if the probability of making an error when H_1: $\theta = \theta_1$ is true *is minimal*:

$$\alpha_1 = \mathbf{P}\{\xi \notin S \mid \theta_1\} = \min. \tag{2.3}$$

Let us show that the *Neyman–Pearson criterion*

$$S_* = \{x\colon L(x \mid \theta_1) > c\}, \tag{2.4}$$

where $c > 0$ is choosen from the condition

$$\alpha_0^* = \mathbf{P}\{\xi \in S_* \mid \theta_0\},$$

is the most powerful.

To do so, let us compare this criterion with any other criterion based on a critical region $S \subseteq \mathbb{R}^n$, which satisfies the corresponding condition (2.2). Applying (2.1) to the indicator function $\varphi(x) = 1_{S_*^c}(x)$, $x \in \mathbb{R}^n$ of the complement S_*^c of S_* in $\mathbb{R}^n$, we have

$$\alpha_1^* = \mathbf{P}\{\xi \notin S_* \mid \theta_1\} = \mathbf{E}_{\theta_1} 1_{S_*^c}(\xi) = \mathbf{E}_{\theta_0} 1_{S_*^c}(\xi)\, L(\xi \mid \theta_1).$$

Next,

$$\alpha_1 - \alpha_1^* = \mathbf{E}_{\theta_0}\left[1_{S^c}(\xi) - 1_{S_*^c}(\xi)\right] L(\xi \mid \theta_1)$$
$$= \mathbf{E}_{\theta_0} 1_{S_*\setminus\Delta}(\xi)\, L(\xi \mid \theta_1) - \mathbf{E}_{\theta_0} 1_{S\setminus\Delta}(\xi)\, L(\xi \mid \theta_1)$$

with $\Delta = S_* \cap S$, where

$$\mathbf{E}_{\theta_0} 1_{S\setminus\Delta}(\xi) L(\xi \mid \theta_1) \leqslant c\mathbf{E}_{\theta_0} 1_{S\setminus\Delta}(\xi)$$
$$\leqslant c\mathbf{E}_{\theta_0} 1_{S_*\setminus\Delta}(\xi) \leqslant \mathbf{E}_{\theta_0} 1_{S_*\setminus\Delta}(\xi)\, L(\xi \mid \theta_1)$$

since $L(x \mid \theta_1) \leqslant c,\ x \in S\setminus\Delta$,

$$\mathbf{E}_{\theta_0} 1_{S\setminus\Delta}(\xi) = \mathbf{E}_{\theta_0} 1_S(\xi) - \mathbf{E}_{\theta_0} 1_\Delta(\xi)$$
$$= \alpha_0 - \mathbf{E}_{\theta_0} 1_\Delta(\xi)$$
$$\leqslant \alpha_0^* - \mathbf{E}_{\theta_0} 1_\Delta(\xi) = \mathbf{E}_{\theta_0} 1_{S_*\setminus\Delta}(\xi)$$

and $L(x \mid \theta_1) > c,\ x \in S_*\setminus\Delta$. Hence

$$\alpha_1^* \leqslant \alpha_1.$$

Let us formulate our result as follows.

THEOREM. *The Neyman–Pearson criterion is the most powerful.*

2.2. SUFFICIENT STATISTICS

Sometimes, one has to make a decision about the unknown parameter $\theta \in \Theta$ of the probability distribution of $\xi = (\xi_1, \ldots, \xi_n)$, by means of incomplete data in the form of a function $\eta = f(\xi) \in \mathbb{R}^m$ of the statistical sample $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$. The corresponding $\eta = f(\xi)$ is considered as a *sufficient* statistic, if it carries the same information about $\theta \in \Theta$ as ξ. This sounds all right, but what does it actually mean? A rigorous answer can be given as follows: for any given $\eta = f(\xi) = y$, the conditional probability distribution of $\xi = (\xi_1, \ldots, \xi_n)$ does not depend on $\theta \in \Theta$, i.e., conditioned on $f(x) = y$, all possible values $\xi = x \in \mathbb{R}^n$ are distributed in the

same way independently of θ. Thus, our knowledge about $\xi = x$ itself, in addition to a given $f(x) = y$, cannot help in making our decision about the true θ.

For a discrete distribution $\mathbf{P}(x \mid \theta)$, $x \in \mathbb{R}^n$ any *sufficient statistic* f of $\xi \in \mathbb{R}^n$ can be characterized by the fact that the corresponding likelihood ratio $L(x \mid \theta)$ is a function of $y = f(x)$, $x \in \mathbb{R}^n$, only:

$$L(x \mid \theta) = g(f(x) \mid \theta). \tag{2.5}$$

Indeed, according to (2.5), or

$$\mathbf{P}(x \mid \theta) = \mathbf{P}(x \mid \theta_0) g(f(x) \mid \theta),$$

the probability distribution of $\eta = f(\xi)$ satisfies

$$\begin{aligned} \mathbf{P}_\eta(y \mid \theta) &= \sum_{x:\, f(x)=y} \mathbf{P}(x \mid \theta_0) g(f(x) \mid \theta) \\ &= \mathbf{P}_\eta(y \mid \theta_0) g(y \mid \theta), \end{aligned}$$

and the conditional probability distribution ξ given $f(\xi) = y$ is

$$\mathbf{P}_\xi(x \mid y) = \frac{\mathbf{P}(x \mid \theta)}{\mathbf{P}_\eta(y \mid \theta)} = \frac{\mathbf{P}(x \mid \theta_0)}{\mathbf{P}_\eta(y \mid \theta_0)},$$

for all x such that $f(x) = y$. On the other hand, for any sufficient statistic f we have

$$\mathbf{P}_\xi(x \mid f(x)) = \frac{\mathbf{P}(x \mid \theta)}{\mathbf{P}_\eta(f(x) \mid \theta)} = \frac{\mathbf{P}(x \mid \theta_0)}{\mathbf{P}_\eta(f(x) \mid \theta_0)},$$

hence

$$L(x \mid \theta) = \frac{\mathbf{P}(x \mid \theta)}{\mathbf{P}(x \mid \theta_0)} = \frac{\mathbf{P}_\eta(f(x) \mid \theta)}{\mathbf{P}_\eta(f(x) \mid \theta_0)} = g(f(x) \mid \theta) \tag{2.5$'$}$$

is actually a function of $y = f(x)$ alone.

EXAMPLE (*Sufficient statistic for a Bernoulli sample*). Let $x = (x_1, \dots, x_n)$ be a statistical sample, representing indicators x_k of 'success' in n Bernoulli trials, with success probability $p = \theta$, $0 < \theta < 1$. Then $f(x) = \sum_{k=1}^{n} x_k$ is a *sufficient statistic*, since

$$\mathbf{P}(x \mid \theta) = \theta^{f(x)}(1-\theta)^{1-f(x)}$$

and the corresponding representation (2.5) holds. This sufficient statistic can be applied, in particular, to the estimate

$$\hat{p} = \frac{1}{n} f(x) = \frac{1}{n} \sum_{k=1}^{n} x_k$$

of the parameter $p = \theta$.

EXAMPLE (*Sufficient statistic for a Poisson sample*). Let $x = (x_1, \dots, x_n)$ be a statistical sample from a Poisson distribution with mean value $a = \theta$, $\theta > 0$. Then $f(x) = \sum_{k=1}^{n} x_k$ is a sufficient statistic, as

$$\mathbf{P}(x \mid \theta) = \frac{\theta^{f(x)}}{x_1! \dots x_n!} \mathrm{e}^{-n\theta}$$

and (2.5) holds again. The above statistic appears in the well-known estimate

$$\hat{a} = \frac{1}{n} f(x) = \frac{1}{n} \sum_{k=1}^{n} x_k$$

of $a = \theta$.

A characterization similar to (2.5) of sufficient statistics can be obtained when $\xi = (\xi_1, \dots, \xi_n)$ has a probability density

$$p(x \mid \theta), \quad x = (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Recall that we defined the conditional distribution of ξ with respect to $\eta = f(\xi)$, by assuming that $\xi \in \mathbb{R}^n$ can be written as $\xi = (\eta, \zeta)$, with the components $\eta \in \mathbb{R}^n$, $\zeta \in \mathbb{R}^{n-m}$ having the joint probability density

$$p_{\eta,\zeta}(y, z \mid \theta) = p(x \mid \theta),$$
$$y \in \mathbb{R}^m, \quad z \in \mathbb{R}^{n-m}, \quad x = (y, z) \in \mathbb{R}^n,$$

so that the corresponding conditional probability density is

$$p_\zeta(z \mid y) = \frac{p_{\eta,\zeta}(y, z \mid \theta)}{p_r(y \mid \theta)}, \quad z \in \mathbb{R}^m$$

(see p. 62). For a *sufficient statistic* $\eta = f(\xi)$, $p_\zeta(z \mid y)$ does not depend on θ and the likelihood ratio is

$$\begin{aligned} L(x \mid \theta) &= \frac{p_{\eta,\zeta}(y, z \mid \theta)}{p_{\eta,\zeta}(y, z \mid \theta_0)} = \frac{p_\eta(y \mid \theta) p_\zeta(z \mid y)}{p_\eta(y \mid \theta_0) p_\zeta(z \mid y)} \\ &= \frac{p_\eta(y \mid \theta)}{p_\eta(y \mid \theta_0)} = g(y \mid \theta), \quad y = f(x). \end{aligned} \tag{2.5}'$$

On the other hand, representation (2.5) with

$$L(x \mid \theta) = \frac{p_{\eta,\zeta}(y, z \mid \theta)}{p_{\eta,\zeta}(y, z \mid \theta_0)} = g(y \mid \theta)$$

gives

$$p_{\eta,\zeta}(y, z \mid \theta) = p_{\eta,\zeta}(y, z \mid \theta_0) g(y \mid \theta)$$

and

$$\begin{aligned} p_\eta(y \mid \theta) &= \int_{\mathbb{R}^{n-m}} p_{\eta,\zeta}(y, z \mid \theta_0) g(y \mid \theta) \, \mathrm{d}z \\ &= p_\eta(y \mid \theta_0) g(y \mid \theta). \end{aligned}$$

Consequently, the conditional probability density

$$p_\zeta(z \mid y) = \frac{p_{\eta,\zeta}(y, z \mid \theta)}{p_\eta(y \mid \theta)} = \frac{p_{\eta,\zeta}(y, z \mid \theta_0)}{p_\eta(y \mid \theta_0)}$$

is independent of θ. Thus, the characterization of sufficient statistics given in (2.5), remains true in this case.

EXAMPLE (*Sufficient statistic for a normal sample*). Let $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ be a statistical sample representing independent normal variables ξ_k, with

$$a = \mathbf{E}\xi_k, \qquad \sigma^2 = \mathbf{D}\xi_k \quad (k = 1, \dots, n).$$

Then

$$f(x) = (f_1(x), f_2(x)), \quad f_1(x) = \sum_{k=1}^{n} x_k, \ f_2(x) = \sum_{k=1}^{n} x_k^2$$

is a sufficient statistic of the parameter

$$\theta = (a, \sigma^2), \quad -\infty < a < \infty, \ \sigma^2 > 0,$$

since

$$p(x \mid \theta) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n} x_k^2 - 2a\sum_{k=1}^{n} x_k + na^2\right)\right\},$$
$$x = (x_1, \dots, x_n) \in \mathbb{R}^n,$$

and the corresponding representation (2.5) holds. The sufficient statistic $f = (f_1, f_2)$ appears in the well-known estimates

$$\hat{a} = \frac{1}{n} f_1 = \frac{1}{n}\sum_{k=1}^{n} x_k,$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\left[f_2(x) - \frac{f_1^2(x)}{n}\right] = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \hat{a}\right)^2.$$

Suppose, we want to apply $\varphi(x)$ as an estimate of a *component* θ of the unknown parameter of the probability distribution of $\xi = (\xi_1, \dots, \xi_n)$, given a statistical sample $x = (x_1, \dots, x_n) \in \mathbb{R}^n$; here, θ is a *scalar* (real) component, and $\varphi = \varphi(x)$, $x \in \mathbb{R}^n$ a (real) function. The accuracy of the estimate can be characterized by the corresponding *mean square error*

$$\| \varphi - \theta \|^2 = \mathbf{E}|\varphi(\xi) - \theta|^2.$$

Consider a *sufficient statistic* $\eta = f(\xi)$ given by a function $y = f(x) \in \mathbb{R}^m$ of $x = (x_1, \ldots, x_n)$. The conditional expectation

$$\psi(y) = \mathbf{E}[\varphi(\xi) \mid f(\xi) = y]$$

does not depend on the unknown parameter and represents a function $\psi = \psi(y)$ of $y = f(x)$, $x \in \mathbb{R}^n$. Clearly,

$$\begin{aligned}&\mathbf{E}\big[|\varphi(\xi) - \theta|^2 \mid f(\xi) = y\big]\\ &\quad = \mathbf{E}\big[|\varphi(\xi) - \psi(y)|^2 \mid f(\xi) = y\big] + |\psi(y) - \theta|^2.\end{aligned}$$

By the total mathematical expectation formula (see p. 60),

$$\begin{aligned}\mathbf{E}|\varphi(\xi) - \theta|^2 &= \mathbf{E}|\psi(\eta) - \theta|^2 + \mathbf{E}|\varphi(\xi) - \psi(\eta)|^2\\ &\geqslant \mathbf{E}|\psi(\eta) - \theta|^2, \quad \eta = f(\xi).\end{aligned} \tag{2.6}$$

This inequality shows that $\psi = \psi(y)$, as function of $y = f(x)$, $x \in \mathbb{R}^n$ gives a *better* estimate of θ than $\varphi = \varphi(x)$, $x \in \mathbb{R}^n$. Moreover, if the *estimate* φ is *unbiased*:

$$\mathbf{E}\varphi(\xi) \equiv \theta,$$

then ψ is of the same type:

$$\mathbf{E}\psi(\eta) \equiv \theta, \tag{2.7}$$

since for

$$\psi(\eta) = \mathbf{E}[\varphi(\xi) \mid \eta = f(\xi)]$$

the total mathematical expectation formula gives

$$\mathbf{E}\mathbf{E}[\varphi(\xi) \mid \eta = f(\xi)] = \mathbf{E}\varphi(\xi).$$

EXAMPLE (*The best estimate for exponential distribution*). Let $x = (x_1, \dots, x_n)$ be a statistical sample representing independent random variables $\xi_k > 0$, distributed according to an exponential law with parameter $\lambda = \theta$, $\theta > 0$, with the joint probability density is given by

$$p(x \mid \theta) = \theta^n \mathrm{e}^{-\theta \sum_{k=1}^n x_k} \quad (x_k > 0,\ k = 1, \dots, n).$$

Then

$$f(x) = \sum_{k=1}^n x_k$$

is a *sufficient statistic*, and the random variable $\eta = f(\xi) > 0$ has the probability density

$$p_n(y \mid \theta) = \frac{1}{(n-1)!}\, \theta^n y^{n-1} \mathrm{e}^{-\theta y}, \quad y > 0,$$

which depends on n (see p. 43 on gamma-distribution). Here, there is only one *unbiased* estimate $\psi(y)$ of the parameter $\theta > 0$ since the function $[\psi(y)\, y^{n-1}]$, $y > 0$, is uniquely determined by its Laplace transform

$$\int_0^\infty \left[\psi(y)\, y^{n-1}\right] \mathrm{e}^{-\theta y}\, \mathrm{d}y = (n-1)!\, \theta^{-n+1}, \quad \theta > 0.$$

One can easily see that

$$\int_0^\infty y^{-1}\, p_n(y \mid \theta)\, \mathrm{d}y = \frac{\theta}{n-1} \int_0^\infty p_{n-1}(y \mid \theta)\, \mathrm{d}y \equiv \frac{\theta}{n-1},$$

hence

$$\psi(y) = (n-1) y^{-1}, \quad y = \sum_{k=1}^n x_k,$$

is the *unbiased* estimate of $\theta > 0$. Actually, it is the *best unbiased estimate* since, for any unbiased estimate $\varphi = \varphi(x)$, a better one is given by

$$\psi(y) = \mathbf{E}\left[\varphi(\xi) \mid \sum_{k=1}^n \xi_k = y\right], \quad y > 0,$$

hence it coincides with $\psi(y) = (n-1) y^{-1}$.

2.3. LOWER BOUND FOR THE MEAN SQUARE ERROR

In our discussion of estimates φ of a *component* θ of the unknown parameter of the probability distribution of $\xi = (\xi_1, \ldots, \xi_n)$, we assume certain regularity conditions on the corresponding likelihood ratio $L(\xi \mid \theta)$, as a function of θ. The likelihood ratio $L(\xi \mid \theta)$ was introduced in the beginning, when θ stood for any (multivariate) parameter of the distribution. Below, θ will denote a real component of this parameter; we hope to avoid confusion with our earlier notation.

According to (2.1), we have

$$\mathbf{E}\varphi(\xi) = \mathbf{E}_0\varphi(\xi)L(\xi \mid \theta) = a(\theta),$$
$$\mathbf{E}1 = \mathbf{E}_0 L(\xi \mid \theta) = 1.$$

Suppose,

$$\begin{aligned} \frac{\partial}{\partial\theta}a(\theta) &= \frac{\partial}{\partial\theta}\mathbf{E}_0\varphi(\xi)L(\xi \mid \theta) = \mathbf{E}_0\varphi(\xi)\frac{\partial}{\partial\theta}L(\xi \mid \theta), \\ \frac{\partial}{\partial\theta}1 &= \frac{\partial}{\partial\theta}\mathbf{E}_0 L(\xi \mid \theta) = \mathbf{E}_0\frac{\partial}{\partial\theta}L(\xi \mid \theta) = 0. \end{aligned} \tag{2.8}$$

Then

$$\mathbf{E}_0[\varphi(\xi) - a(\theta)]\frac{\partial}{\partial\theta}L(\xi \mid \theta) = \frac{\partial}{\partial\theta}a(\theta).$$

Suppose,

$$\frac{\partial}{\partial\theta}\log L(x \mid \theta) = \frac{1}{L(x \mid \theta)}\frac{\partial}{\partial\theta}L(x \mid \theta)$$

satisfies

$$I(\theta) = \mathbf{E}\left[\frac{\partial}{\partial\theta}\log L(\xi \mid \theta)\right]^2 < \infty. \tag{2.9}$$

The quantity $I(\theta)$ is called the *Fisher information* on the parameter θ.

Consider the random variables

$$\eta_1 = [\varphi(\xi) - a(\theta)]\sqrt{L(\xi \mid \theta)},$$
$$\eta_2 = \frac{\partial}{\partial\theta}\log L(\xi \mid \theta)\sqrt{L(\xi \mid \theta)},$$

with

$$\mathbf{E}_0\eta_1\eta_2 = \frac{\partial}{\partial\theta}a(\theta).$$

Using the inequality

$$|\mathbf{E}_0\eta_1\eta_2| \leqslant \left(\mathbf{E}_0\eta_1^2\right)^{1/2}\left(\mathbf{E}_0\eta_2^2\right)^{1/2},$$

we obtain

$$\mathbf{E}_0\eta_1^2 \cdot \mathbf{E}_0\eta_2^2 \geqslant \left[\frac{\partial}{\partial\theta}a(\theta)\right]^2,$$

where, according to our basic assumption (2.1),

$$\begin{aligned}\mathbf{E}_0\eta_1^2 &= \mathbf{E}_0[\varphi(\xi) - a(\theta)]^2\, L(\xi \mid \theta)\\ &= \mathbf{E}[\varphi(\xi) - a(\theta)]^2 = \mathbf{D}\varphi(\xi)\end{aligned}$$

and

$$\begin{aligned}\mathbf{E}_0\eta_2^2 &= \mathbf{E}_0\left[\frac{\partial}{\partial\theta}\log L(\xi \mid \theta)\right]^2 L(\xi \mid \theta)\\ &= \mathbf{E}\left[\frac{\partial}{\partial\theta}\log L(\xi \mid \theta)\right]^2 = I(\theta).\end{aligned}$$

Thus, we get the following inequality for the variance $\mathbf{D}\varphi(\xi)$:

$$\mathbf{D}\varphi(\xi) \geqslant \left[\frac{\partial}{\partial\theta}a(\theta)\right]^2 I(\theta)^{-1}. \tag{2.10}$$

(2.10) is known as the *Rao–Cramér inequality*. In particular, if φ is an *unbiased* estimate with

$$a(\theta) = \mathbf{E}\varphi(\xi) \equiv \theta, \qquad \frac{\partial}{\partial\theta}a(\theta) \equiv 1,$$

then (2.10) gives the lower bound for the mean square error:

$$\mathbf{E}|\varphi(\xi) - \theta|^2 \geqslant I(\theta)^{-1}. \tag{2.11}$$

Let us formulate this result as follows.

THEOREM. *Under the regularity conditions* (2.8), (2.9), *the variance of an estimate of the parameter* θ *satisfies the Rao–Cramér inequality* (2.10).

Consider the Fisher information $I(\theta)$ introduced in (2.9). From the regularity conditions (2.8) we have

$$\mathbf{E}\frac{\partial}{\partial\theta}\log L(\xi \mid \theta) = \mathbf{E}_0\frac{\partial}{\partial\theta}L(\xi \mid \theta) = 0,$$

so that

$$I(\theta) = \mathbf{D}\frac{\partial}{\partial\theta}\log L(\xi \mid \theta) \tag{2.12}$$

is the variance of $\frac{\partial}{\partial\theta}\log L(\xi \mid \theta)$. Note that, according to the definition of $L(x \mid \theta)$,

$$\frac{\partial}{\partial\theta}\log L(\xi \mid \theta) = \frac{\partial}{\partial\theta}\log \mathbf{P}(\xi \mid \theta) \tag{2.13}$$

for discrete probability distribution $\mathbf{P}(x \mid \theta)$, $x \in \mathbb{R}^n$, and

$$\frac{\partial}{\partial\theta}\log L(\xi \mid \theta) = \frac{\partial}{\partial\theta}\log p(\xi \mid \theta) \tag{2.14}$$

for probability density $p(x \mid \theta)$, $x \in \mathbb{R}^n$.

How does $I(\theta) = I_n(\theta)$ depend on n? Assuming that $\xi = (\xi_1, \ldots, \xi_n)$ consists of independent identically distributed random variables ξ_k, $k = 1, \ldots, n$, in both cases (2.13), (2.14) we have

$$I_n(\theta) = nI_1(\theta) \tag{2.15}$$

(why?).

Let us analyse the inequality (2.10)/(2.11). Obviously, both (2.10) and (2.11) become equalities if and only if

$$\frac{\partial}{\partial\theta}\log L(\xi \mid \theta) = C(\theta)[\varphi(\xi) - a(\theta)], \tag{2.16}$$

where $C(\theta)$ is a constant depending on the full parameter of the probability distribution, since (2.16) means linear dependence between the random variables η_1, η_2

defined in the proof of (2.10). The estimate φ which satisfies (2.16) and the equality in (2.10), has the *minimal* variance

$$\mathbf{D}\varphi(\xi) = \left[\frac{\partial}{\partial\theta} a(\theta)\right]^2 I(\theta)^{-1} \tag{2.17}$$

and, for *unbiased* $\varphi(\xi)$, the *minimal* mean square error

$$\mathbf{E}|\varphi(\xi) - \theta|^2 = I(\theta)^{-1}. \tag{2.18}$$

An estimate $\varphi(\xi)$ satisfying (2.18) is called *efficient.*

EXAMPLE (*Efficient estimation of the mean value*). Let $\xi = (\xi_1, \ldots, \xi_n)$ consist of independent normal variables ξ_k, with

$$\mathbf{E}\xi_k = a, \qquad \mathbf{D}\xi_k = \sigma^2 \quad (k = 1, \ldots, n).$$

The estimate

$$\varphi(\xi) = \frac{1}{n}\sum_{k=1}^{n} \xi_k$$

of the component $\theta = a$ of the full parameter (a, σ^2) is unbiased and efficient, thanks to representation (2.16) in the form

$$\frac{\partial}{\partial\theta} \log p(\xi \mid \theta) = \frac{n}{\sigma^2}[\varphi(\xi) - \theta],$$

with $\theta = a$ on the right hand side.

2.4. ASYMPTOTIC NORMALITY AND EFFICIENCY OF THE MAXIMUM LIKELIHOOD ESTIMATE

Recall that the maximum likelihood estimate $\hat{\theta}$ of the unknown parameter $\theta \in \Theta$ of the probability distribution of $\xi = (\xi_1, \ldots, \xi_n)$ was defined as the maximum point of the corresponding likelihood ratio $L(x \mid \theta)$ (see p. 147). We are going to study some properties of such estimates $\hat{\theta}$ of a scalar (real) parameter θ, obtained by solving the equation

$$\frac{\partial}{\partial\theta} \log L(\xi \mid \theta) = 0, \tag{2.19}$$

in the case when $\xi = (\xi_1, \ldots, \xi_n)$ consists of independent identically distributed ξ_k, $k = 1, \ldots, n$. Here,

$$\frac{\partial}{\partial\theta} \log L(x \mid \theta) = \sum_{k=1}^{n} \frac{\partial}{\partial\theta} \log \mathbf{P}(x_k \mid \theta)$$

if the ξ_k's are discrete, and

$$\frac{\partial}{\partial\theta} \log L(x \mid \theta) = \sum_{k=1}^{n} \frac{\partial}{\partial\theta} \log p(x_k \mid \theta)$$

if the ξ_k's have a density $p(\cdot \mid \theta)$.

In any case,

$$\frac{\partial}{\partial\theta} \log L(\xi \mid \theta) = \sum_{k=1}^{n} \frac{\partial}{\partial\theta} \log L_1(\xi_k \mid \theta) \tag{2.20}$$

is the sum of independent identically distributed random variables

$$\frac{\partial}{\partial\theta} \log L_1(\xi_k \mid \theta), \quad k = 1, \ldots, n,$$

where $L(\xi \mid \theta) = L_n(\xi \mid \theta)$, $n = 1, 2, \ldots$. We impose the following regularity conditions for $n = 1$:

$$\mathbf{E}_0 \frac{\partial}{\partial\theta} L_1(\xi \mid \theta) = \frac{\partial}{\partial\theta} \mathbf{E}_0 L_1(\xi \mid \theta) = \frac{\partial}{\partial\theta} 1 = 0,$$

$$\mathbf{E}_0 \frac{\partial^2}{\partial\theta^2} L_1(\xi \mid \theta) = \frac{\partial^2}{\partial\theta^2} \mathbf{E}_0 L_1(\xi \mid \theta) = 0$$

and

$$\left| \frac{\partial^3}{\partial\theta^3} \log L_1(\xi \mid \theta) \right| \leqslant \varphi(\xi), \tag{2.21}$$

where $\varphi(\xi) \geqslant 0$, $\mathbf{E}\varphi(\xi) < \infty$.

In particular, the above assumptions imply

$$\mathbf{E}\frac{\partial}{\partial\theta}\log L_1(\xi \mid \theta) = \mathbf{E}_0\frac{\partial}{\partial\theta}L_1(\xi \mid \theta) = 0,$$

$$\mathbf{E}\frac{\partial^2}{\partial\theta^2}\log L_1(\xi \mid \theta) \tag{2.22}$$

$$= \mathbf{E}_0\frac{\partial^2}{\partial\theta^2}L_1(\xi \mid \theta) - \mathbf{E}\left[\frac{\partial}{\partial\theta}\log L_1(\xi \mid \theta)\right]^2 = -I_1(\theta),$$

where $I_1(\theta)$ is the Fisher information corresponding to $n = 1$ (see (2.8), (2.9), (2.15)); we assume also that

$$I_1(\theta) = \mathbf{E}\left[\frac{\partial}{\partial\theta}\log L_1(\xi \mid \theta)\right]^2 \neq 0.$$

With all these assumptions made, the following result holds true.

THEOREM. *For sufficiently large* n $(n \to \infty)$, *with probability* 1 *there is a solution* $\theta = \hat{\theta}$ *of the likelihood equation* (2.19), *which gives a consistent estimate* $\hat{\theta}$ *of the parameter* θ:

$$\hat{\theta} \to \theta. \tag{2.23}$$

Moreover,

$$\lim_{n\to\infty} \mathbf{P}\{t_1 \leqslant (\hat{\theta} - \theta)\sqrt{I_n(\theta)} \leqslant t_2\}$$

$$= \frac{1}{\sqrt{2\pi}}\int_{t_1}^{t_2} \mathrm{e}^{(-t^2)/2}\,\mathrm{d}t, \quad -\infty \leqslant t_1 < t_2 \leqslant \infty. \tag{2.24}$$

Let us note at once that the asymptotic normality (2.24) implies the so-called *asymptotic efficiency*. Namely, $(\hat{\theta} - \theta)\sqrt{I_n(\theta)}$ *asymptotically* has mean zero and variance 1, or $\hat{\theta}$ asymptotically is unbiased and has variance $I_n(\theta)^{-1}$, similarly as if it were an *efficient estimate* (see p. 166).

Proof of the theorem. Suppose, the true value of the parameter is $\theta = \theta_*$. Using the regularity conditions on the likelihood ratio, we obtain

$$\frac{\partial}{\partial\theta} \log L_1(\xi_k \mid \theta)$$
$$= \frac{\partial}{\partial\theta} \log L_1(\xi_k \mid \theta_*) + (\theta - \theta_*)\frac{\partial^2}{\partial\theta^2} \log L_1(\xi_k \mid \theta_*) +$$
$$+ \frac{1}{2}(\theta - \theta_*)^2\, \delta\varphi(\xi_k)$$

with some $\delta = \delta(\xi_k, \theta)$, $|\delta| \leqslant 1$. Together with (2.20), this implies

$$\frac{1}{n}\frac{\partial}{\partial\theta} \log L(\xi \mid \theta) = \lambda_{0n} + (\theta - \theta_*)\lambda_{1n} + \frac{1}{2}(\theta - \theta_*)^2\delta\lambda_{2n}$$

with some δ, $|\delta| \leqslant \max_k |\delta(\xi_k, \theta)| \leqslant 1$, and

$$\lambda_{0n} = \frac{1}{n}\sum_{k=1}^{n} \frac{\partial}{\partial\theta} \log L_1(\xi_k \mid \theta_*) \to \mathbf{E}\frac{\partial}{\partial\theta} \log L_1(\xi_k \mid \theta_*) = 0,$$

$$\lambda_{1n} = \frac{1}{n}\sum_{k=1}^{n} \frac{\partial^2}{\partial\theta^2} \log L_1(\xi_k \mid \theta_*) \to \mathbf{E}\frac{\partial^2}{\partial\theta^2} \log L_1(\xi_k \mid \theta_*) = -I_1(\theta_*) < 0,$$

$$\lambda_{2n} = \frac{1}{n}\sum_{k=1}^{n} \varphi(\xi_k) \to C = \mathbf{E}\varphi(\xi_k) \quad (n \to \infty),$$

thanks to (2.21), (2.22) and the law of large numbers. Now, it is easy to see that, for any arbitrary small $\varepsilon > 0$, with probability 1 there is a point $\theta = \hat{\theta}$ which satisfies the inequalities

$$\theta_* - \varepsilon \leqslant \hat{\theta} \leqslant \theta_* + \varepsilon,$$

and the likelihood equation (2.19), since the continuous function $\frac{\partial}{\partial\theta} \log L(\xi \mid \theta)$ of θ changes sign at the end points of the interval $\theta_* - \varepsilon \leqslant \theta \leqslant \theta_* + \varepsilon$, for sufficiently large n ($n \to \infty$). Obviously, this proves the consistency property (2.23) of the estimate $\hat{\theta}$. Equation (2.19) for $\theta = \hat{\theta}$ gives us

$$\lambda_{0n} + (\hat{\theta} - \theta_*)\lambda_{1n} + \frac{1}{2}(\hat{\theta} - \theta_*)^2\delta\lambda_{2n} = 0;$$

hence

$$(\hat{\theta} - \theta_*)\sqrt{nI_1(\theta_*)} = \frac{\sqrt{n}\frac{\lambda_{0n}}{\sqrt{I_1(\theta_*)}}}{-\frac{\lambda_{1n}}{I_1(\theta_*)} - \frac{1}{2}(\hat{\theta} - \theta_*)\delta\lambda_{2n}},$$

where

$$\sqrt{n}\,\frac{\lambda_{0n}}{\sqrt{I_1(\theta_*)}} = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{\frac{\partial}{\partial\theta}\log L_1(\xi_k \mid \theta_*)}{\sqrt{I_1(\theta_*)}}$$

obeys the central limit theorem, and

$$-\frac{\lambda_{1n}}{I_1(\theta_*)} - \frac{1}{2}(\hat{\theta} - \theta_*)\delta\lambda_{2n} \longrightarrow 1,$$

since $\lambda_{1n} \to -I_1(\theta_*)$, $\hat{\theta} \to \theta_*$ $(n \to \infty)$ with probability 1. This proves the asymptotic normality (2.24) of the estimate $\hat{\theta}$.

CHAPTER 4

Basic Elements of Probability Theory

1. General Probability Distributions

1.1. MAPPINGS AND σ-ALGEBRAS

Suppose, we are given a probability model $(\Omega, \mathfrak{A}, \mathbf{P})$, with all possible outcomes $\omega \in \Omega$ as elementary events, a σ-algebra $\mathfrak{A}$ of events $A \subseteq \Omega$, and probabilities $\mathbf{P}(A)$, $A \in \mathfrak{A}$.

In other words, $\mathbf{P}(A)$, $A \in \mathfrak{A}$, is a (probability) *measure* on Ω, satisfying the *σ-additivity* property.

Suppose,

$$\Omega \ni \omega \xrightarrow{\xi} x \in X$$

is a mapping from Ω to a set X, and

$$X \supseteq B \xrightarrow{\xi^{-1}} A \subseteq \Omega$$

is the corresponding set inverse defined by

$$\xi^{-1}B = \{\xi \in B\} = \{\omega : \xi(\omega) \in B\}.$$

The inverse mapping ξ^{-1} preserves relationships between sets $B \subseteq X$, such as

$$\xi^{-1}(B^c) = (\xi^{-1}B)^c,$$

$$\xi^{-1}\Big(\bigcup_k B_k\Big) = \bigcup_k (\xi^{-1}B_k),$$

$$\xi^{-1}\Big(\bigcap_k B_k\Big) = \bigcap_k (\xi^{-1}B_k),$$

etc. This simple observation shows at once that the family of all sets $B \subseteq X$, with the property

$$\xi^{-1}B = \{\xi \in B\} \in \mathfrak{A}, \tag{1.1}$$

forms a *σ-algebra* which we denote by $\mathfrak{B}_\xi$. With $\omega \in \Omega$ representing a *random* outcome, we treat $\xi = \xi(\omega)$, $\omega \in \Omega$, as a random element in X which generates the family $\mathfrak{B}_\xi$ of events, with the probabilities

$$\mathbf{P}_\xi(B) = \mathbf{P}\{\xi \in B\}, \quad B \in \mathfrak{B}_\xi. \tag{1.2}$$

Here, we have the *probability distribution* of ξ with the 'phase space' X, given by the *probability measure* $\mathbf{P}_\xi = \mathbf{P}_\xi(B)$, $B \in \mathfrak{B}_\xi$ of (1.2) on X. The triplet $(X, \mathfrak{B}_\xi, \mathbf{P}_\xi)$ serves as the probability model associated with the random element ξ, when we are interested in ξ alone and consider any event $\{\xi = x\}$, $x \in X$, as a possible outcome (*elementary event*).

Suppose, we are interested in the *events* of the form

$$A = \{\xi \in B\} \in \mathfrak{A}, \quad B \in \mathfrak{B}_0, \tag{1.3}$$

where $\mathfrak{B}_0$ is a family of sets $B \subseteq X$. More general sets $B \subseteq X$ and $\{\xi \in B\} \in \mathfrak{A}$ may appear as a result of various combinations of the initial ones (and their limits). More precisely, one has to consider the whole *σ-algebra* $\mathfrak{B}$ *generated* by $B \in \mathfrak{B}_0$, and the corresponding σ-algebra

$$\mathfrak{A}_\xi = \xi^{-1}\mathfrak{B}$$

of all $\xi^{-1}B = \{\xi \in B\}$, $B \in \mathfrak{B}$. Formally, $\mathfrak{B}$ can be defined as the *minimal* σ-algebra containing $\mathfrak{B}_0$. According to the definition of $\mathfrak{B}$, we have $\mathfrak{B} \subseteq \mathfrak{B}_\xi$ since the σ-algebra $\mathfrak{B}_\xi$ contains the sets in (1.3). Thus, all events

$$A = \{\xi \in B\}, \quad B \in \mathfrak{B}, \tag{1.4}$$

form the *σ-algebra* $\mathfrak{A}_\xi \subseteq \mathfrak{A}$, *generated* by the initial events (1.3), and we can consider the corresponding *probability measure*

$$\mathbf{P}_\xi(B) = \mathbf{P}\{\xi \in B\}, \quad B \in \mathfrak{B}, \tag{1.5}$$

representing the *probability distribution* of ξ on the σ-algebra $\mathfrak{B} \subseteq \mathfrak{B}_\xi$.

The triplet $(X, \mathfrak{B}, \mathbf{P}_\xi)$ can also serve as the probability model associated with ξ.

EXAMPLE (*Random variables*). In the general probability model $(\Omega, \mathfrak{A}, \mathbf{P})$, we actually defined a random variable $\xi = \xi(\omega)$, $\omega \in \Omega$ as a mapping from Ω to $X = \mathbb{R}$ such that the events of the type (1.3) are well-defined for all $B \in \mathfrak{B}_0$, where $\mathfrak{B}_0$ is the family of all finite unions of *disjoint* intervals $(x', x'']$ (see p. 44). Note that this family $\mathfrak{B}_0$, consisting of finite unions of all *disjoint* intervals $(x', x'']$, $-\infty \leqslant x' < x'' \leqslant \infty$ is an *algebra* on $\mathbb{R}$, which generates *Borel sets* $B \subseteq \mathbb{R}$, forming the *minimal* σ-algebra $\mathfrak{B} \supseteq \mathfrak{B}_0$. Thus, a random variable ξ determines the σ-algebra $\mathfrak{A}_\xi \subseteq \mathfrak{A}$ (1.4) of events and the *probability distribution* $\mathbf{P}_\xi = \mathbf{P}_\xi(B)$, $B \in \mathfrak{B}$, (1.5) on *Borel sets* $B \subseteq \mathbb{R}$.

EXAMPLE (*Joint probability distributions*). Considering several random variables $\xi_1, \ldots, \xi_n$, we deal with the mapping

$$\Omega \ni \omega \to \xi(\omega) = \{\xi_1(\omega), \ldots, \xi_n(\omega)\} \in X = \mathbb{R}^n,$$

such that events of the type (1.3) are well-defined for all $B \in \mathfrak{B}_0$, where $\mathfrak{B}_0$ is the family of all finite unions of *disjoint* 'rectangles'

$$(x_1', x_1''] \times \cdots \times (x_n', x_n''], \quad -\infty \leqslant x_k' < x_k'' < \infty, \ k = 1, \ldots, n,$$

including the whole space $\mathbb{R}^n$ (see p. 35). $\mathfrak{B}_0$ is an *algebra* which generates all *Borel sets* $B \subseteq \mathbb{R}^n$ forming the *minimal* σ-algebra $\mathfrak{B} \supseteq \mathfrak{B}_0$. Thus, random variables $\xi_1, \ldots, \xi_n$ determine the corresponding *joint probability distribution* $\mathbf{P}_\xi = \mathbf{P}_\xi(B)$, $B \in \mathfrak{B}$, such that, for $B = B_1 \times \cdots \times B_n$ with Borel sets $B_k \subseteq \mathbb{R}$, $k = 1, \ldots, n$,

$$\mathbf{P}_{\xi_1, \ldots, \xi_n}(B_1 \times \ldots \times B_n) = \mathbf{P}\{\xi_1 \in B_1, \ldots, \xi_n \in B_n\}; \quad B_1, \ldots, B_n \subseteq \mathbb{R}. \tag{1.6}$$

The triplet $(X, \mathfrak{B}, \mathbf{P}_\xi)$, $X = \mathbb{R}^n$, can serve as the probability space for the random variables $\xi_1, \ldots, \xi_n$, when we are interested in the probabilities of $\xi = (\xi_1, \ldots, \xi_n)$ alone. A random outcome

$$\{\xi_1(\omega) = x_1, \ldots, \xi_n(\omega) = x_n\}$$

can be identified with $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. The random variables $\xi_1, \ldots, \xi_n$ themselves can be defined, as functions of elementary event $x \in \mathbb{R}^n$ in the probability model $(\mathbb{R}^n, \mathfrak{B}, \mathbf{P}_\xi)$, by

$$\xi_1(x) = x_1, \ldots, \xi_n(x) = x_n, \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

EXAMPLE (*Probability distributions in functional spaces*). Suppose, we deal with a family ξ_t, $t \in T$, of random variables indexed by an arbitrary set T. For any possible outcome $\omega \in \Omega$,

$$\xi(\omega) = \{\xi_t(\omega),\ t \in T\}$$

represents a corresponding *trajectory*, which formally can be defined as a function of $t \in T$. Let us introduce the space $X = \mathbb{R}^T$ of *all* functions

$$x = \{x_t,\ t \in T\}$$

with values $x_t \in \mathbb{R}$. Consider the so-called *cylinder sets* in X of the form

$$\{x\colon (x_{t_1}, \dots, x_{t_n}) \in B^{(n)}\}, \quad B^{(n)} \subseteq \mathbb{R}^n, \tag{1.7}$$

for any finite collection $t_1, \dots, t_n \in T$, and any Borel set $B^{(n)} \subseteq \mathbb{R}^n$ (the cylinder set (1.7) will be denoted $B^{(n)} \subseteq X$, by the same symbol as the corresponding set $B^{(n)} \subseteq \mathbb{R}^n$). Obviously, the family $\mathfrak{B}_0$ of all such cylinder sets is an algebra, which generates a *σ-algebra* $\mathfrak{B}$ in X.

The mapping

$$\Omega \ni \omega \xrightarrow{\xi} x \in X = \mathbb{R}^T,$$

with $x = \xi(\omega)$, is such that

$$A = \xi^{-1}B^{(n)} = \{\omega\colon (\xi_{t_1}, \dots, \xi_{t_n}) \in B^{(n)}\}$$

is an *event* for any cylinder set $B^{(n)} \subseteq X$. It determines the *probability* measure

$$\mathbf{P}_\xi = \mathbf{P}_\xi(B), \quad B \in \mathfrak{B},$$

on the σ-algebra $\mathfrak{B}$ in the functional space $X = \mathbb{R}^T$, which is given on the algebra $\mathfrak{B}_0$ of all cylinder sets $B \subseteq X$ by

$$\mathbf{P}_\xi(B) = \mathbf{P}(A), \quad A = \xi^{-1}B.$$

Dealing with random variables ξ_t, $t \in T$, characterized by their joint probability distributions alone, we can apply $(X, \mathfrak{B}, \mathbf{P}_\xi)$, $X = \mathbb{R}^T$, as the corresponding probability model, with

$$\xi_t = \xi_t(x) \equiv x_t, \quad t \in T,$$

being functions of elementary event $x \in X = \mathbb{R}^T$, $x = \{x_t,\ t \in T\}$. Of course, in the framework of the probability model $(\mathbb{R}^T, \mathfrak{B}, \mathbf{P}_\xi)$, the *joint probability distributions*

$$P_{t_1,\dots,t_n}(B^{(n)}) = \mathbf{P}_{\xi_{t_1},\dots,\xi_{t_n}}(B^{(n)}), \quad B^{(n)} \subseteq \mathbb{R}^n,$$

altogether determine the probability measure $\mathbf{P}_\xi$ on X; in particular, for any cylinder set $B \subseteq \mathbb{R}^T$,

$$\mathbf{P}_\xi(B) = \mathbf{P}_{t_1,\dots,t_n}(B^{(n)}), \tag{1.8}$$

where $B^{(n)}$ is the corresponding set in $\mathbb{R}^n$. Observe that the $\mathbf{P}_{t_1,\dots,t_n}$'s are consistent in the following sense: $\mathbf{P}_{t_1,\dots,t_n}(B_1 \times \cdots \times B_n)$ is invariant with respect to a simultaneous permutation of $t_1, \dots, t_n$ and $B_1, \dots, B_n$, moreover,

$$\mathbf{P}_{t_1,\dots,t_{n+1}}(B_1 \times \cdots \times B_n \times \mathbb{R}) = \mathbf{P}_{t_1,\dots,t_n}(B_1 \times \cdots \times B_n).$$

The above properties define a *consistent family of finite dimensional distributions* $\mathbf{P}_{t_1,\dots,t_n}$; $t_1, \dots, t_n \in T$, $n = 1, 2, \dots$.

1.2. APPROXIMATION OF EVENTS

In the framework of the general probability model $(\Omega, \mathfrak{A}, \mathbf{P})$, any event $A \subseteq \Omega$ can be described by means of its indicator

$$1_A = 1_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

We recall that A is equivalent to B if they coincide almost surely, i.e., if the event

$$\{\omega\colon 1_A(\omega) \neq 1_B(\omega)\} = AB^c + BA^c \equiv A \triangle B$$

has *zero* probability; here, $A \triangle B$ is the *symmetric difference* of the events A, B.

The probability

$$\mathbf{P}(A\triangle B) = \mathbf{E}|1_A - 1_B| = \| 1_A - 1_B \| \tag{1.9}$$

defines the *distance* between events A and B which coincides with the *mean distance*

$$\| 1_A - 1_B \| = \mathbf{E}|1_A - 1_B|$$

between their indicators; obviously

$$|\mathbf{P}(A) - \mathbf{P}(B)| = |\mathbf{E}\,1_A - \mathbf{E}\,1_B| \leqslant \mathbf{E}|1_A - 1_B| = \mathbf{P}(A\triangle B). \tag{1.10}$$

LEMMA (Approximation of events). *Let $\mathfrak{B} \subseteq \mathfrak{A}$ be the σ-algebra generated by an algebra $\mathfrak{B}_0$ of events. Then $\mathfrak{B}$ belongs to the closure of $\mathfrak{B}_0$ with respect to the distance* (1.9), *i.e. for any $B \in \mathfrak{B}$ and $\varepsilon > 0$ there is $B_\varepsilon \in \mathfrak{B}_0$ such that*

$$\mathbf{P}(B\triangle B_\varepsilon) \leqslant \varepsilon. \tag{1.11}$$

Proof. Let $[\mathfrak{B}_0] \subseteq \mathfrak{A}$ be the closure of $\mathfrak{B}_0$; we show that it is a σ-algebra. For any $B \in [\mathfrak{B}_0]$, the complement $B^c \in [\mathfrak{B}_0]$, since for a corresponding $B_\epsilon \in \mathfrak{B}_0$ and its complement $B_\epsilon^c \in \mathfrak{B}_0$ we have

$$B^c\triangle B_\epsilon^c = B^c B_\epsilon \bigcup B B_\epsilon^c = B\triangle B_\epsilon,$$

$$\mathbf{P}(B^c\triangle B_\epsilon^c) = \mathbf{P}(B\triangle B_\epsilon) \leqslant \varepsilon.$$

Consider any $B_1, B_2 \in [\mathfrak{B}_0]$ and their product B_1, B_2, with the indicator $1_{B_1 B_2} = 1_{B_1} \cdot 1_{B_2}$. Then

$$\begin{aligned}
&\| 1_{B_1} \cdot 1_{B_2} - 1_{B_{\epsilon 1}} \cdot 1_{B_{\epsilon 2}} \| \\
&\quad \leqslant \| 1_{B_1} \cdot 1_{B_2} - 1_{B_{\epsilon 1}} \cdot 1_{B_2} \| + \| 1_{B_{\epsilon 1}} \cdot 1_{B_2} - 1_{B_{\epsilon 1}} \cdot 1_{B_{\epsilon 2}} \| \\
&\quad \leqslant \| 1_{B_1} - 1_{B_{\epsilon 1}} \| + \| 1_{B_2} - 1_{B_{\epsilon 2}} \| \leqslant 2\varepsilon,
\end{aligned}$$

where $B_{\epsilon 1}, B_{\epsilon 2} \in \mathfrak{B}_0$ are corresponding approximations; hence have $B_1 \cdot B_2 \in [\mathfrak{B}_0]$. Thus, $[\mathfrak{B}_0]$ is an *algebra*. Consider an increasing sequence $B_n \in [\mathfrak{B}_0]$, $n = 1, 2, \ldots$, and its limit

$$B = \lim_{n\to\infty} B_n \quad \left(= \bigcup_n B_n\right).$$

According to

$$\| 1_B - 1_{B_n} \| = \mathbf{P}(B) - \mathbf{P}(B_n) \to 0,$$

for proper approximations $B_{\epsilon n} \in \mathfrak{B}_0$ we have

$$\| 1_B - 1_{B_{\epsilon n}} \| \leqslant \| 1_B - 1_{B_n} \| + \| 1_{B_n} - 1_{B_{\epsilon n}} \| \leqslant \varepsilon$$

provided n is sufficiently large. Thus, $B \in [\mathfrak{B}_0]$ and $[\mathfrak{B}_0]$ is a *σ-algebra*. Hence, $[\mathfrak{B}_0] \supseteq \mathfrak{B}$ by the definition of the *minimal* σ-algebra $\mathfrak{B} \supseteq \mathfrak{B}_0$, since $[\mathfrak{B}_0]$ contains $\mathfrak{B}_0$. Thus, any $B \in \mathfrak{B}$ can be approximated by a corresponding, $B_\varepsilon \in \mathfrak{B}_0$ as stated in (1.11).

Of course, according to (1.10), (1.11), all probabilities $\mathbf{P}(B)$, $B \in \mathfrak{B}$, can be obtained as corresponding limits $\lim \mathbf{P}(B)$, $B \in \mathfrak{B}_0$.

EXAMPLE (*Probability distributions*). At the very beginning, we actually introduced the probability distribution,

$$\mathbf{P}_\xi(B) = \mathbf{P}\{\xi \in B\}, \quad B \subseteq \mathfrak{B}_0,$$

of a random variable ξ on the algebra $\mathfrak{B}_0$ of finite unions of disjoint intervals $(x', x'']$, $-\infty < x' < x'' < \infty$ (see p. 34). According to (1.10), (1.11), it uniquely determines the probability measure

$$\mathbf{P}_\xi(B) = \mathbf{P}\{\xi \in B\}, \quad B \in \mathfrak{B},$$

on the σ-algebra $\mathfrak{B}$ of Borel sets $B \subseteq \mathbb{R}$. In a similar way, for several random variables $\xi_1, \ldots, \xi_n$, we introduced their joint probability distribution

$$\mathbf{P}_{\xi_1,\ldots,\xi_n}(B_1 \times \cdots \times B_n) = \mathbf{P}\{\xi_1 \in B_1, \ldots, \xi_n \in B_n\}, \quad B_1, \ldots, B_n \subseteq \mathbb{R},$$

on the algebra $\mathfrak{B}_0$ of finite unions of all *disjoint* rectangles $(x_1', x_1''] \times \cdots \times (x_n', x_n'']$, $-\infty < x_k' < x_k'' < \infty$, $k = 1, \ldots, n$ (see p. 36); according to (1.10), (1.11) it uniquely determines the probability measure

$$\mathbf{P}_\xi(B) = \mathbf{P}\{\xi \in B\}, \quad B \in \mathfrak{B},$$

on the σ-algebra $\mathfrak{B}$ of Borel sets $B \subseteq \mathbb{R}^n$.

EXAMPLE (*Independent random variables*). Earlier, we characterized independent random variables $\xi_1, \dots, \xi_n$ by the equality

$$\mathbf{P}_{\xi_1,\dots,\xi_n}(B_1 \times \cdots \times B_n) = \mathbf{P}_{\xi_1}(B_1) \cdots \mathbf{P}_{\xi_n}(B_n) \tag{1.12}$$

for all intervals $B_k = (x_k', x_k''] \subseteq \mathbb{R}$, $k = 1, \dots, n$. The equality (1.12) can be immediately extended to all finite unions of *disjoint* intervals $(x_k', x_k'']$, $-\infty \leqslant x_k' < x_k'' \leqslant \infty$, and then, according to (1.10), (1.11), to all Borel sets $B_k \subseteq \mathbb{R}$, $k = 1, \dots, n$. This shows that *independent random variables* $\xi_1, \dots, \xi_n$ can be characterized in such a way that, for any Borel sets $B_k \subseteq \mathbb{R}$, the events $\{\xi_k \in B_k\}$, $k = 1, \dots, n$, are independent. (1.12) defines the joint probability distribution $\mathbf{P}_\xi(B)$, $B \subseteq \mathfrak{B}$, on Borel sets $B \subseteq \mathbb{R}^n$, as the product of the marginal probability distributions of ξ_k, $k = 1, \dots, n$. □

In further development of the notion of independence, we introduce the following formal definition. A random variable ξ is said *independent* of an *algebra (σ-algebra)* $\mathfrak{B}$ of events if ξ is independent of all indicators 1_B of events $B \in \mathfrak{B}$.

As we know, for independent events A_k, $k = 1, \dots, n$, any A_k is independent of the algebra generated by A_j, $j \neq k$ (see p. 23). In a similar way, if random variables ξ_k, $k = 1, \dots, n$, are independent, any ξ_k is independent of the *σ-algebra*, generated by ξ_j, $j \neq k$ (why?).

1.3. 0-1 LAW

Let us call algebras (σ-algebras) $\mathfrak{A}_1$ and $\mathfrak{A}_2$ independent if any events $A_1 \in \mathfrak{A}_1$ and $A_2 \in \mathfrak{A}_2$ are independent. For example, independent $\mathfrak{A}_1$, $\mathfrak{A}_2$ can be generated by given independent events, independent random variables, etc.

Suppose, we deal with an infinite sequence $\mathfrak{A}(k)$, $k = 1, 2, \dots$, of *independent* σ-algebras. Consider an event A whose occurrence is completely determined by the 'tail' $\mathfrak{A}(k)$, $k \geqslant n \to \infty$; more precisely, A is an event from the σ-algebra

$$\mathfrak{A}^\infty = \lim_{n\to\infty} \mathfrak{A}^n \quad \left(= \bigcap_n \mathfrak{A}^n\right),$$

where $\mathfrak{A}^n$ is the σ-algebra generated by $\mathfrak{A}(k)$, $k \geqslant n$.

THEOREM. *For any event* $A \in \mathfrak{A}^\infty$,

$$\mathbf{P}(A) = 0 \text{ or } 1. \tag{1.13}$$

Proof. Let us introduce the σ-algebra$\mathfrak{A}_n$ generated by all events from $\mathfrak{A}(k)$, $k \leqslant n$; the union

$$\bigcup_n \mathfrak{A}_n$$

is an *algebra* which generates the σ-algebra $\mathfrak{A} \supseteq \mathfrak{A}^{\infty}$. According to the lemma on approximation of events, any $A \in \mathfrak{A}$ can be approximated by a corresponding event A_ε from the union $\bigcup_n \mathfrak{A}_n$, $A_\varepsilon \in \mathfrak{A}_n$ for some $n = n(\varepsilon)$, such that

$$\lim_{\varepsilon \to 0} \mathbf{P}(A_\varepsilon) = \mathbf{P}(A), \qquad \lim_{\varepsilon \to 0} \mathbf{P}(AA_\varepsilon) = \mathbf{P}(A).$$

In particular, this is true for $A \in \mathfrak{A}^{\infty}$ which does not depend on $A_\varepsilon \in \mathfrak{A}_n$,

$$\mathbf{P}(AA_\varepsilon) = \mathbf{P}(A)\mathbf{P}(A_\varepsilon).$$

Consequently,

$$\mathbf{P}(A) = \lim_{\varepsilon \to 0} \mathbf{P}(AA_\varepsilon) = \mathbf{P}(A) \cdot \lim_{\varepsilon \to 0} \mathbf{P}(A_\varepsilon) = \mathbf{P}(A)^2,$$

which can be true only if $\mathbf{P}(A) = 0$ or 1.

The above theorem is called *Kolmogorov's* 0-1 *law*. (For example, the statements of the Borel–Cantelli lemmas, on the occurrence of independent events A_k, $k = 1, 2, \ldots$, can be interpreted in terms of the 0-1 law; see p. 22–23)

1.3. MATHEMATICAL EXPECTATION AS THE LEBESGUE INTEGRAL

We defined the mathematical expectation $\mathbf{E}\xi$ of a random variable $\xi \in \mathbb{R}$ as the *Lebesgue–Stieltjes integral*

$$\mathbf{E}\xi = \int_{-\infty}^{\infty} x \, \mathrm{d}\, F_\xi(x)$$

with respect to the corresponding distribution function

$$F_\xi(x) = \mathbf{P}\{\xi \leqslant x\} = \mathbf{P}_\xi(-\infty, x], \quad -\infty < x < \infty$$

(see p. 45). Using the probability distribution $\mathbf{P}_\xi = \mathbf{P}_\xi(B)$ as the *probability measure* on Borel sets $B \subseteq \mathbb{R}$, we can write

$$\mathbf{E}\xi = \int_{-\infty}^{\infty} x \mathbf{P}_\xi \, (\mathrm{d}\, x) \tag{1.14}$$

as the *Lebesgue integral*.

Given random variables $(\xi_1, \dots, \xi_n) = \xi \in \mathbb{R}^n$ we often deal with various functions $\varphi = \varphi(\xi)$. What conditions on a real function

$$\varphi = \varphi(x), \quad x = (x_1, \dots, x_n) \in \mathbb{R}^n,$$

guarantee that $\eta = \varphi(\xi)$ is a *random variable* so that

$$\{\eta \leqslant y\} \in \mathfrak{A}, \quad y \in \mathbb{R},$$

are well-defined events, say, in the framework of the general probability model $(\Omega, \mathfrak{A}, \mathbf{P})$? As a matter of fact, one need not worry about this question in the case when φ is a *Borel function*, i.e., such that

$$B = \{x\colon \varphi(x) \leqslant y\} \subseteq \mathbb{R}^n$$

is a *Borel set*, for any $y \in \mathbb{R}$. Indeed, in such a case,

$$\{\eta \leqslant y\} = \{\xi \in B\} \in \mathfrak{A}$$

as we already know. Using the joint probability distribution $\mathbf{P}_\xi = \mathbf{P}_\xi(B)$ of $\xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$ on Borel sets $B \subseteq \mathbb{R}^n$ as the *probability measure*, we can write

$$\mathbf{E}\varphi(\xi) = \int_{\mathbb{R}^n} \varphi(x)\, \mathbf{P}_\xi\,(\mathrm{d}\,x)$$

as the *Lebesgue integral*, or, in the coordinate form,

$$\mathbf{E}\varphi(\xi_1, \dots, \xi_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \varphi(x_1, \dots, x_n) \mathbf{P}_{\xi_1, \dots, \xi_n}\,(\mathrm{d}\,x_1 \dots \mathrm{d}\,x_n). \tag{1.15}$$

From (1.15) it follows that, for any *independent* $\xi_1, \dots, \xi_n$ with the joint probability distribution

$$\mathbf{P}_{\xi_1, \dots, \xi_n}\,(\mathrm{d}\,x_1 \dots \mathrm{d}\,x_n) = \mathbf{P}_{\xi_1}\,(\mathrm{d}\,x_1) \times \cdots \times \mathbf{P}_{\xi_n}\,(\mathrm{d}\,x_n),$$

the multiplicative formula

$$\mathbf{E}\big[\varphi_1(\xi_1) \cdots \varphi_n(\xi_n)\big] = \mathbf{E}\varphi_1(\xi_1) \cdots \mathbf{E}\varphi_n(\xi_n)$$

holds; here, for any Borel functions $\varphi_1, \dots, \varphi_n$, the random variables $\eta_1 = \varphi_1(\xi_1), \dots, \eta_n = \varphi_n(\xi_n)$ are actually independent.

As it was mentioned earlier, in the general probability model one has a *probability measure*

$$\mathbf{P}(A) = \int_A \mathbf{P}(\mathrm{d}\omega), \quad A \in \mathfrak{A},$$

on a σ-algebra $\mathfrak{A}$ in the space of elementary events $\omega \in \Omega$, and the *mathematical expectation* $\mathbf{E}\xi$ of a random variable

$$\xi = \xi(\omega), \quad \omega \in \Omega,$$

is given by the corresponding *Lebesgue integral*

$$\mathbf{E}\xi = \int_\Omega \xi(\omega)\,\mathbf{P}(\mathrm{d}\omega). \tag{1.16}$$

1.5. $\mathcal{L}_p$-SPACES

In the framework of the general probability model $(\Omega, \mathfrak{A}, \mathbf{P})$, the corresponding $\mathcal{L}_1$-space consists of all random variables ξ, $\mathbf{E}|\xi| < \infty$; it is a *linear* space equipped with the *mean norm*

$$\| \xi \| = \mathbf{E}|\xi|, \quad \xi \in \mathcal{L}_1.$$

We actually applied this norm when considering the convergence $\xi_n \to \xi$ in mean, which coincides with the convergence

$$\| \xi_n - \xi \| \to 0$$

in $\mathcal{L}_1$-space (see p. 55).

The corresponding *$\mathcal{L}_2$-space* is formed by all random variables ξ, $\mathbf{E}|\xi|^2 < \infty$; it is a *linear* space equipped with the *square mean norm*

$$\| \xi \| = \left(\mathbf{E}|\xi|^2\right)^{1/2}, \quad \xi \in \mathcal{L}_2,$$

which we applied earlier as well (see p. 63).

The following property is known for $\mathcal{L}_p$-spaces ($p = 1, 2$).

THEOREM. *$\mathcal{L}_p$-space is complete, i.e., the convergence*

$$\xi_n - \xi_m \to 0, \quad n, m \to \infty \tag{1.17}$$

(in the $\mathcal{L}_p$-space) implies

$$\xi_n \to \xi \in \mathcal{L}_p.$$

To prove the theorem, note that (1.17) implies the convergence in probability which in turn implies the same convergence $\xi_n \to \xi$, hence the existence of the limit random variable ξ. Indeed, for $\varepsilon = 1/2^k$ one can find n_k, $k = 1, 2, \dots$, such that

$$\mathbf{P}\left\{|\xi_{n_{k+1}} - \xi_{n_k}| > \frac{1}{2^k}\right\} \leqslant \frac{1}{2^k}.$$

According to the Borel–Cantelli lemma, for all sufficiently large $k = 1, 2, \dots$,

$$|\xi_{n_{k+1}} - \xi_{n_k}| \leqslant \frac{1}{2^k}.$$

Hence

$$\xi_{n_k}(\omega) \to \xi(\omega) = \xi_{n_1}(\omega) + \sum_{k=1}^{\infty} [\xi_{n_{k+1}}(\omega) - \xi_{n_k}(\omega)]$$

with probability 1. Now,

$$\mathbf{P}\{|\xi_n - \xi| > \varepsilon\} \leqslant \mathbf{P}\{|\xi_n - \xi_{n_k}| > \varepsilon/2\} + \mathbf{P}\{|\xi_{n_k} - \xi| > \varepsilon/2\} \to 0$$

for any $\varepsilon > 0$, i.e., $\xi_n \to \xi$ in probability.

By the discussion above,

$$\xi_{n_k} \to \xi, \qquad |\xi_n - \xi_{n_k}|^p \to |\xi_n - \xi|^p, \quad n_k \to \infty,$$

with probability 1. Moreover, according to (1.17), for any $\varepsilon > 0$

$$\mathbf{E}|\xi_n - \xi_{n_k}|^p \leqslant \varepsilon$$

provided n, n_k are sufficiently large. Now, by the well-known limit properties of mean values (see Chapter 1, (4.19), (4.19)′), we conclude that

$$\mathbf{E}|\xi_n - \xi|^p \leqslant \varepsilon.$$

This ends the proof of the theorem, as ξ_n, $\xi_n - \xi \in \mathcal{L}_p$ $(p = 1, 2)$ imply $\xi \in \mathcal{L}_p$. □

(Of course, $\mathcal{L}_p$-spaces and their completness are well-known in *Functional Analysis*. Namely, in the $(\Omega, \mathfrak{A}, \mathbf{P})$ model with the probability measure $\mathbf{P}(A) = \int_A \mathbf{P}(\mathrm{d}\,\omega)$, $A \in \mathfrak{A}$, the real $\mathcal{L}_p$-space is formed by all functions $\xi = \xi(\omega)$, $\omega \in \Omega$, which are measurable with respect to the σ-algebra $\mathfrak{A}$ ($\mathfrak{A}$-measurable, for short), with the corresponding norm $\| \xi \| < \infty$ given by means of the Lebesgue integral

$$\| \xi \|^p = \int_\Omega |\xi(\omega)|^p \mathbf{P}(\mathrm{d}\,\omega),$$

$p = 1, 2$. For any $\xi_n \in \mathcal{L}_p$, $n = 1, 2, \ldots$, satisfying

$$\int_\Omega \left|\xi_n(\omega) - \xi_m(\omega)\right|^p \mathbf{P}(\mathrm{d}\,\omega) \longrightarrow 0, \quad n, m \to \infty,$$

there is $\xi \in \mathcal{L}_p$, which represents the limit $\lim_{n\to\infty} \xi_n = \xi$ in $\mathcal{L}_p$-space:

$$\int_\Omega \left|\xi_n(\omega) - \xi(\omega)\right|^p \mathbf{P}(\mathrm{d}\omega)\mathrm{d}\,\omega \longrightarrow 0, \quad n \to \infty.)$$

In $\mathcal{L}_p$-spaces ($p = 1, 2$), we make no difference between any elements $\xi, \widetilde{\xi}$ with $\| \xi - \widetilde{\xi} \| = 0$, which means that the random variables $\xi, \widetilde{\xi}$ are *equivalent*, i.e., $\xi - \widetilde{\xi} = 0$ with probability 1.

We write

$$\mathcal{L}_p = \mathcal{L}_p(\Omega, \mathfrak{A}, \mathbf{P}), \quad p = 1, 2,$$

to emphasize the dependence on the triplet $(\Omega, \mathfrak{A}, \mathbf{P})$. Note that different σ-algebras $\mathfrak{A}$ correspond to different $\mathcal{L}_p$-spaces. Obviously,

$$\mathcal{L}_p(\Omega, \mathfrak{A}_1, \mathbf{P}) \subseteq \mathcal{L}_p(\Omega, \mathfrak{A}_2, \mathbf{P})$$

whenever $\mathfrak{A}_1 \subseteq \mathfrak{A}_2$. In particular, for any σ-algebra $\mathfrak{B} \subseteq \mathfrak{A}$ in the probability model $(\Omega, \mathfrak{A}, \mathbf{P})$,

$$\mathcal{L}_p(\Omega, \mathfrak{B}, \mathbf{P})$$

is a *subspace* of $\mathcal{L}_p = \mathcal{L}_p(\Omega, \mathfrak{A}, \mathbf{P})$.

Similarly to the fact that any $\xi \in \mathcal{L}_p(\Omega, \mathfrak{A}, \mathbf{P})$ generates events $\{\xi \in B\} \in \mathfrak{A}$, with Borel sets $B \subseteq \mathbb{R}$, we have that, in the probability model $(\Omega, \mathfrak{A}, \mathbf{P})$, any $\xi \in \mathcal{L}_p(\Omega, \mathfrak{B}, \mathbf{P})$ generates events

$$\{\xi \in B\} \in \mathfrak{B} \tag{1.18}$$

for Borel sets $B \subseteq \mathbb{R}$. In such a case, we say that ξ is *measurable* with respect to the σ-algebra $\mathfrak{B}$ ($\mathfrak{B}$-measurable).

Projection in Hilbert space. $\mathcal{L}_2$-space can be equipped with the *inner product*

$$(\xi, \eta) = \mathbf{E}\xi\eta; \quad \xi, \eta \in \mathcal{L}_2$$

(see also p. 63). This makes $\mathcal{L}_2$-space a *Hilbert space*, with the square mean norm

$$\| \xi \| = |(\xi, \xi)|^{1/2}, \quad \xi \in \mathcal{L}_2.$$

According to the well-known inequality

$$|(\xi, \eta)| \leqslant \| \xi \| \cdot \| \eta \|; \quad \xi, \eta \in \mathcal{L}_2,$$

which we already applied, see p. 63, we have

$$\mathbf{E}|\xi| \leqslant \left(\mathbf{E}|\xi|^2\right)^{1/2}.$$

Consequently,

$$\mathcal{L}_2 \subseteq \mathcal{L}_1; \tag{1.19}$$

moreover, $\mathcal{L}_1$ coincides with the closure of $\mathcal{L}_2$ in $\mathcal{L}_1$, as any $\xi \in \mathcal{L}_1$ is the limit in $\mathcal{L}_1$ of 'continuous' approximations $\varphi_a(\xi) \in \mathcal{L}_2$ (see Figure 17).

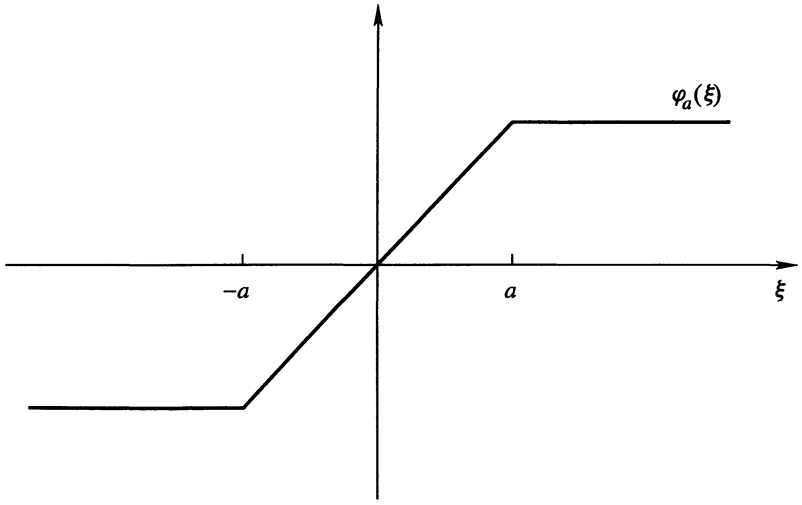

Fig. 17.

Of course, the above property holds for any σ-algebra $\mathfrak{B} \subseteq \mathfrak{A}$, i.e.,

$$\mathcal{L}_2 = \mathcal{L}_2(\Omega, \mathfrak{B}, \mathbf{P}) \subseteq \mathcal{L}_1(\Omega, \mathfrak{B}, \mathbf{P}) = \mathcal{L}_1.$$

□

We recall now the following remarkable property of the Hilbert space $H(= \mathcal{L}_2)$.

THEOREM. *For any element $\xi \in H$, one can define its projection $\hat{\xi} \in H_0$ on an arbitrary subspace $H_0 \subseteq H$, such that*

$$\| \xi - \hat{\xi} \| = \inf_{\varphi \in H_0} \| \xi - \varphi \| . \tag{1.20}$$

The difference $\xi - \hat{\xi}$ is orthogonal to H_0, and the condition

$$(\xi - \hat{\xi}, \varphi) = [\mathbf{E}(\xi - \hat{\xi})\varphi] = 0, \quad \varphi \in H_0, \tag{1.21}$$

uniquely characterizes the projection (see Figure 18).

This result is well-known in finite dimensions, and can be easily verified in the general case. Namely, consider a sequence φ_n, $n = 1, 2, \ldots$, leading to the infimum in (1.20), and the finite dimensional subspace H_m generated by $\varphi_1, \ldots, \varphi_m$ $(m = 1, 2, \ldots)$. The projection $\hat{\xi}_m$ of ξ on H_m is at the same time the projection of $\hat{\xi}_n$ on H_m for $n \geqslant m$, since $H_m \subseteq H_n$. We have

$$\| \hat{\xi}_n - \hat{\xi}_m \|^2 = \| \hat{\xi}_n \|^2 - \| \hat{\xi}_m \|^2 \longrightarrow 0, \quad m, n \to \infty,$$

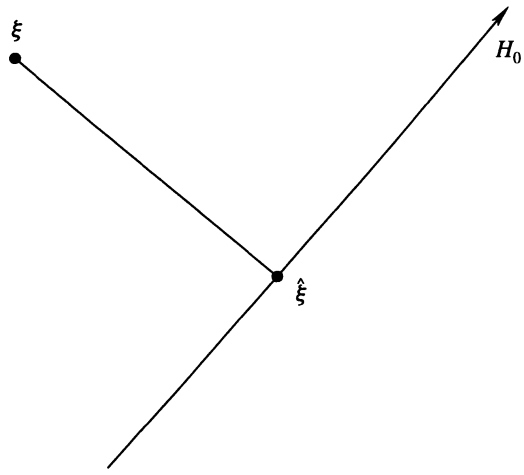

Fig. 18.

as $\| \hat{\xi}_n \|$ *increase* with n and are bounded:

$$\| \hat{\xi}_m \|^2 \leqslant \| \hat{\xi}_n \|^2 \leqslant \| \xi \|^2 \quad (n \geqslant m).$$

Hence, the limit

$$\lim_{n \to \infty} \hat{\xi}_n = \hat{\xi} \in H_0$$

exists, and

$$\begin{aligned} \| \xi - \hat{\xi} \| &= \lim_{n \to \infty} \| \xi - \hat{\xi}_n \| \\ &\leqslant \lim_{n \to \infty} \| \xi - \varphi_n \| = \inf_{\varphi \in H_0} \| \xi - \varphi \| . \end{aligned}$$

□

An application of the projection method is given in the following

EXAMPLE (*The best forecast problem*). Suppose, we are interested in a random variable ξ, $\mathbf{E}|\xi|^2 < \infty$, which is not *observable* (at present, say). We want to forecast ξ using available information, which is represented by a *random* element $\eta = y \in Y$ in a *measurable space* $(Y, \mathfrak{B})$. In other words, the space Y is equipped with a σ-algebra $\mathfrak{B}$ of sets $B \subseteq Y$, and, for any $B \in \mathfrak{B}$, the corresponding *event*

$$\{\eta \in B\} \in \mathfrak{A}.$$

The forecast will be given in the form $\varphi(\eta)$, where $\varphi = \varphi(y)$, $y \in Y$, is an arbitrary function satisfying $\varphi = \varphi(\eta) \in \mathcal{L}_2$. Let

$$\| \xi - \varphi \|^2 = \mathbf{E}|\xi - \varphi(\eta)|^2$$

be the corresponding *mean square error*. We can proceed in the following way. For a convenience of notation, we identify an *event* $\{\eta \in B\} \in \mathfrak{A}$ with the corresponding set $B \in \mathfrak{B}$, which lets us formally introduce the corresponding σ-algebra $\mathfrak{B} \subseteq \mathfrak{A}$ of *events* $B \in \mathfrak{B}$. Then, by a forecast $\varphi(\eta)$ we mean any

$$\varphi \in \mathcal{L}_2(\Omega, \mathfrak{B}, \mathbf{P}).$$

Obviously, the *best forecast* $\hat{\xi} = \varphi_0(\eta)$ is given by the *projection* of ξ in the space $H = \mathcal{L}_2(\Omega, \mathfrak{A}, \mathbf{P})$ onto the subspace $H_0 = \mathcal{L}_2(\Omega, \mathfrak{B}, \mathbf{P})$, since it yields the *minimal mean square error*:

$$\| \xi - \hat{\xi} \| = \min_{\varphi \in H_0} \| \xi - \varphi \|,$$

see (1.20). Moreover, $\hat{\xi} = \varphi_0$ can be identified by means of the orthogonality condition (1.21), or the condition

$$\mathbf{E}\hat{\xi}\, 1_B = \mathbf{E}\xi\, 1_B, \quad B \in \mathfrak{B},$$

since the indicators 1_B of *events* $B \in \mathfrak{B}$ form a *complete system* in the subspace $H_0 = \mathcal{L}_2(\Omega, \mathfrak{B}, \mathbf{P})$, i.e., any $\varphi \in H_0$ is the limit $\varphi = \lim \sum_k c_k 1_{B_k}$ of linear combinations of 1_B, $B \in \mathfrak{B}$.

Finally, let us note that, dealing with complex random variables, one can apply complex $\mathcal{L}_p$-spaces ($p = 1, 2$), where $\mathcal{L}_1$ does not need any additional comments, while $\mathcal{L}_2$ is equipped with the inner product

$$(\xi, \eta) = \mathbf{E}\xi\bar{\eta}, \quad \xi, \eta \in \mathcal{L}_2$$

(where, as usual, $\bar{\eta}$ stands for the complex conjugate of η).

2. Conditional Probabilities and Expectations

2.1. PRELIMINARY REMARKS

There is no need to explain how important is to have an instrument to characterize the dependence of various events, random variables, or any random phenomena we are interested in. The most obvious dependence is between events $A \supseteq B$, when we know that the occurrence of B implies the occurrence of A. In a more complicated situation, one can use e.g. the conditional probability $\mathbf{P}(A \mid B)$ of A given B. In a similar way, the conditional expectation $\mathbf{E}(\xi \mid \eta)$ characterizes how a random variable ξ depends on another random variable η. In general, one can be interested in the dependence of an event or a random variable on a family of events and random variables, and the problem is how to define this dependence rigorously. Here, one can proceed in the following way.

Suppose, the information about the occurrence of certain events, or random variables taking certain values etc. can be given by means of an element $y = \eta \in Y$ of

some space Y, which *a priori* is *random*. To be more precise, $\eta \in Y$ is a random element in a measurable space $(Y, \mathfrak{B})$, generating the σ-algebra $\mathfrak{A}$ of events:

$$\{\eta \in B\} \in \mathfrak{A}, \quad B \subseteq Y, \ B \in \mathfrak{B}, \tag{2.1}$$

in our probability model $(\Omega, \mathfrak{B}, \mathbf{P})$. It is convenient to use the notation $B \in \mathfrak{B}$ for the corresponding *event* $\{\eta \in B\}$. With this agreement, we can say that the random element $\eta \in Y$ generates the σ-algebra

$$\mathfrak{B} \subseteq \mathfrak{A} \tag{2.2}$$

of *events* $B \subseteq \Omega$, $B \in \mathfrak{B}$. (Actually, this formal scheme can be applied to any σ-algebra $\mathfrak{B} \subseteq \mathfrak{A}$, given in advance). We are going to define the *conditional probability* $\mathbf{P}(A \mid \mathfrak{B})$ and the *conditional expectation* $\mathbf{E}(\xi \mid \mathfrak{B})$ as functions of $\eta \in Y$, the same as the *conditional probability* $\mathbf{P}(A \mid \eta)$ and the *conditional expectation* $\mathbf{E}(\xi \mid \eta)$, respectively.

Consider any events A and B $(= \{\eta \in B\}$, with $B \in \mathfrak{B})$. Suppose, it happens $\eta = y$; what is the *a posteriori* probability $\mathbf{P}(AB \mid \eta = y)$ of the joint occurrence of A and B? Obviously, in the case $y \in B$ $(B \subseteq Y)$ we can treat B $(B \subseteq \Omega)$ as the *certain* event, hence

$$\mathbf{P}(AB \mid \eta = y) = \mathbf{P}(A \mid \eta = y), \quad y \in B.$$

If $y \notin B$, then B is impossible and

$$\mathbf{P}\{AB \mid \eta = y\} = 0, \quad y \notin B.$$

Together, we can write

$$\mathbf{P}\{AB \mid \eta\} = \mathbf{P}(A \mid \eta)\, 1_B, \tag{2.3}$$

where 1_B is the indicator of the event $B \in \mathfrak{B}$. How can one return to the *a priori* probability $\mathbf{P}(AB)$? Of course, one has to average $\mathbf{P}(AB \mid \eta)$ over all possible outcomes of $\eta \in Y$, i.e., according to (2.3), one obtains $\mathbf{P}(AB)$ as

$$\mathbf{P}(AB) = \mathbf{E}[\mathbf{P}(A \mid \eta)\, 1_B].$$

The above equality can be written as

$$\mathbf{E}\xi 1_B = \mathbf{E}\hat{\xi} 1_B, \quad B \in \mathfrak{B}, \tag{2.4}$$

where $\xi = 1_A$, $\hat{\xi} = \mathbf{P}(A \mid \eta)$. One can recognize in (2.4) the equation which appeared in the best forecast problem (see p. 186). It defines the *projection* $\hat{\xi}$ of the random variable $\xi = 1_A$ in the $\mathcal{L}_2$-space on the subspace $\mathcal{L}_2(\Omega, \mathfrak{B}, \mathbf{P})$, or the best approximation of ξ by means of all functions $\varphi = \varphi(\eta)$. In other words, $\hat{\xi} = \mathbf{P}(A \mid \eta)$ represents the most we can say about the event A, given the 'η-information.'

One can use equation (2.4) to define the conditional expectation $\hat{\xi} = \mathbf{E}(\xi \mid \mathfrak{B})$ of any random variable ξ, $\mathbf{E}|\xi| < \infty$.

2.2. CONDITIONAL EXPECTATION AND ITS PROPERTIES

Consider a random variable ξ, $\mathbf{E}|\xi|^2 < \infty$, as an element of the $\mathcal{L}_2$-space,

$$\mathcal{L}_2 = \mathcal{L}_2(\Omega, \mathfrak{A}, \mathbf{P}).$$

As we know, equation (2.4) identifies the projection $\hat{\xi}$ of $\xi \in \mathcal{L}_2$ onto the subspace $\mathcal{L}_2(\Omega, \mathfrak{B}, \mathbf{P})$, where $\mathfrak{B} \subseteq \mathfrak{A}$ is the σ-algebra, generated by the random variable η (see p. 188). Introduce $\mathbf{E}(\xi \mid \mathfrak{B})$ as the corresponding *projection operator.*

Of course, $\mathbf{E}(\cdot \mid \mathfrak{B})$ is a *linear* operator, and its operator norm in $\mathcal{L}_1$ is

$$\sup_{\|\xi\|=1} \| \mathbf{E}(\xi \mid \mathfrak{B}) \| = \sup_{\mathbf{E}|\xi| \leqslant 1} \mathbf{E} | \mathbf{E}(\xi \mid \mathfrak{B}) | \leqslant 1.$$

Indeed, according to equation (2.4) with $B = \{\hat{\xi} \leqslant 0\}$, $\{\hat{\xi} > 0\}$, we have

$$\begin{aligned} \mathbf{E}|\hat{\xi}| &= -\mathbf{E}\hat{\xi} 1_{\{\hat{\xi} \leqslant 0\}} + \mathbf{E}\hat{\xi} 1_{\{\hat{\xi} > 0\}} \\ &= -\mathbf{E}\xi 1_{\{\hat{\xi} \leqslant 0\}} + \mathbf{E}\xi 1_{\{\hat{\xi} > 0\}} \leqslant \mathbf{E}|\xi|. \end{aligned}$$

Thus, $\mathbf{E}(\xi \mid \mathfrak{B})$ extends to a linear *continuous* (bounded) operator in $\mathcal{L}_1$ since $\mathcal{L}_2$ is *dense* in the $\mathcal{L}_1$-space. Equation (2.4) is valid for any $\xi \in \mathcal{L}_1$ as well, as the limit of the corresponding equality in $\mathcal{L}_2$; of course, the limit of conditional expectations is $\mathfrak{B}$-measurable and satisfies

$$\hat{\xi} \in \mathcal{L}_1(\Omega, \mathfrak{B}, \mathbf{P}). \tag{2.5}$$

Let us show that *equation* (2.4) *uniquely determines* $\hat{\xi} = \mathbf{E}(\xi \mid \mathfrak{B})$, *for any given* $\xi \in \mathcal{L}_1$. Indeed, in the opposite case $\hat{\xi} = \hat{\xi}_1, \hat{\xi}_2$, their difference $\Delta = \hat{\xi}_1 - \hat{\xi}_2$ is $\mathfrak{B}$-measurable and satisfies

$$\mathbf{E}\Delta 1_B = 0, \quad B \in \mathfrak{B}.$$

In particular, for $B = \{\Delta \leqslant 0\}$, $\{\Delta > 0\}$, this leads to

$$-\mathbf{E}\Delta 1_{\{\Delta \leqslant 0\}} + \mathbf{E}\Delta 1_{\{\Delta > 0\}} = \mathbf{E}|\Delta| = 0,$$

which implies $\Delta = 0$ with probability 1, i.e., $\Delta = 0$ as an element of $\mathcal{L}_1$. □

The operator $\mathbf{E}(\xi \mid \mathfrak{B})$ for any $\xi \in \mathcal{L}_1$ defines the corresponding *conditional mathematical expectation* with respect to the σ-algebra $\mathfrak{B}$. In the case when $\mathfrak{B}$ is generated by some random element η, it is also known as the *conditional mathematical expectation* $\mathbf{E}(\xi \mid \eta)$ with respect to η (see p. 60).

Obviously, for any $\mathfrak{B}$-measurable ξ,

$$\mathbf{E}(\xi \mid \mathfrak{B}) = \xi, \tag{2.6}$$

since $\hat{\xi} = \xi$ is a solution of equation (2.4); in particular,

$$\mathbf{E}(1 \mid \mathfrak{B}) = 1.$$

We have already mentioned that the conditional mathematical expectation $\mathbf{E}(\xi \mid \mathfrak{B})$ is *linear* in random variables ξ, i.e.,

$$\mathbf{E}\left[(c_1\xi_1 + c_2\xi_2) \mid \mathfrak{B}\right] = c_1\mathbf{E}(\xi_1 \mid \mathfrak{B}) + c_2\mathbf{E}(\xi_2 \mid \mathfrak{B}) \tag{2.7}$$

for any linear combination of ξ_1, ξ_2.

It is easy to see that, if ξ is independent of all events of the σ-algebra $\mathfrak{B}$ (i.e., ξ is independent of all random variables 1_B, $B \in \mathfrak{B}$), then

$$\mathbf{E}(\xi \mid \mathfrak{B}) = \mathbf{E}\xi. \tag{2.8}$$

Indeed,

$$\mathbf{E}\,\xi\, 1_B = \mathbf{E}\,\xi \cdot \mathbf{E}\, 1_B = \mathbf{E}[(\mathbf{E}\xi)1_B], \quad B \in \mathfrak{B}.$$

Suppose, $\xi \geqslant 0$; then

$$\mathbf{E}(\xi \mid \mathfrak{B}) \geqslant 0. \tag{2.9}$$

Indeed, according to equation (2.4) with $B = \{\hat{\xi} < 0\}$, we have

$$\mathbf{E}\,\hat{\xi}\,1_{\{\hat{\xi}<0\}} = \mathbf{E}\,\xi\,1_{\{\hat{\xi}<0\}} \geqslant 0,$$

which implies $\mathbf{P}\{\hat{\xi} < 0\} = 0$, or $\hat{\xi} \geqslant 0$ as an element of $\mathcal{L}_1$. Of course, inequality (2.9) implies

$$\mathbf{E}(\xi_1 \mid \mathfrak{B}) \leqslant \mathbf{E}(\xi_2 \mid \mathfrak{B}) \tag{2.10}$$

for $\xi_1 \leqslant \xi_2$; in particular,

$$c_1 \leqslant \mathbf{E}(\xi \mid \mathfrak{B}) \leqslant c_2 \tag{2.11}$$

if

$$c_1 \leqslant \xi \leqslant c_2,$$

c_1, c_2 being some *constants*.

Finally, we have the *total mathematical expectation* formula

$$\mathbf{E}\xi = \mathbf{E}[\mathbf{E}(\xi \mid \mathfrak{B})], \tag{2.12}$$

according to equation (2.4) with $B = \Omega$.

2.3. CONDITIONAL PROBABILITY

For any event $A \in \mathfrak{A}$, its *conditional probability* $\mathbf{P}(A \mid \mathfrak{B})$ *with respect to the* σ-algebra $\mathfrak{B}$ *of events* is defined by

$$\mathbf{P}(A \mid \mathfrak{B}) = \mathbf{E}(1_A \mid \mathfrak{B}). \tag{2.13}$$

In the case when $\mathfrak{B}$ is generated by a random element η, (2.13) becomes the *conditional probability with respect to* η:

$$\mathbf{P}(A \mid \mathfrak{B}) = \mathbf{P}(A \mid \eta).$$

According to (2.13),

$$0 \leqslant \mathbf{P}(A \mid \mathfrak{B}) \leqslant 1, \tag{2.14}$$

since $0 \leqslant 1_A \leqslant 1$. For a countable number of any *disjoint* events A_k, $k = 1, 2, \ldots$,

$$\mathbf{P}(A \mid \mathfrak{B}) = \sum_k \mathbf{P}(A_k \mid \mathfrak{B}), \quad A = \bigcup_k A_k, \tag{2.15}$$

since

$$1_A = \sum_k 1_{A_k} = \lim_{n\to\infty} \sum_{k\leqslant n} 1_{A_k}$$

in the $\mathcal{L}_1$-space.

According to (2.6) and (2.8), for any $A \in \mathfrak{B}$

$$\mathbf{P}(A \mid \mathfrak{B}) = 1_A, \tag{2.16}$$

while

$$\mathbf{P}(A \mid \mathfrak{B}) = \mathbf{P}(A) \tag{2.17}$$

for any event A *independent* of the σ-algebra $\mathfrak{B}$. Moreover, for any $A \in \mathfrak{A}$, we have the *total probability formula*:

$$\mathbf{P}(A) = \mathbf{E}\,\mathbf{P}(A \mid \mathfrak{B}). \tag{2.18}$$

Note that the conditional mathematical expectation of a random variable $\xi \in \mathcal{L}_1$ can be determined as

$$\mathbf{E}(\xi \mid \mathfrak{B}) = \lim_{n\to\infty} \mathbf{E}(\xi_n \mid \mathfrak{B}) = \lim_{n\to\infty} \sum_k \frac{k}{n} \mathbf{P}\left\{\frac{k-1}{n} < \xi \leqslant \frac{k}{n} \mid \mathfrak{B}\right\}$$

by means of the corresponding discrete approximations

$$\xi_n = \frac{k}{n}, \qquad \frac{k-1}{n} < \xi \leqslant \frac{k}{n}, \quad k = 0, \pm 1, \ldots,$$

similarly to the definition of

$$\mathbf{E}\xi = \lim \mathbf{E}\xi_n = \lim_{n\to\infty} \sum_k \frac{k}{n} \mathbf{P}\left\{\frac{k-1}{n} < \xi \leqslant \frac{k}{n}\right\},$$

see p. 51.

EXAMPLE. Let $\mathfrak{B}$ be generated by a discrete random variable η taking a countable number of possible values y_k, $k = 1, 2, \ldots$, and let

$$B_k = \{\eta = y_k\}, \quad k = 1, 2, \ldots, \quad \bigcup_k B_k = \Omega,$$

the corresponding disjoint events. Then equation (2.4) with $\xi = 1_A$, $\hat{\xi} = \mathbf{P}(A \mid \mathfrak{B})$, $B = B_k$, $k = 1, 2, \ldots$, gives at once

$$\mathbf{P}(A \mid \mathfrak{B}) = \frac{\mathbf{P}(AB_k)}{\mathbf{P}(B_k)}, \quad \omega \in B_k, \quad k = 1, 2, \ldots, \tag{2.19}$$

as a function of $\omega \in \Omega$ (or outcome B_k, $k = 1, 2, \ldots$).

EXAMPLE (*Discrete conditional distribution*). Suppose, we want to determine the *conditional probability distribution* of ξ with respect to η, where ξ, η are discrete random variables with a given joint probability distribution. I.e., we have to determine the conditional probabilities

$$\mathbf{P}_\xi(x \mid \eta) = \mathbf{P}\{\xi = x \mid \eta\}$$

for all possible values x. It is easy to verify by (2.4) that

$$\mathbf{P}_\xi(x \mid \eta) = \frac{\mathbf{P}_{\xi,\eta}(x, y)}{\mathbf{P}_\eta(y)}, \quad \omega \in \{\eta = y\}. \tag{2.20}$$

EXAMPLE (*Conditional probability density*). Consider random variables ξ, η having the joint probability density; we want to determine the *conditional probability distribution* of ξ given η:

$$\mathbf{P}_\xi(B \mid \eta) = \mathbf{P}\{\xi \in B \mid \eta\}, \quad B \subseteq \mathbb{R}.$$

It is easy to verify by (2.4) that

$$\mathbf{P}_\xi(B \mid \eta) = \int_B p_\xi(x \mid y)\, \mathrm{d}\, x, \quad \omega \in \{\eta = y\}, \tag{2.21}$$

where

$$p_\xi(x \mid y) = \frac{p_{\xi,\eta}(x, y)}{p_\eta(y)}, \quad -\infty < x < \infty,$$

is the conditional probability density.

3. Conditional Expectations and Martingales

3.1. GENERAL PROPERTIES

We start with the following *multiplicative formula*:

$$\mathbf{E}(\varphi\xi \mid \mathfrak{B}) = \varphi\mathbf{E}(\xi \mid \mathfrak{B}), \tag{3.1}$$

valid for any $\mathfrak{B}$-measurable random variable φ; of course, we assume $\varphi\xi \in \mathcal{L}_1$. This formula is obvious for φ of the form

$$\varphi = \sum_k c_k 1_{B_k},$$

involving a finite number of sets $B_k \in \mathfrak{B}$, $k = 1, 2, \ldots$, since, in equation (2.4), we have

$$\mathbf{E}(1_{B_k}\xi)1_B = \mathbf{E}\xi 1_{B_k B} = \mathbf{E}\hat{\xi}1_{B_k B} = \mathbf{E}(1_{B_k}\hat{\xi})1_B,$$

with B, $B_k B \in \mathfrak{B}$. For general φ, use an approximation φ_n of φ of the above type such that $|\varphi_n| \leqslant \varphi$ and $\varphi_n \to \varphi$ with probability 1; then $\varphi_n\xi \to \varphi\xi$ in $\mathcal{L}_1$ by the dominated convergence theorem. Therefore

$$\mathbf{E}(\varphi_n\xi \mid \mathfrak{B}) = \varphi_n\mathbf{E}(\xi \mid \mathfrak{B}) \to \mathbf{E}(\varphi\xi \mid \mathfrak{B})$$

and, simultaneously,

$$\varphi_n\mathbf{E}(\xi \mid \mathfrak{B}) \to \varphi\mathbf{E}(\xi \mid \mathfrak{B})$$

with probability 1, which proves (3.1).

Iterated conditional expectations. Let $\mathfrak{B}' \supset \mathfrak{B}''$ be two σ-algebras. Then

$$\mathbf{E}[\mathbf{E}(\xi \mid \mathfrak{B}') \mid \mathfrak{B}''] = \mathbf{E}(\xi \mid \mathfrak{B}''). \tag{3.2}$$

(3.2) is obvious for $\xi \in \mathcal{L}_2$ as the superposition of two consecutive projections on $H' = \mathcal{L}_2(\Omega, \mathfrak{B}', \mathbf{P})$ and $H'' = \mathcal{L}_2(\Omega, \mathfrak{B}'', \mathbf{P}) \subseteq H'$ is the projection on H''. For $\xi \in \mathcal{L}_1$, (3.2) can be extended by taking a limit in $\mathcal{L}_1$ of elements from $\mathcal{L}_2$.

Increasing and decreasing σ-algebras. Suppose, we deal with *increasing* σ-algebras $\mathfrak{B}_1 \subseteq \mathfrak{B}_2 \subseteq \ldots$. Their limit

$$\mathfrak{B} = \lim_{n\to\infty} \mathfrak{B}_n$$

is the *minimal σ-algebra* $\mathfrak{B}$ containing all $\mathfrak{B}_n$; $\mathfrak{B}$ is actually generated by the *algebra*

$$\bigcup_m \mathfrak{B}_m,$$

which is the union of all $\mathfrak{B}_m$, $m = 1, 2, \ldots$. Then

$$\mathbf{E}(\xi \mid \mathfrak{B}) = \lim_{n\to\infty} \mathbf{E}(\xi \mid \mathfrak{B}_n). \tag{3.3}$$

To prove it, let us first assume $\xi \in \mathcal{L}_2$, then

$$\hat{\xi}_m = \mathbf{E}(\xi \mid \mathfrak{B}_m) = \mathbf{E}[\mathbf{E}(\xi \mid \mathfrak{B}_n) \mid \mathfrak{B}_m] = \mathbf{E}(\hat{\xi}_n \mid \mathfrak{B}_m), \quad m < n,$$

and

$$\| \hat{\xi}_n - \hat{\xi}_m \|^2 = \| \hat{\xi}_n \|^2 - \| \hat{\xi}_m \|^2 \longrightarrow 0, \quad n, m \to \infty,$$

as $\| \hat{\xi}_n \|^2$ increase with n and are bounded by $\| \hat{\xi} \|^2$. Hence, there is a limit

$$\hat{\xi} = \lim_{n\to\infty} \hat{\xi}_n.$$

Moreover, $\hat{\xi} = \mathbf{E}(\xi \mid \mathfrak{B})$, since, for any $B \in \mathfrak{B}_m$ ($m = 1, 2, \ldots$), we have

$$\mathbf{E}\hat{\xi}\, 1_B = \lim_{n\to\infty} \mathbf{E}\hat{\xi}_n\, 1_B = \mathbf{E}\xi\, 1_B.$$

or equation (2.4) for all B from the *algebra* $\bigcup_m \mathfrak{B}_m$, and it can be extended to all $B \in \mathfrak{B}$ using the lemma on events approximation on p. 176. Thus, (3.3) holds for $\xi \in \mathcal{L}_2$; to prove it for $\xi \in \mathcal{L}_1$, one can use an appropriate approximation by elements of the $\mathcal{L}_2$-space (see p. 185).

In a similar way, one can verify (3.3) for any *decreasing* σ-algebras $\mathfrak{B}_1 \supseteq \mathfrak{B}_2 \supseteq \ldots$, with

$$\mathfrak{B} = \lim_{n\to\infty} \mathfrak{B}_n \quad \left(= \bigcap_n \mathfrak{B}_n\right).$$

Martingales. This term* refers, in particular, to random variables ξ_t, $t = 0, 1, \ldots$, having the following property:

$$\mathbf{E}(\xi_t \mid \mathfrak{B}_s) = \xi_s, \quad s \leqslant t, \tag{3.4}$$

with respect to some given *increasing* σ-algebras $\mathfrak{B}_s$, $s = 0, 1, \ldots$.

For example, this property holds for the σ-algebras $\mathfrak{B}_t$ generated by ξ_s, $s \leqslant t$, in the summation scheme

$$\xi_t = \xi_0 + \sum_{0\leqslant u\leqslant t-1} \Delta\xi_u \tag{3.5}$$

with $\xi_0 = 0$ and *independent* increments

$$\Delta\xi_u = \xi_{u+1} - \xi_u, \quad u = 0, 1, \ldots,$$

having zero expectation $\mathbf{E}\Delta\xi_u = 0$. Indeed, since, for $u \geqslant s$, $\Delta\xi_u$ is independent of $\mathfrak{B}_s$, so

$$\mathbf{E}(\Delta\xi_u \mid \mathfrak{B}_s) = \mathbf{E}\Delta\xi_u = 0,$$

and, according to the representation (3.5),

$$\mathbf{E}(\xi_t \mid \mathfrak{B}_s) = \mathbf{E}(\xi_s \mid \mathfrak{B}_s) + \sum_{s\leqslant u\leqslant t-1} \mathbf{E}(\Delta\xi_u \mid \mathfrak{B}_s) = \xi_s.$$

Actually, the scheme (3.5) defines a martingale in the case of arbitrary random variables $\Delta\xi_t$, $t = 0, 1, \ldots$, satisfying

$$\mathbf{E}(\Delta\xi_t \mid \mathfrak{B}_t) = 0, \quad t = 1, 2, \ldots \tag{3.6}$$

* This term is of French origin and describes part of horses' harness.

(why?).

We are going to apply the martingale approach to the problem of finding the mathematical expectation $\mathbf{E}\xi_\tau$ of the random variable

$$\xi_\tau = \xi_0 + \sum_{0 \leqslant u \leqslant \tau - 1} \Delta\xi_u,$$

where τ is a *stopping time*, i.e., a random variable τ with possible values $t = 0, 1, \ldots$ such that, for any t,

$$\{\tau \leqslant t\} \in \mathfrak{B}_t. \tag{3.7}$$

(Roughly speaking, a stopping time τ is a random variable such that the occurrence of any event $\{\tau \leqslant t\}$ is determined by observation of $\xi(s)$, $s \leqslant t$, alone.)

EXAMPLE (*The gambler's ruin problem* (see p. 26)). This problem can be reduced to the scheme (3.5) with $\Delta\xi_t$ equal to the gambler's win at time $t = 0, 1, \ldots$, $\Delta\xi_t = \pm 1$ with equal probability $1/2$. If the initial gambler's capital is x, $0 < x < a$, the game continues up to time τ which is the first time ξ_t, $t = 0, 1, \ldots$, hits the points $a - x$ or $-x$, and the final win is $\xi_\tau = a - x$ with the probability x/a, $\xi_\tau = -x$ with the probability $1 - x/a$. Hence

$$\mathbf{E}\xi_\tau = (a - x)\frac{x}{a} - x\frac{a - x}{a} = 0,$$

which reflects the fact that

$$\mathbf{E}\xi_\tau = \mathbf{E}\xi_0, \tag{3.8}$$

with $\xi_0 = 0$. Suppose now that the gambler's capital is *infinite*, and the game continues up to the first time τ the gambler wins some amount $a > 0$. As we know, $\tau < \infty$ with probability 1 (see p. 29), and the gambler's win at the end of the game is exactly

$$\xi_\tau = a,$$

which contradicts (3.8). □

What can we say about ξ_τ for a *random* τ in the general scheme (3.5), (3.6)? Let us introduce

$$\xi_{t\wedge\tau} = \begin{cases} \xi_t, & t < \tau, \\ \xi_\tau, & t \geqslant \tau, \end{cases} \tag{3.9}$$

where $t \wedge \tau = \min(t, \tau)$ and $\xi_\tau = \xi_s$ for $\tau = s$, $s = 0, 1, \ldots$.

LEMMA. *The random variables* $\xi_{t\wedge\tau}$, $t = 0, 1, \ldots$, *form a martingale.*

To prove it, we apply the following representation:

$$\xi_{t\wedge\tau} = \sum_{s=0}^{t-1} \xi_s 1_{\{\tau=s\}} + \xi_t 1_{\{\tau\geqslant t\}}.$$

Consider the increment

$$\begin{aligned} \Delta\xi_{t\wedge\tau} &= \xi_t 1_{\{\tau=t\}} + \xi_{t+1} 1_{\{\tau\geqslant t+1\}} - \xi_t 1_{\{\tau\geqslant t\}} \\ &= 1_{\{\tau\geqslant t+1\}} \Delta\xi_t, \end{aligned}$$

where the complement $\{\tau \geqslant t+1\}$ to the event $\{\tau \leqslant t\}$ is contained in the σ-algebra $\mathfrak{B}_t$. Hence

$$\mathbf{E}\left[\Delta\xi_{t\wedge\tau} \mid \mathfrak{B}_t\right] = 1_{\{\tau\geqslant t+1\}} \mathbf{E}\left[\Delta\xi_t \mid \mathfrak{B}_t\right] = 0,$$

which is what we need to show only.

Suppose now that $\tau < \infty$ with probability 1; then in (3.9) we obviously have

$$\xi_{t\wedge\tau} \to \xi_\tau, \quad t \to \infty,$$

with probability 1. If, in addition,

$$\mathbf{E}\,\xi_\tau = \lim_{t\to\infty} \mathbf{E}\,\xi_{t\wedge\tau}, \tag{3.10}$$

then we have equality (3.8), since

$$\mathbf{E}\,\xi_{t\wedge\tau} = \mathbf{E}\left[\mathbf{E}(\xi_{t\wedge\tau} \mid \mathfrak{B}_0)\right] = \mathbf{E}\,\xi_0 = 0$$

for the *martingale* $\xi_{t\wedge\tau}$.

For example, (3.10) is true when the random variables $\Delta\xi_t$ are bounded:

$$|\Delta\xi_t| \leqslant C, \quad t = 0, 1, \ldots, \tag{3.11}$$

and the stopping time τ satisfies

$$\mathbf{E}\,\tau < \infty. \tag{3.12}$$

Indeed, we have

$$\xi_{t\wedge\tau} = \sum_{0\leqslant s\leqslant t-1} \Delta\xi_{s\wedge\tau},$$

with increments

$$\Delta\xi_{s\wedge\tau} = 1_{\{\tau\geqslant s+1\}}\Delta\xi_s,$$

and therefore

$$|\xi_{t\wedge\tau}| \leqslant C\sum_{s=1}^{\infty} 1_{\{\tau\geqslant s\}},$$

where the random variable $\eta = \sum 1_{\{\tau\geqslant s\}} \geqslant 0$ has finite expectation:

$$\mathbf{E}\,\eta = \sum_{s=1}^{\infty}\mathbf{E}\,1_{\{\tau\geqslant s\}} = \sum_{s=1}^{\infty}\sum_{u=s}^{\infty}\mathbf{P}\{\tau = u\}$$

$$= \sum_{u=1}^{\infty} u\,\mathbf{P}\{\tau = u\} = \mathbf{E}\,\tau < \infty.$$

This proves (3.10), according to the dominated convergence theorem (see p. 58). □

EXAMPLE (*Wald's identity*). Suppose, we deal with independent random variables η_t, $t = 1, 2, \ldots$, having the same mean value

$$\mathbf{E}\,\eta_t = a.$$

Then

$$\mathbf{E} \sum_{1 \leqslant t \leqslant \tau} \eta_t = a \mathbf{E} \tau \tag{3.13}$$

for any stopping time τ, $\mathbf{E}\tau < \infty$ (with respect to the σ-algebras $\mathfrak{B}_t$ generated by η_s, $s \leqslant t$, $t = 1, 2, \ldots$). To prove (3.13), one can apply the above result to the martingale

$$\xi_0 = 0, \qquad \xi_t = \sum_{1 \leqslant u \leqslant t} (\eta_u - a), \quad t = 1, 2, \ldots,$$

with $\Delta\xi_{t-1} = \eta_t - a$.

CHAPTER 5

Elements of Stochastic Analysis and Stochastic Differential Equations

1. Stochastic Series

1.1. SERIES OF INDEPENDENT RANDOM VARIABLES

It is often very difficult to decide about the convergence of a series $\sum_k x_k$ in the case when $\sum_k |x_k| = \infty$ and the $\pm$ signs of x_k, $k = 1, 2, \ldots$, do not form a regular pattern; of course one can apply the general criterion $\sum_{k=m}^{n} x_k \to 0$, $m, n \to \infty$, but actually nothing else. In such a case, Stochastic Analysis can be helpful provided the signs of the summands follow a typical 'head' or 'tail' sequences in a series of independent coin tossings. For example, let

$$\sum_k \xi_k$$

be a series of independent random variables. According to the 0-1 law, the series $\sum_k \xi_k(\omega)$ converges for outcomes $\omega \in \Omega$ whose total probability is either 0 or 1.

THEOREM. *Let the numerical series*

$$\sum_k \mathbf{E}\xi_k, \qquad \sum_k \mathbf{D}\xi_k$$

converge. Then the series $\sum_k \xi_k$ *of independent variables converges with probability* 1.

To prove the theorem we need the following

LEMMA. *The Kolmogorov inequality*

$$\mathbf{P}\left\{ \max_{1 \leqslant m \leqslant n} \left| \sum_{k=1}^{m} (\xi_k - \mathbf{E}\xi_k) \right| > \varepsilon \right\} \leqslant \frac{1}{\varepsilon^2} \sum_{k=1}^{n} \mathbf{D}\xi_k \tag{1.1}$$

holds for any $\varepsilon > 0$ *and any independent variables* ξ_k, $k = 1, \ldots, n$, *such that* $\mathbf{D}\xi_k < \infty$.

Proof. Without loss of generality, assume $\mathbf{E}\xi_k = 0$. Set

$$S_m = \sum_{k=1}^{m} \xi_k, \quad m = 1, \ldots, n.$$

Write $\nu = 1$ whenever $|S_1| > \varepsilon$ and $\nu = m$ whenever

$$|S_1| \leqslant \varepsilon, \ldots, |S_{m-1}| \leqslant \varepsilon, \quad |S_m| > \varepsilon$$

occurs, $m > 1$. Consider the indicator $1_{\{\nu=m\}}$, $m = 1, \ldots, n$, then $S_m 1_{\{\nu=m\}}$ and $S_n - S_m$ are independent. Hence

$$\mathbf{E}\big[(S_m 1_{\{\nu=m\}})(S_n - S_m)\big] = \mathbf{E}\big(S_m 1_{\{\nu=m\}}\big)\mathbf{E}(S_n - S_m) = 0$$

according to $\mathbf{E}(S_n - S_m) = 0$, which implies

$$\begin{aligned}\mathbf{E}\big(S_n^2 1_{\{\nu=m\}}\big) &= \mathbf{E}\big(S_m^2 1_{\{\nu=m\}}\big) + \mathbf{E}\big[(S_n - S_m)^2 1_{\{\nu=m\}}\big] \\ &\geqslant \mathbf{E}(S_m^2 1_{\{\nu=m\}}), \quad m = 1, \ldots, n.\end{aligned}$$

Hence

$$\begin{aligned}\mathbf{E}S_n^2 &\geqslant \mathbf{E}\Big(S_n^2 \sum_{m=1}^{n} 1_{\{\nu=m\}}\Big) = \sum_{m=1}^{n} \mathbf{E}\Big(S_n^2 1_{\{\nu=m\}}\Big) \\ &\geqslant \sum_{m=1}^{n} \mathbf{E}\Big(S_m^2 1_{\{\nu=m\}}\Big) \geqslant \varepsilon^2 \sum_{m=1}^{n} \mathbf{E}1_{\{\nu=m\}} \\ &= \varepsilon^2 \sum_{m=1}^{n} \mathbf{P}\{\nu = m\} = \varepsilon^2 \mathbf{P}\{\nu \leqslant n\},\end{aligned}$$

since $\nu = m$ implies $S_m^2 > \varepsilon^2$. Thus, for the event

$$\Big\{\max_{1 \leqslant m \leqslant n} |S_m| > \varepsilon\Big\} = \{\nu \leqslant n\}$$

we obtain the corresponding inequality (1.1), as

$$\sum_{k=1}^{n} \mathbf{D}\xi_k = \mathbf{E}S_n^2.$$

Now, we can prove the theorem. As $\sum_k \mathbf{E}\xi_k$ converges, we need to prove the convergence of $\sum_k(\xi_k - \mathbf{E}\xi_k)$. Indeed, for

$$S_n = \sum_{k=1}^{n}(\xi_k - \mathbf{E}\xi_k),$$

according to the Kolmogorov inequality, we have

$$\mathbf{P}\Big\{\sup_{n>m}|S_n - S_m| > \varepsilon\Big\} = \lim_{n\to\infty}\mathbf{P}\Big\{\max_{m\leqslant k\leqslant n}|S_k - S_m| > \varepsilon\Big\}$$
$$\leqslant \frac{1}{\varepsilon^2}\sum_{k\geqslant m}\mathbf{D}\xi_k \longrightarrow 0, \quad m\to\infty,$$

which implies the existence of the limit

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty}\sum_{k=1}^{n}\xi_k$$

with probability 1 (see p. 57).

1.2. THREE SERIES' CRITERION

We start with the following simple observation: the convergence of $\sum_k \xi_k$ (for any particular outcome $\omega \in \Omega$) implies the convergence of

$$\sum_k \xi_k 1_{\{|\xi_k|\leqslant a\}}$$

for any $a > 0$. Indeed, since $\xi_k \to 0$, so

$$\xi_k 1_{\{|\xi_k|\leqslant a\}} = \xi_k, \quad k \geqslant n,$$

for some $n = n(\omega)$. Moreover, for *independent* ξ_k, $k = 1, 2 \ldots$, the convergence $\sum_k \xi_k$ with probability 1 is equivalent to the convergence of

$$\sum_k \xi_k 1_{\{|\xi_k| \leqslant a\}}$$

and

$$\sum_k \mathbf{P}\{|\xi_k| > a\} < \infty,$$

since the latter implies the coincidence of ξ_k and $\xi_k 1_{\{|\xi_k| \leqslant a\}}$, $k \geqslant n$, starting with some *finite* $n = n(\omega)$, according to the Borel–Cantelli lemmas.

Let us consider *bounded* variables, assuming that $|\xi_k| \leqslant a$, $k = 1, 2, \ldots$.

THEOREM. *A series $\sum_k \xi_k$ of bounded independent variables converges with probability* 1 *if and only if*

$$\sum_k \mathbf{E}\xi_k, \qquad \sum_k \mathbf{D}\xi_k$$

converge.

To prove the theorem, we need the *inverse Kolmogorov inequality*

$$\mathbf{P}\left\{ \max_{1 \leqslant m \leqslant n} \left| \sum_{k=1}^{m} (\xi_k - \mathbf{E}\xi_k) \right| > \varepsilon \right\} \geqslant 1 - \frac{(a+\varepsilon)^2}{\sum_{k=1}^{n} \mathbf{D}\xi_k}, \tag{1.2}$$

which proves the theorem in the zero mean case: $\mathbf{E}\xi_k \equiv 0$, since the necessary condition

$$\mathbf{P}\left\{ \sup_{n \geqslant m} |S_n - S_m| > \varepsilon \right\} \longrightarrow 0, \quad m \to \infty,$$

for the convergence of

$$S_n = \sum_{k=1}^{n} \xi_k, \quad n \to \infty,$$

with probability 1 (see p. 57), is satisfied only if $\sum_k \mathbf{D}\xi_k < \infty$ which follows from the inequality

$$\mathbf{P}\Big\{\max_{m\leqslant k\leqslant n} |S_k - S_m| > \varepsilon\Big\} \geqslant 1 - \frac{(a+\varepsilon)^2}{\sum_{k=m}^{n} \mathbf{D}\xi_k}$$

for *all* $n > m$. In the general case, consider a sequence $\widetilde{\xi}_k$, $k = 1, 2, \ldots$, having the same probability distribution as the original one ξ_k, $k = 1, 2 \ldots$, then the convergence of $\sum_k \xi_k$ with probability 1 obviously implies the same for $\sum_k \widetilde{\xi}_k$ hence the convergence of $\sum_k (\xi_k - \widetilde{\xi}_k)$ with

$$\mathbf{E}(\xi_k - \widetilde{\xi}_k) = \mathbf{E}\xi_k - \mathbf{E}\widetilde{\xi}_k = 0.$$

Consequently, the series

$$\sum_k \mathbf{D}(\xi_k - \widetilde{\xi}_k) = 2\sum_k \mathbf{D}\xi_k$$

converges, which implies also the convergence of $\sum_k (\xi_k - \mathbf{E}\xi_k)$ (see p. 201). Finally, the convergence of

$$\sum_k \xi_k, \qquad \sum_k (\xi_k - \mathbf{E}\xi_k)$$

implies the convergence of

$$\sum_k \mathbf{E}\xi_k = \sum_k \big[\xi_k - (\xi_k - \mathbf{E}\xi_k)\big].$$

Now we prove the very inequality (1.2). Assuming $\mathbf{E}\xi_k \equiv 0$ for convenience, we proceed similarly as in the proof of inequality (1.1). Namely,

$$\begin{aligned}
\mathbf{E}\big(S_n^2 1_{\{\nu=m\}}\big) &= \mathbf{E}(S_n - S_m)^2 \cdot \mathbf{E}1_{\{\nu=m\}} + \mathbf{E}\big(S_m^2 1_{\{\nu=m\}}\big) \\
&\leqslant \mathbf{E}S_n^2 \cdot \mathbf{P}\{\nu = m\} + (a+\varepsilon^2)\mathbf{P}\{\nu = m\} \\
&= \mathbf{P}\{\nu = m\}\big[\mathbf{E}S_n^2 + (a+\varepsilon^2)\big]
\end{aligned}$$

as

$$\mathbf{E}(S_n - S_m)^2 = \sum_{k=m+1}^{n} \mathbf{D}\xi_k \leqslant \sum_{k=1}^{n} \mathbf{D}\xi_k = \mathbf{E}S_n^2,$$

and

$$S_m^2 1_{\{\nu=m\}} = \left(S_{m-1}1_{\{\nu=m\}} + \xi_m 1_{\{\nu=m\}}\right)^2 \leqslant (\varepsilon + a)^2$$

because of $|S_{m-1}1_{\{\nu=m\}}| \leqslant \varepsilon$, $|\xi_m| \leqslant a$. Summing up the above inequality over $m = 1, \ldots, n$ brings us to

$$\begin{aligned}
&\mathbf{P}\{\nu \leqslant n\}\left[\mathbf{E}S_n^2 + (a+\varepsilon)^2\right] \geqslant \mathbf{E}S_n^2 1_{\{\nu \leqslant n\}} \\
&= \mathbf{E}S_n^2 - \mathbf{E}S_n^2 1_{\{\nu > n\}} \geqslant \mathbf{E}S_n^2 - \varepsilon^2 \mathbf{P}\{\nu > n\} \\
&= \mathbf{E}S_n^2 - \varepsilon^2 + \varepsilon^2 \mathbf{P}\{\nu \leqslant n\},
\end{aligned}$$

and, finally,

$$\mathbf{P}\{\nu \leqslant n\} \geqslant \frac{\mathbf{E}S_n^2 - \varepsilon^2}{\mathbf{E}S_n^2 + (a+\varepsilon)^2 - \varepsilon^2} \geqslant 1 - \frac{(a+\varepsilon)^2}{\mathbf{E}S_n^2}.$$

□

For arbitrary independent variables ξ_k, $k = 1, 2, \ldots$, the following *criterion* applies: *in order that the series* $\sum_k \xi_k$ *converges with probability* 1, *it is necessary and sufficient that the three series*

$$\sum_k \mathbf{P}\{|\xi_k| > a\}, \qquad \sum_k \mathbf{E}\left(\xi_k 1_{\{|\xi_k| \leqslant a\}}\right), \qquad \sum_k \mathbf{D}\left(\xi_k 1_{\{|\xi_k| \leqslant a\}}\right) \tag{1.3}$$

converge. Actually, this is a *comparison criterion* between the convergence of $\sum_k \xi_k$ and the 'truncated series'

$$\sum_k \xi_k 1_{\{|\xi_n| \leqslant a\}}$$

discussed in the beginning of this section.

EXAMPLE (*Series with 'independent signs'*). Let us consider independent variables ξ_k taking only two values $\pm x_k$, with probabilities p and $q = 1 - p$, correspondingly, $k = 1, 2, \ldots$. According to the three series criterion, in the symmetric case $p = q = 1/2$ ($\mathbf{E}\xi_k \equiv 0$) the series $\sum_k \xi_k$ converges if and only if

$$\sum_k x_k^2 < \infty,$$

while for $p \neq q$, this is true only if

$$\sum_k |x_k| < \infty$$

in addition (why?).

2. Stochastic Integrals

2.1. RANDOM FUNCTIONS (PRELIMINARY REMARKS)

We have encountered before random functions, describing the time evolution of a random process $\xi(t)$, $t \in T$, $T = [0, \infty)$, such as Poisson process, Brownian motion, etc. (see Chapter 2).

There are two basic interpretations of a random function. Firstly, we can treat it as a function $\xi(t)$, $t \in T$, defined on a given set T and taking *values* $\xi(t) \in \mathbb{R}$, which are *random variables*

$$\xi(t) = \xi(\omega, t), \quad \omega \in \Omega,$$

on a probability space $(\Omega, \mathfrak{A}, \mathbf{P})$. Secondly, for any outcome $\omega \in \Omega$ we can consider the real-valued function

$$\xi(\omega, \cdot) = \xi(\omega, t), \quad t \in T,$$

defined on the set T, which is random insofar it depends on the 'random' $\omega \in \Omega$; for any particular $\omega \in \Omega$ this function is usually called *trajectory*, or *realization*, of the random function $\xi(t)$, $t \in T$. The first approach is convenient when we are interested in some properties of $\xi(t)$, $t \in T$, which are determined by the *joint probability distributions*

$$\mathbf{P}_{t_1,\ldots,t_n}(B_1 \times \cdots \times B_n) = \mathbf{P}\{\xi(t_1) \in B_1, \ldots, \xi(t_n) \in B_n\}$$

for various $t_1, \dots, t_n \in T$ and $B_1, \dots, B_n \subseteq \mathbb{R}$; the second one is preferable when we are interested in trajectories having certain desirable properties (recall the Brownian motion model with *continuous* trajectories discussed on p. 119).

Dealing with the properties of $\xi = \xi(t)$, $t \in T$, which are determined by the joint probability distributions $\mathbf{P}_{t_1,\dots,t_n}$, one can always apply the corresponding funcctional model $(X, \mathfrak{B}, \mathbf{P}_\xi)$, $X = \mathbb{R}^T$, which was actually considered when we discussed $\xi(t)$, $t \in T$, as a family of random variables (see p. 174). Having in mind certain desirable properties of the trajectories (such as continuity or integrability), one usually has to replace the original random variables $\xi(t)$ by properly chosen equivalent ones $\widetilde{\xi}(t)$; of course such a replacement does not affect the joint probability distributions $\mathbf{P}_{t_1,\dots,t_n}$ (why?). □

A random function $\xi = \xi(t)$ on $T \subseteq \mathbb{R}$ with $t \in T$ interpreted as *time*, can be considered as a *random process* $\xi(t)$, $t \in T$, which describes the evolution of the random variables $\xi(t)$ (formally, the term 'random process' is equivalent to 'random function').

Dealing with $\xi(t)$, $\mathbf{E}|\xi(t)| < \infty$, one can consider $\xi = \xi(t)$, $t \in T$, as a function with values in the Banach space

$$\mathcal{L}_1 = \mathcal{L}_1(\Omega, \mathfrak{A}, \mathbf{P}).$$

A similar observation applies to $\xi(t)$, $\mathbf{E}|\xi(t)|^2 < \infty$, and the Hilbert space

$$\mathcal{L}_2 = \mathcal{L}_2(\Omega, \mathfrak{A}, \mathbf{P}).$$

Continuity, differentiability, or integrability of a random function with values in $\mathcal{L}_p$ will be referred to as the corresponding property 'in mean' ($p = 1$), and 'in square mean' ($p = 2$), similarly as we used these terms for the corresponding convergence (see p. 55).

Here, we introduce the following characteristics of a random process $\xi(t)$, $t \in T$: the *mean value*

$$A(t) = \mathbf{E}\xi(t), \quad t \in T,$$

and the *correlation function*

$$B(s,t) = \mathbf{E}\big[\xi(s) - A(s)\big]\big[\xi(t) - A(t)\big], \quad s, t \in T.$$

EXAMPLE (*Normal, or Gaussian, random functions*). These are random functions $\xi(t)$, $t \in T$, having normal (Gaussian) joint probability distributions

$$\mathbf{P}_{t_1, \ldots, t_n}(B_1 \times \cdots \times B_n) = \int_{B_1} \cdots \int_{B_n} p_{t_1, \ldots, t_n}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n, \quad B_1, \ldots, B_n \subseteq \mathbb{R},$$

where

$$p_{t_1, \ldots, t_n}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}|B|} \exp\left\{ -\frac{1}{2} \sum_{k,j=1}^{n} b_{kj}(x_k - a_k)(x_j - a_j) \right\},$$

$(x_1, \ldots, x_k) \in \mathbb{R}^n$.

Here,

$$a_k = \mathbf{E}\xi(t_k) = A(t_k),$$
$$|B| = \det\{B_{kj}\} \neq 0, \qquad \{b_{kj}\} = \{B_{kj}\}^{-1},$$
$$B_{kj} = \mathbf{E}\big[\xi(t_k) - A(t_k)\big]\big[\xi(t_j) - A(t_j)\big] = B(t_k, t_j), \quad k, j = 1, \ldots, n$$

(see p. 67). The mean value $A(t)$, $t \in T$, and the correlation function $B(s,t)$, $s, t \in T$, completely determine all joint probability distributions $\mathbf{P}_{t_1, \ldots, t_n}$, $t_1, \ldots, t_n \in T$. □

Let us remark that many interesting probability models deal with *non-differentiable* random functions (such as the Brownian motion), which can be studied using the methods of Stochastic Analysis, the proper stochastic calculus for non-differentiable functions.

2 2 INTEGRATION IN $\mathcal{L}_1$-SPACE

Consider a random function $\xi(t)$, $t \in T$, in the $\mathcal{L}_1$-space, assuming $\mathbf{E}|\xi(t)| < \infty$ and $T \subseteq \mathbb{R}$ an interval.

If $\xi(t)$ is a step function:

$$\xi(t) = \xi_k, \quad t \in \Delta_k, \tag{2.1}$$

taking constant values on a *finite* number of *disjoint* intervals $\Delta_k = (s_k, t_k]$, $\xi(t) = 0$ elsewhere, we set

$$\int_T \xi(t)\,\mathrm{d}t = \sum_k \xi_k |\Delta_k|,$$

with $|\Delta| = t - s$ for any $\Delta = (s, t]$. In the general case, call $\xi(t)$, $t \in T$, *integrable in mean*, if there exist step functions $\xi_n(t)$, $t \in T$, of the form (2.1) approximating $\xi(t)$, $t \in T$:

$$\lim_{n\to\infty} \int_T \|\xi(t) - \xi_n(t)\|\,\mathrm{d}t = 0, \tag{2.2}$$

where $\|\cdot\|$ is the $\mathcal{L}_1$-norm. We have

$$\left\| \int_T \xi_n(t)\,\mathrm{d}t - \int_T \xi_m(t)\,\mathrm{d}t \right\| = \left\| \int_T [\xi_n(t) - \xi_m(t)]\,\mathrm{d}t \right\|$$
$$\leqslant \int_T \|\xi_n(t) - \xi_m(t)\|\,\mathrm{d}t \to 0, \quad m, n \to \infty,$$

so there is the limit

$$\lim_{n\to\infty} \int_T \xi_n(t)\,\mathrm{d}t = \int_T \xi(t)\,\mathrm{d}t$$

which uniquely determines the corresponding *integral*

$$\int_T \xi(t)\,\mathrm{d}t \in \mathcal{L}_1 \tag{2.3}$$

in the $\mathcal{L}_1$-space. □

Observe that any step function

$$\xi(t) = \xi(\omega, t), \quad \omega \in \Omega,\ t \in T,$$

of the form (2.1) is *integrable* with respect to the product measure $\mathbf{P}(\mathrm{d}\omega) \times \mathrm{d}t$, as the function of $(\omega, t) \in \Omega \times T$. The $\mathcal{L}_1$-integral

$$\int_T \xi(t)\,\mathrm{d}t = \int_T \xi(\omega, t)\,\mathrm{d}t, \quad \omega \in \Omega, \tag{2.4}$$

as an element of $\mathcal{L}_1$ can be represented by the random variable which, for any particular outcome $\omega \in \Omega$, is the integral of the corresponding *trajectory* $\xi(\omega, t)$, $t \in T$. This remarkable property can be extended to general random functions $\xi(t)$, $t \in T$, as follows. For approximating step functions $\xi_n(t)$, $t \in T$, condition (2.2) implies

$$\int_T \left[\mathbf{E}|\xi_n(t) - \xi_m(t)|\right] \mathrm{d}t$$
$$= \int_T \int_\Omega |\xi_n(\omega, t) - \xi_m(\omega, t)| \mathbf{P}(\mathrm{d}\omega) \times \mathrm{d}t \longrightarrow 0, \quad n, m \to \infty,$$

which implies the existence of a jointly integrable function

$$\widetilde{\xi}(\omega, t), \quad (\omega, t) \in \Omega \times T,$$

with the finite integral

$$\int_T \int_\Omega \widetilde{\xi}(\omega, t) \mathbf{P}(\mathrm{d}\omega \times \mathrm{d}t), \quad t \in T,$$

such that

$$\int_T \int_\Omega |\widetilde{\xi}(\omega, t) - \xi_n(\omega, t)| \mathbf{P}(\mathrm{d}\omega) \times \mathrm{d}t$$
$$= \int_T \mathbf{E}|\widetilde{\xi}(t) - \xi_n(t)| \, \mathrm{d}t$$
$$= \int_T \|\widetilde{\xi}(t) - \xi(t)\| \, \mathrm{d}t \longrightarrow 0, \quad n \to \infty.$$

Hence, according to condition (2.2), we have

$$\int_T \|\widetilde{\xi}(t) - \xi(t)\| \, \mathrm{d}t = 0,$$

which implies $\widetilde{\xi}(t) = \xi(t)$ with probability 1 for almost all $t \in T$. Substituting $\xi(t)$ by equivalent variables $\widetilde{\xi}(t)$, we obtain a *new* random function

$$\xi(t) = \xi(\omega, t), \quad \omega \in \Omega, \ t \in T,$$

denoted by the same letter as the old one, for simplicity, which is jointly integrable on $\Omega \times T$ with respect to the product measure $\mathbf{P}(d\omega) \times dt$, and such that its $\mathcal{L}_1$-*integral is represented for almost all outcomes* $\omega \in \Omega$ (with probability 1, in other words) *by the trajectory integral*, as in (2.4). The *new* random function $\xi(t)$, $t \in T$, is called *equivalent* to the initial random function, in the sense that, for every $t \in T$, the corresponding random variables are equivalent. □

In a similar way, one can discuss *integration* in the $\mathcal{L}_2$-space, the corresponding integral being defined as an element of $\mathcal{L}_2 \subseteq \mathcal{L}_1$. □

2.3. STOCHASTIC INTEGRALS IN $\mathcal{L}_2$-SPACE

Suppose we want to study a random processes whose local behaviour is given by the equation

$$d\xi(t) = \varphi(t)\, d\eta(t),$$

describing the relationship between infinitesimal time increments of $\eta(t)$ and $\xi(t)$. If $\xi = \xi(t)$ and $\eta = \eta(t)$ are differentiable functions of time $t \geqslant t_0$, one can apply the integral

$$\xi(t) - \xi(t_0) = \int_{t_0}^{t} \varphi(s)\, d\eta(s), \quad t \geqslant t_0.$$

However, if $\eta = \eta(t)$, $t \geqslant t_0$, has *unbounded variation* (as in the case of the Brownian motion), the above integral has to be replaced by an appropriate stochastic integral, which is defined below.

We start with the definition of the stochastic integral for deterministic (non-random) functions $\varphi = \varphi(t)$ and a random process $\eta = \eta(t)$,

$$\eta(t_0) = 0, \qquad \mathbf{E}\eta(t) = 0, \qquad \mathbf{E}|\eta(t)|^2 < \infty,$$

corresponding to a right-continuous function in $\mathcal{L}_2$ with *uncorrelated* (*orthogonal*) increments

$$\Delta\eta = \eta(t) - \eta(s)$$

on disjoint intervals $\Delta = (s, t]$, $s < t$. Such a function is characterized by the right-continuous increasing function

$$F(t) \equiv \mathbf{E}|\eta(t)|^2 = \|\eta(t)\|^2, \quad t \geqslant t_0,$$

with*

$$\Delta F = F(t) - F(s) = \mathbf{E}|\Delta\eta|^2 = \|\Delta\eta\|^2, \quad \Delta = (s,t]. \tag{2.5}$$

For example, $\eta(t)$, $t \geqslant t_0 = 0$, can be the Brownian motion process, with *independent* increments $\Delta\eta(s)$ on disjoint intervals, and

$$F(t) = \mathbf{E}\eta(t)^2 = \sigma^2 t, \quad t \geqslant 0,$$

see p. 115.

Consider an interval $T \subseteq [t_0, \infty)$. Given a step function $\varphi = \varphi(t)$, $t \in T$, taking constant values

$$\varphi(t) = \varphi_k, \quad t \in \Delta_k, \tag{2.6}$$

on a finite number of disjoint intervals $\Delta_k = (s_k, t_k] \subseteq T$, $\varphi(t) = 0$ elsewhere, set

$$\int_T \varphi \, d\eta = \int_T \varphi(t) \, d\eta(t) = \sum_k \varphi_k \Delta_k \eta, \tag{2.7}$$

where $\Delta_k\eta = \eta(t_k) - \eta(s_k)$, $k = 1, \ldots, n$.

Obviously,

$$\mathbf{E} \int_T \varphi \, d\eta = 0 \tag{2.8}$$

and

$$\int_T (c_1\varphi_1 + c_2\varphi_2) \, d\eta = c_1 \int_T \varphi_1 \, d\eta + c_2 \int_T \varphi_2 \, d\eta \tag{2.9}$$

---

* Obviously, for *uncorrelated* $\eta(s) = \eta(s) - \eta(t_0)$ and $\Delta\eta = \eta(t) - \eta(s)$, we have

$$F(t) = \mathbf{E}|\eta(s) + \Delta\eta|^2 = \mathbf{E}|\eta(s)|^2 + \mathbf{E}|\Delta\eta|^2 = F(s) + \|\Delta\eta\|^2.$$

for any linear combination of step functions considered above. Moreover, the stochastic integral satisfies the following properties:

$$\begin{aligned}
&\mathbf{E}\left|\int_T \varphi\,\mathrm{d}\eta\right|^2 = \left\|\int_T \varphi\,\mathrm{d}\eta\right\|^2 = \int_T |\varphi|^2\,\mathrm{d}F,\\
&\mathbf{E}\left[\int_T \varphi_1\,\mathrm{d}\eta\cdot\int_T \varphi_2\,\mathrm{d}\eta\right] = \left(\int_T \varphi_1\,\mathrm{d}\eta,\int_T \varphi_2\,\mathrm{d}\eta\right) = \int_T \varphi_1\varphi_2\,\mathrm{d}F,
\end{aligned} \tag{2.10}$$

where the integrals on the right side are the Lebesgue integrals with respect to the corresponding measure $\mathrm{d}F$. (2.10) can be justified by

$$\left\|\sum_k \varphi_k\Delta_k\eta\right\|^2 = \sum_k |\varphi_k|^2\|\Delta_k\eta\|^2 = \sum_k |\varphi_k|^2\Delta_k F,$$

where the random variables $\varphi_k\Delta_k\eta$ are uncorrelated (orthogonal) and $\Delta_k F = F(t_k) - F(s_k)$, $k = 1,\ldots,n$.

In the general case, consider $\varphi = \varphi(t)$ which can be approximated by step functions $\varphi_n = \varphi_n(t)$ of the form (2.6), in the sense that

$$\int_T |\varphi(t) - \varphi_n(t)|^2\,\mathrm{d}F(t) \to 0. \tag{2.11}$$

Then

$$\begin{aligned}
&\left\|\int_T \varphi_n\,\mathrm{d}\eta - \int_T \varphi_m\,\mathrm{d}\eta\right\|^2 = \left\|\int_T (\varphi_n - \varphi_m)\,\mathrm{d}\eta\right\|^2\\
&\quad = \int_T |\varphi_n - \varphi_m|^2\,\mathrm{d}F \longrightarrow 0, \quad n,m\to\infty,
\end{aligned}$$

so there is the limit in the $\mathcal{L}_2$-space

$$\int_T \varphi\,\mathrm{d}\eta = \int_T \varphi(t)\,\mathrm{d}\eta(t) = \lim_{n\to\infty}\int_T \varphi_n\,\mathrm{d}\eta, \tag{2.12}$$

which satisfies the properties (2.8)–(2.10) as well, for general functions $\varphi = \varphi(t)$ considered above.

EXAMPLE (*Stochastic integral with respect to the Poisson process*). Let $\eta(t)$, $t \geqslant t_0$, be a Poisson process starting at $t_0 = 0$. Here, increments $\Delta\eta$ on disjoint intervals $\Delta = (s, t]$ are independent, but it does not fit into our construction of the stochastic integral since

$$\mathbf{E}\eta(t) = \lambda t, \quad t \geqslant 0.$$

To apply the general scheme (2.6)–(2.12), introduce

$$\eta_0(t) = \eta(t) - \mathbf{E}\eta(t), \quad t \geqslant 0,$$

with the corresponding

$$F(t) = \mathbf{E}\left|\eta_0(t)\right|^2 = \mathbf{D}\eta(t) = \lambda t, \quad t \geqslant 0.$$

If $\varphi = \varphi(t)$ is continuous on $T = (a, b]$, say, one can easily see that

$$\int_a^b \varphi \, \mathrm{d}\eta_0 = \int_a^b \varphi(t) \, \mathrm{d}\eta(t) - \lambda \int_a^b \varphi(t) \, \mathrm{d}t$$

where

$$\int_a^b \varphi(t) \, \mathrm{d}\eta(t) = \sum_{a < \tau_k \leqslant b} \varphi(\tau_k)$$

is the usual Lebesgue–Stieltjes integral with respect to the Poisson process *trajectory* $\eta(t)$, $t \geqslant 0$, which is a right-continuous increasing function having jumps $\eta(\tau_k) - \eta(\tau_k - 0) = 1$ at some points τ_k, $a < \tau_k \leqslant b$ (see p. 92). □

Note that the scheme (2.6)–(2.12) can be applied to a family of random variables $\Delta\eta$, $\mathbf{E}\Delta\eta = 0$, indexed by intervals $\Delta = (s, t) \subseteq T$, respectively, such that

$$\Delta\eta = \sum_k \Delta_k \eta$$

for any finite partition $\Delta = \cup\Delta_k$ by disjoint $\Delta_k = (s_k, t_k]$, $k = 1, 2, \ldots$, and such that for any disjoint intervals $\Delta = (s, t]$ the corresponding random variables $\Delta\eta$ are uncorrelated (orthogonal in the $\mathcal{L}_2$-space), with

$$\mathbf{E}|\Delta\eta|^2 = \|\Delta\eta\|^2 = \Delta F = \int_\Delta \mathrm{d}F$$

given by a *measure* $\mathrm{d}F$. Such a family is called a *stochastic measure* $\mathrm{d}\eta$ with a *structure measure* $\mathrm{d}F$ which is written as

$$\mathbf{E}\mathrm{d}\eta = 0, \qquad \mathbf{E}|\mathrm{d}\eta|^2 = \mathrm{d}F. \tag{2.13}$$

For any given stochastic measure $\mathrm{d}\eta$ satisfying (2.13), we can define the *stochastic integral*

$$\int_T \varphi \,\mathrm{d}\eta = \int_T \varphi(t)\,\mathrm{d}\eta(t)$$

as it was done in (2.6)–(2.12).

EXAMPLE (*Stochastic measure with a given structure measure*). Given an arbitrary structure measure

$$\mathrm{d}F = \mathbf{E}|\mathrm{d}\eta|^2,$$

with

$$\int_{-\infty}^{\infty} \mathrm{d}F = 1,$$

say, we can construct the corresponding stochastic measure as follows. Let η be a random variable with the distribution function

$$F_\eta(t) = \mathbf{P}\{\eta \leqslant t\} = \int_{-\infty}^{t} \mathrm{d}F, \quad -\infty < t < \infty.$$

Set

$$\Delta\eta = I1_\Delta(\eta), \quad \Delta = (s,t], \tag{2.14}$$

where I is any *isometric* operator in the $\mathcal{L}_2$-space mapping indicators $1_\Delta(\eta)$ into the subspace of random variables with zero means. Hence

$$\mathbf{E}\Delta\eta = 0,$$

$$\mathbf{E}|\Delta\eta|^2 = \mathbf{E}\left|1_\Delta(\eta)\right|^2 = \mathbf{P}\{s < \eta \leqslant t\} = \int_\Delta \mathrm{d}F, \quad \Delta = (s,t].$$

Moreover, $\Delta\eta$ are uncorrelated (orthogonal) on disjoint intervals $\Delta = (s, t]$, since

$$\mathbf{E}\Delta_1\eta \cdot \Delta_2\eta = \mathbf{E}1_{\Delta_1}(\eta)1_{\Delta_2}(\eta) = \mathbf{E}1_{\Delta_1\cap\Delta_2}(\eta)$$

for any $\Delta_1 = (s_1, t_1]$, $\Delta_2 = (s_2, t_2]$. □

The scheme (2.6)–(2.12) can be obviously applied to any *stochastic measure* $\mathrm{d}\eta$ on a domain $T \subseteq \mathbb{R}^n$ and satisfying (2.13). One can start with $\Delta\eta$ defined on multiintervals

$$\Delta = (s_1, t_1] \times \cdots \times (s_n, t_n],$$

say, and then extend it to the stochastic integral

$$\int_T \varphi \,\mathrm{d}\eta = \int_T \varphi(t)\,\mathrm{d}\eta(t)$$

on $T \subseteq \mathbb{R}^n$ satisfying all properties (2.8)–(2.9). □

In fact, the stochastic integral (2.12) is well defined for any function $\varphi = \varphi(t)$, $t \in T$, with

$$\int_T \left|\varphi(t)\right|^2 \mathrm{d}F(t) < \infty, \tag{2.15}$$

since any such φ can be approximated in the sense of (2.11) by appropriate step functions φ_n, $n = 1, 2, \ldots$.

Under condition (2.15), the corresponding stochastic integral satisfies the relation

$$\int_\Delta \varphi(t)\,\mathrm{d}\eta(t) \equiv \int_T \left[\varphi 1_\Delta(t)\right] \mathrm{d}\eta(t), \quad \Delta \subseteq T.$$

As a function of $\Delta \subseteq T$, the last integral represents a *stochastic measure* on T, denoted by $\varphi\,\mathrm{d}\eta$, such that

$$\mathbf{E}[\varphi\,\mathrm{d}\eta] = 0, \qquad \mathbf{E}|\varphi\,\mathrm{d}\eta|^2 = |\varphi|^2\,\mathrm{d}F. \tag{2.16}$$

□

Finally, the (2.6)–(2.16) can be discussed in the *complex* $\mathcal{L}_2$-space as well, assuming that the corresponding stochastic measure $d\eta$ is complex-valued, and that its values $\Delta\eta$ on disjoint $\Delta \subseteq T$ are *orthogonal* in this space, with the only change that the second part of (2.10) reads now

$$\mathbf{E}\left[\int_T \varphi_1 \, d\eta \cdot \overline{\int_T \varphi_2 \, d\eta}\right] = \left(\int_T \varphi_1 \, d\eta, \int_T \varphi_2 \, d\eta\right) = \int_T \varphi_1 \overline{\varphi}_2 \, dF. \tag{2.10'}$$

2.4. STOCHASTIC ITO INTEGRAL IN $\mathcal{L}_2$-SPACE

In (2.6)–(2.12), we dealt with deterministic (non-random) functions $\varphi = \varphi(t)$. Can one discuss *random* functions $\varphi = \varphi(t)$, $t \geqslant t_0$, and define the corresponding stochastic integral

$$\int_a^b \varphi \, d\eta = \int_a^b \varphi(t) \, d\eta(t)$$

on $T = (a, b]$, say? The following construction yields the so-called *stochastic Ito integral* (in the $\mathcal{L}_2$-space).

The main point is to consider *non-anticipating* random functions $\varphi = \varphi(t)$, $\mathbf{E}|\varphi(t)|^2 < \infty$, measurable with respect to the corresponding σ-algebra $\mathfrak{B}_t$ which, roughly speaking, represents all events up to time t; the increasing σ-algebras

$$\mathfrak{B}_s \subseteq \mathfrak{B}_t, \quad t_0 \leqslant s \leqslant t,$$

are assumed to be right-continuous, i.e.,

$$\mathfrak{B}_s = \lim_{t \to s} \mathfrak{B}_t \left(= \bigcap_{t > s} \mathfrak{B}_t \right).$$

In addition to the initial scheme (2.6)–(2.12), $\eta = \eta(t)$, $t \geqslant t_0$, is supposed to be a *random process with independent increments* on disjoint intervals such that

$$\Delta\eta = \eta(t) - \eta(s), \qquad \Delta = (s, t],$$

for all 'future' times $t > s$ does not depend on the 'past' $\mathfrak{B}_s$, representing on intervals $\Delta = (s, t]$ a stochastic measure $d\eta$ with zero mean $\mathbf{E}d\eta = 0$ and the structure measure

$$\mathbf{E}|d\eta|^2 = dF.$$

Similarly to the scheme (2.6)–(2.12), one can start with *random non-anticipating* functions of the form (2.6), i.e.,

$$\varphi(t) = \varphi_k, \quad t \in \Delta_k,$$

on disjoint intervals $\Delta_k = (s_k, t_k] \subseteq T$, where random variables φ_k, $\mathbf{E}|\varphi_k|^2 < \infty$, are measurable with respect to the corresponding σ-algebra

$$\lim_{h \to +0} \mathfrak{B}_{s_k + h} = \mathfrak{B}_{s_k}, \quad k = 1, \ldots, n.$$

Of course, for the stochastic integral of the form (2.7), (2.9) is trivial, and (2.8) follows from

$$\mathbf{E} \int_T \varphi \, d\eta = \sum_k \left[\mathbf{E}\varphi_k \cdot \mathbf{E}\Delta_k \eta \right] = 0$$

since the increment $\Delta_k \eta = \eta(t_k) - \eta(s_k)$ does not depend on the random variable φ_k which is measurable with respect to $\mathfrak{B}_{s_k}$, $k = 1, \ldots, n$ (see p. 48 on the multiplicative property of the mathematical expectation). Equations (2.10) for the norm and inner product in the $\mathcal{L}_2$-space become now

$$\begin{aligned} \mathbf{E}\left| \int_T \varphi \, d\eta \right|^2 &= \left\| \int_T \varphi \, d\eta \right\|^2 = \int_T \left(\mathbf{E}|\varphi|^2 \right) dF \\ &= \int_T \left\| \varphi(t) \right\|^2 dF(t), \end{aligned} \tag{2.17}$$

$$\mathbf{E}\left(\int_T \varphi_1 \, d\eta \cdot \int_T \varphi_2 \, d\eta \right) = \left(\int_T \varphi_1 \, d\eta, \int_T \varphi_2 \, d\eta \right)$$

$$= \int_T \mathbf{E}(\varphi_1 \varphi_2) \, dF = \int_T \left(\varphi_1(t), \varphi_2(t) \right) dF(t).$$

Indeed,

$$\begin{aligned}
&\mathbf{E}\Big|\sum_k \varphi_k \cdot \Delta_k\eta\Big|^2 \\
&= \sum_k \mathbf{E}\big|\varphi_k \cdot (\Delta_k\eta)\big|^2 + 2\sum_{k<j} \mathbf{E}\big[\varphi_k(\Delta_k\eta)\varphi_j(\Delta_j\eta)\big] \\
&= \sum_k \mathbf{E}|\varphi_k|^2 \cdot \mathbf{E}\big|\Delta_k\eta\big|^2 + 2\sum_{k<j} \mathbf{E}\big[\varphi_k(\Delta_k\eta)\varphi_j\big]\mathbf{E}(\Delta_j\eta) \\
&= \sum_k \big(\mathbf{E}|\varphi_k|^2\big)\Delta_k F = \int_T \big(\mathbf{E}|\varphi|^2\big)\,\mathrm{d}F.
\end{aligned}$$

In the general case consider a non-anticipating function $\varphi = \varphi(t)$, $\mathbf{E}|\varphi(t)|^2 < \infty$, which can be approximated by non-anticipating functions $\varphi_k = \varphi_k(t)$ of the form (2.6), in the sense that

$$\begin{aligned}
&\int_T \mathbf{E}\big|\varphi(t) - \varphi_n(t)\big|^2\,\mathrm{d}F(t) \\
&= \int_T \big\|\varphi(t) - \varphi_n(t)\big\|^2\,\mathrm{d}F(t) \longrightarrow 0, \quad n \to \infty.
\end{aligned}$$

The corresponding stochastic integrals satisfy

$$\begin{aligned}
&\Big\|\int_T \varphi_n\,\mathrm{d}\eta - \int_T \varphi_m\,\mathrm{d}\eta\Big\|^2 = \Big\|\int_T (\varphi_n - \varphi_m)\,\mathrm{d}\eta\Big\|^2 \\
&= \int_T \big\|\varphi_n(t) - \varphi_m(t)\big\|^2\,\mathrm{d}F(t) \longrightarrow 0, \quad n, m \to \infty,
\end{aligned}$$

and therefore converge in the $\mathcal{L}_2$-space, giving the *stochastic integral*

$$\int_T \varphi\,\mathrm{d}\eta = \int_T \varphi(t)\,\mathrm{d}\eta(t) = \lim_{n\to\infty} \int_T \varphi_n\,\mathrm{d}\eta. \tag{2.18}$$

Obviously, this stochastic integral inherits all properties claimed earlier in (2.8), (2.9) and (2.17).

In particular, the stochastic integral (2.18) exists for any non-anticipating continuous (in the $\mathcal{L}_2$-space) function $\varphi = \varphi(t)$ on $T = (a, b]$, as

$$\int_a^b \varphi \, d\eta = \lim_{n\to\infty} \sum_{k=1}^{n} \varphi(t_{k-1})\big[\eta(t_k) - \eta(t_{k-1})\big] \tag{2.19}$$

for any sequence of partitions

$$T = \bigcup_{k=1}^{n} (t_{k-1}, t_k]$$

with

$$\max_k (t_k - t_{k-1}) \to 0.$$

EXAMPLE (*Stochastic Ito integral with respect to the Brownian motion*). Integration of the Brownian motion process $\varphi(t) = \eta(t)$, $t \geqslant t_0$, with respect to itself yields

$$\int_a^b \eta(t) \, d\eta(t) = \frac{1}{2}\big[\eta(b)^2 - \eta(a)^2\big] - \frac{1}{2}(b - a) \tag{2.20}$$

instead of the formula

$$\int_a^b \eta(s) \, d\eta(s) = \frac{1}{2}\big[\eta(b)^2 - \eta(a)^2\big],$$

which is true in the case of a differentiable function (however, the Brownian motion is *not* differentiable). To verify (2.20), write

$$\begin{aligned}&\eta(t_{k-1})\big[\eta(t_k) - \eta(t_{k-1})\big] \\ &= \frac{1}{2}\big[\eta(t_k)^2 - \eta(t_{k-1})^2\big] - \frac{1}{2}\big[\eta(t_k) - \eta(t_{k-1})\big]^2, \quad k = 1, \ldots, n,\end{aligned}$$

and then use formula (2.19) with $\varphi(t) = \eta(t)$. We obtain

$$\begin{aligned}&\lim_{n\to\infty} \sum_{k=1}^{n} \eta(t_{k-1})\big[\eta(t_k) - \eta(t_{k-1})\big] \\ &= \frac{1}{2}\big[\eta(b)^2 - \eta(a)^2\big] - \lim_{n\to\infty} \frac{1}{2} \sum_{k=1}^{n} \big[\eta(t_k) - \eta(t_{k-1})\big]^2 \\ &= \frac{1}{2}\big[\eta(b)^2 - \eta(a)^2\big] - \frac{1}{2}(b - a),\end{aligned}$$

according to the formula for the 'quadratic variation'

$$\lim_{n\to\infty} \sum_{k=1}^{n} \left[\eta(t_k) - \eta(t_{k-1})\right]^2 = b - a$$

of the Brownian motion (see p. 120).

3. Stochastic Integral Representations

3.1. CANONICAL REPRESENTATIONS

We call a canonical representation any stochastic integral representation

$$\xi(t) = \int_{\Lambda} \varphi(t, \lambda)\, \mathrm{d}\eta(\lambda), \quad t \in T, \tag{3.1}$$

of a given (real or complex) random process $\xi(t)$, $t \in T$, where $\mathrm{d}\eta(\lambda)$, $\lambda \in \Lambda$, is a stochastic measure on a parameter space $\Lambda \subseteq \mathbb{R}$, and $\varphi(t, \lambda)$ is a deterministic function of $t \in T$, $\lambda \in \Lambda$. If $\mathrm{d}\eta$ is a stochastic measure with

$$\mathbf{E}\mathrm{d}\eta = 0, \qquad \mathbf{E}|\mathrm{d}\eta|^2 = \mathrm{d}F, \tag{3.2}$$

then the mean value

$$\mathbf{E}\xi(t) = 0, \quad t \in T,$$

and the *correlation function*

$$B(s, t) = \mathbf{E}\xi(s)\bar{\xi}(t) = \int_{\Lambda} \varphi(s, \lambda)\overline{\varphi(t, \lambda)}\, \mathrm{d}F(\lambda), \quad s, t \in T, \tag{3.3}$$

according to (3.1) and general properties (2.8), (2.10) of stochastic integrals.

As a matter of fact, one can obtain a representation (3.1) for any random process $\xi(t)$, $t \in T$, with zero mean and the correlation function of the form (3.3). Namely, assume that the probability model $(\Omega, \mathfrak{A}, \mathbf{P})$ is sufficiently rich so that one can define a stochastic measure $\mathrm{d}\tilde{\eta}$ of the given structure

$$\mathbf{E}\mathrm{d}\tilde{\eta} = 0, \qquad \mathbf{E}|\mathrm{d}\tilde{\eta}|^2 = \mathrm{d}F$$

(see (2.14), for example). Then the random process $\xi(t)$, $t \in T$, is *isometric* in the $\mathcal{L}_2$-space to the random process

$$\widetilde{\xi}(t) = \int_{\Lambda} \varphi(t, \lambda)\, \mathrm{d}\widetilde{\eta}(\lambda), \quad t \in T,$$

since

$$\begin{aligned}\big(\xi(s), \xi(t)\big) &= \mathbf{E}\xi(s)\overline{\xi(t)} = \mathbf{E}\widetilde{\xi}(s)\overline{\widetilde{\xi}(t)} \\ &= \big(\widetilde{\xi}(s), \widetilde{\xi}(t)\big), \quad t \in T.\end{aligned}$$

The stochastic measure

$$\mathrm{d}\eta(\lambda) = I\mathrm{d}\widetilde{\eta}(\lambda)$$

can be obtained by means of the corresponding *isometric operator* I in the $\mathcal{L}_2$-space such that

$$I\widetilde{\xi}(t) = \xi(t), \quad t \in T.$$

Consequently,

$$\begin{aligned}\xi(t) = I\widetilde{\xi}(t) &= I \int_{\Lambda} \varphi(t, \lambda)\, \mathrm{d}\widetilde{\eta}(\lambda) \\ &= I\Big[\lim \sum_{k} \varphi_k \Delta_k \widetilde{\eta}\Big] \\ &= \lim \sum_{k} \varphi_k \big[I\Delta_k \widetilde{\eta}\big] = \lim \sum_{k} \varphi_k \Delta_k \eta = \int_{\Lambda} \varphi(t, \lambda)\, \mathrm{d}\eta(\lambda), \quad t \in T;\end{aligned}$$

here we used a corresponding stochastic integral approximation (see (2.7), (2.12)) to make things clear.

EXAMPLE (*Canonical representation by stochastic functional series*). Let $\xi(t)$, $t \in T$, be a random process with zero mean and *continuous* correlation function, defined on a finite interval $T \subseteq \mathbb{R}$. The correlation function $B(s, t)$, $s, t \in T$, being continuous, *symmetric*:

$$B(s, t) = \overline{B(t, s)}, \quad s, t \in T,$$

and *positive definite*: for any $c_k,\ t_k,\ k = 1, \ldots, n$,

$$\sum_{k,j=1}^{n} c_k \bar{c}_j B(t_k, t_j) = \mathbf{E}\left|\sum_{k=1}^{n} c_k \xi(t_k)\right|^2 \geqslant 0,$$

there exists a complete system of *eigenfunctions*

$$\varphi_k(t) = \varphi(t, \lambda_k), \quad k = 1, 2, \ldots,$$

which satisfy the equation

$$\int_T B(s,t)\varphi(t,\lambda)\,\mathrm{d}t = \lambda\varphi(s,\lambda), \quad s \in T,$$

and the orthogonality conditions

$$\int_T \varphi_k(t)\overline{\varphi_j(t)}\,\mathrm{d}t = \begin{cases} 1, & k = j, \\ 0, & k \neq j. \end{cases}$$

This leads us to the representation

$$\begin{aligned} B(s,t) &= \sum_k \lambda_k \varphi(s,\lambda_k)\varphi(t,\lambda_k) \\ &= \int_\Lambda \varphi(s,\lambda)\varphi(t,\lambda)\,\mathrm{d}F(\lambda), \quad s,t \in T, \end{aligned} \tag{3.4}$$

of the type (3.3), with the *spectral measure* $\mathrm{d}F(\lambda) = \lambda$ concentrated on the discrete set $\Lambda \subseteq \mathbb{R}$ of all *eigenvalues* $\lambda = \lambda_k \geqslant 0$,

$$\sum_k \lambda_k < \infty.$$

The above representation of the correlation function, known as Mercer's theorem, leads to the canonical representation

$$\xi(t) = \sum_k \eta_k \varphi_k(t) = \int_\Lambda \varphi(t,\lambda)\,\mathrm{d}\eta(\lambda), \quad t \in T, \tag{3.5}$$

of $\xi(t)$, as a stochastic functional series of $\varphi_k(t) = \varphi(t, \lambda)$ with *uncorrelated* random coefficients η_k as atoms of the stochastic measure $\mathrm{d}\eta(\lambda)$ at $\lambda = \lambda_k$,

$$\mathbf{E}\eta_k = 0, \qquad \mathbf{E}|\eta_k|^2 = \lambda_k, \quad k = 1, 2, \ldots .$$

EXAMPLE (*Spectral representation of a stationary process*). Let $\xi(t)$, $-\infty < t < \infty$, be a random process with *constant mean value* $\mathbf{E}\xi(t) = a$ (below, we assume $a = 0$) and such that the correlation between $\xi(s), \xi(t)$ depends only on $s - t$:

$$B(s, t) = \mathbf{E}\xi(s)\overline{\xi(t)} = B(s - t), \quad -\infty < s, t < \infty.$$

Such random processes, corresponding to a (*continuous*) *correlation* function

$$B(t) = \mathbf{E}\xi(t + s)\overline{\xi(s)} = \mathbf{E}\xi(t)\overline{\xi(0)}, \quad -\infty < t < \infty,$$

are usually called *stationary* (*in the wide sense*). Being *positive definite*:

$$\sum_{k,j=1}^{n} c_k \bar{c}_j B(t_k - t_j) = \mathbf{E}\left|\sum_{k=1}^{n} c_k \xi(t_k)\right|^2 \geqslant 0,$$

see above, the continuous function $B(t)$, $-\infty < t < \infty$, can be written as the *Fourier integral*

$$B(t) = \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda t}\, \mathrm{d}F(\lambda), \quad -\infty < t < \infty, \tag{3.6}$$

with some bounded measure $\mathrm{d}F(\lambda) \geqslant 0$, due to the well-known *Bochner–Khinchin theorem*. The representation

$$B(s, t) = B(s - t) = \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda(s-t)}\, \mathrm{d}F(\lambda), \quad -\infty < s, t < \infty,$$

being of the type (3.3), we obtain the corresponding canonical representation (3.1):

$$\xi(t) = \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda t}\, \mathrm{d}\eta(\lambda), \quad -\infty < t < \infty, \tag{3.7}$$

called the *spectral representation* of the stationary random process $\xi(t), \infty < t < \infty$. The stochastic measure $d\eta(\lambda)$ of the structure

$$\mathbf{E}\, d\eta = 0, \qquad \mathbf{E}|d\eta|^2 = dF$$

is called the *stochastic spectral measure*, and dF itself is called the *spectral measure* of the stationary process. The spectral representation (3.7) suggests the following approximation

$$\xi(t) = \lim \sum_k e^{i\lambda_k t} \Delta_k \eta$$

of $\xi(t)$ by means of uncorrelated *random* oscillations $e^{i\lambda_k t}\Delta_k\eta$, $-\infty < t < \infty$, with the corresponding frequencies $|\lambda_k|$, $-\infty < \lambda_k < \infty$. Roughly speaking, (3.7) itself is a superposition of *stochastic* uncorrelated *harmonics*

$$e^{i\lambda t} d\eta(\lambda), \quad -\infty < t < \infty.$$

The spectral measure $dF(\lambda)$ gives the distribution of the total 'mean energy',

$$\mathbf{E}|\xi(t)|^2 = \int_{-\infty}^{\infty} dF(\lambda), \quad -\infty < t < \infty,$$

over various harmonics $e^{i\lambda t} d\eta(\lambda)$, $-\infty < t < \infty$, with the corresponding amplitude square mean equal to

$$\mathbf{E}|d\eta(\lambda)|^2 = dF(\lambda), \quad -\infty < \lambda < \infty.$$

□

Interchange of the order of integration. Given a canonical representation (3.1), one often has to deal with the integral

$$\int_T \left[\int_\Lambda \varphi(t, \lambda) \eta(d\lambda) \right] dt$$

in the $\mathcal{L}_2$-space, say. Can we interchange here the integration order? To simplify the analysis of this question, assume that

$$\int_T \int_\Lambda |\varphi(t, \lambda)|\, dF(\lambda)\, dt < \infty.$$

The answer is positive, i.e.

$$\eta_1 = \int_T \left[\int_\Lambda \varphi(t,\lambda)\, d\eta(\lambda) \right] dt = \int_\Lambda \left[\int_T \varphi(t,\lambda)\, dt \right] d\eta(\lambda) = \eta_2. \tag{3.8}$$

Indeed, η_1 and η_2 are the limits in the $\mathfrak{L}_2$-space of linear combinations of $\Delta\eta$'s on various $\Delta \subseteq \Lambda$, and we need only to verify that the inner product

$$(\eta_1, \Delta\eta) = \int_T \left(\int_\Lambda \varphi(t,\lambda)\, d\eta(\lambda), \Delta\eta \right) dt = \int_T \left[\int_\Delta \varphi(t,\lambda)\, dF(\lambda) \right] dt$$

is the same as

$$(\eta_2, \Delta\eta) = \int_\Delta \left[\int_T \varphi(t,\lambda)\, dt \right] dF(\lambda).$$

Obviously, this is true, thanks to the Fubini theorem on the interchange of the integration order.

3.2. SPECTRAL REPRESENTATION OF A STATIONARY PROCESS AND ITS APPLICATIONS

The stochastic integral representation

$$\xi(t) = \int_{-\infty}^{\infty} e^{i\lambda t}\, d\eta(\lambda), \quad -\infty < t < \infty,$$

of a stationary process (see (3.7)) is very useful in the study of its *linear transformations*

$$\xi_\varphi(t) = \int_{-\infty}^{\infty} e^{i\lambda t} \varphi(\lambda)\, d\eta(\lambda), \quad -\infty < t < \infty. \tag{3.9}$$

Given $\varphi(\lambda)$, $-\infty < \lambda < \infty$, (3.9) transforms harmonic components of the initial process by multiplying them by the corresponding 'weights' $\varphi(\lambda)$, depending on the frequency λ. Such a transformation can give more 'weight' to some components and less 'weight' to another ones. Of course, the weight function $\varphi(\lambda)$ has to fulfill the condition

$$\int_{-\infty}^{\infty} |\varphi(\lambda)|^2\, dF(\lambda) < \infty,$$

under which the stochastic integral on the right side of (3.9) is defined for each t (see (2.15)).

EXAMPLE (*Differentiation*). Let the spectral measure $\mathrm{d}F$ satisfy

$$\int_{-\infty}^{\infty} |\lambda|^2 \,\mathrm{d}F(\lambda) < \infty.$$

Then the stationary process $\xi(t)$ has the derivative in the square mean:

$$\begin{aligned}\xi'(t) &= \lim_{h\to 0} \frac{\xi(t+h) - \xi(t)}{h} \\ &= \int_{-\infty}^{\infty} \left[\lim_{h\to\infty} \frac{\mathrm{e}^{i\lambda(t+h)} - \mathrm{e}^{i\lambda t}}{h}\right] \mathrm{d}\eta(\lambda) \\ &= \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda t}(i\lambda)\,\mathrm{d}\eta(\lambda) = \xi_{\varphi}(t), \quad -\infty < t < \infty, \quad \varphi(\lambda) = i\lambda.\end{aligned}$$

EXAMPLE (*Integration*). Consider an integrable function $c(t)$, $-\infty < t < \infty$, then, with

$$\int_{-\infty}^{\infty} \left|c(t)\right| \mathrm{d}t < \infty, \ \varphi(\lambda) = \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda t} c(t)\,\mathrm{d}t,$$

we have

$$\begin{aligned}\xi_{\varphi}(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda t}\varphi(\lambda)\,\mathrm{d}\eta(\lambda) = \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda t}\left[\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda t} c(s)\,\mathrm{d}s\right] \mathrm{d}\eta(\lambda) \\ &= \int_{-\infty}^{\infty} c(s)\left[\int_{-\infty}^{\infty} \mathrm{e}^{i\lambda(t-s)}\,\mathrm{d}\eta(\lambda)\right] \mathrm{d}s = \int_{-\infty}^{\infty} c(s)\,\xi(t-s)\,\mathrm{d}s \\ &= \int_{-\infty}^{\infty} c(t-s)\,\xi(s)\,\mathrm{d}s, \quad -\infty < t < \infty.\end{aligned}$$

□

Let us consider a random process $\xi(t)$, $-\infty < t < \infty$, with a *constant* mean value

$$\mathbf{E}\xi(t) = a,$$

such that

$$\xi_0(t) = \xi(t) - a, \quad -\infty < t < \infty,$$

is a stationary process with *zero* mean and a spectral measure $F_0(\mathrm{d}\lambda)$ which is *continuous* at $\lambda = 0$: $\mathrm{d}F_0(0) = 0$. Introduce *the sample average*

$$\frac{1}{T}\int_0^T \xi(t)\,\mathrm{d}t = \frac{1}{T}\int_0^T \xi_0(t)\,\mathrm{d}t + a$$

over the time interval $0 \leqslant t \leqslant T$, $T \to \infty$.

According to the spectral representation

$$\xi_0(t) = \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda t}\,\mathrm{d}\eta_0(\lambda), \quad -\infty < t < \infty,$$

we have

$$\frac{1}{T}\int_0^T \xi_0(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} \frac{\mathrm{e}^{i\lambda T} - 1}{i\lambda T}\,\mathrm{d}\eta_0(\lambda).$$

Therefore

$$\left|\frac{1}{T}\int_0^T \xi_0(t)\,\mathrm{d}t\right|^2 = \int_{-\infty}^{\infty} \left|\frac{\mathrm{e}^{i\lambda T} - 1}{i\lambda T}\right|^2 \mathrm{d}F_0(\lambda)$$

$$= \int_{\lambda \neq 0} \left|\frac{\mathrm{e}^{i\lambda T} - 1}{i\lambda T}\right|^2 \mathrm{d}F_0(\lambda) \longrightarrow 0, \quad T \to \infty,$$

since $\mathrm{d}F_0(\lambda) = 0$ and

$$\left|\frac{\mathrm{e}^{i\lambda T} - 1}{i\lambda T}\right| \longrightarrow 0, \quad T \to \infty,$$

boundedly for any $\lambda \neq 0$. Thus, we get the following result.

THEOREM. *Given a stationary process* $\xi(t)$, $-\infty < t < \infty$, *there is the limit*

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \xi(t)\,\mathrm{d}t = a \tag{3.10}$$

(*in the square mean*), *which coincides with the mean value* $\mathbf{E}\xi(t) \equiv a$.

This result is similar to the *law of large numbers* and is known as the *Ergodic Theorem* for stationary processes (in the $\mathcal{L}_2$-space).

Stationary processes in the strict sense. This term usually indicates that, for any $t_1, \dots, t_n$, the joint probability distribution of random variables $\xi(t_1), \dots, \xi(t_n)$ coincides with the joint probability distribution of $\xi(t_1 + t), \dots, \xi(t_n + t)$, for any time shift t, $-\infty < t < \infty$.

Consider a *nonlinear transformation* of $\xi(t)$ of the form

$$\xi_\varphi(t) = \varphi\big[\xi(t_1 + t), \dots, \xi(t_n + t)\big], \quad -\infty < t < \infty.$$

If

$$\mathbf{E}\big|\varphi\big[\xi(t_1), \dots, \xi(t_n)\big]\big|^2 < \infty,$$

then $\xi_\varphi(t)$ is a stationary process in the wide sense, with the *constant* mean value

$$\mathbf{E}\xi_\varphi(t) \equiv \mathbf{E}\varphi\big[\xi(t_1), \dots, \xi(t_n)\big].$$

The process $\xi(t)$, $-\infty < t < \infty$, itself is called *ergodic*, if any stationary process $\xi_\varphi(t)$, $-\infty < t < \infty$, of the above form satisfies

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \xi_\varphi(t)\, \mathrm{d}t = \mathbf{E}\varphi\big[\xi(t_1), \dots, \xi(t_n)\big]. \tag{3.11}$$

EXAMPLE (*Estimation of the correlation function and spectral measure*). Suppose, we observe, on a time interval $0 \leqslant t \leqslant T$, an ergodic stationary process $\xi(t)$, $-\infty < t < \infty$, with zero mean and unknown correlation function

$$B(t) = \mathbf{E}\xi(t + s)\xi(s) = \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda t}\, \mathrm{d}F(\lambda), \quad -\infty < t < \infty.$$

One can apply

$$\widehat{B}(t) = \frac{1}{T} \int_{-\infty}^{\infty} \widehat{\xi}(t + s)\widehat{\xi}(s)\, \mathrm{d}s$$

as an estimate of the correlation function, where $\widehat{\xi}(t) = \xi(t)$ for $0 \leqslant t \leqslant T$, $\widehat{\xi}(t) = 0$ otherwise. According to the general ergodic property (3.11), for any $-\infty < t < \infty$

$$\begin{aligned} \lim_{T\to\infty} \widehat{B}(t) &= \lim_{T\to\infty} \frac{1}{T} \int_0^T \xi(t + s)\overline{\xi(s)}\, \mathrm{d}s \\ &= \mathbf{E}\xi(t)\overline{\xi(0)} = B(t). \end{aligned} \tag{3.12}$$

Moreover, for any $\varphi \in C_0^\infty$

$$\lim_{T\to\infty} \int_{-\infty}^{\infty} c(t)\widehat{B}(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} c(t)B(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} \varphi(\lambda)\,\mathrm{d}F(\lambda), \tag{3.13}$$

where

$$c(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda t}\varphi(\lambda)\,\mathrm{d}\lambda, \quad -\infty < t < \infty,$$

is the Fourier transform.

3.3. STOCHASTIC INTEGRAL REPRESENTATION OF A PROCESS WITH INDEPENDENT INCREMENTS

Examples of this type of random processes are the Poisson process and the Brownian motion. Let us forget the Brownian motion for a while, and suppose that we have at our disposal a family of *independent* Poisson processes $\eta(t, \Delta x)$, $t \geqslant 0$, indexed by some (intervals) Δx, $-\infty < x < \infty$, and corresponding to parameters $\lambda = \Delta F(x)$; recall that

$$\mathbf{E}\eta(t, \Delta x) = t \cdot \Delta F(x), \qquad \mathbf{D}\eta(t, \Delta x) = t \cdot \Delta F(x).$$

Set

$$\xi(t) = \sum_x x\eta(t, \Delta x), \quad t \geqslant 0,$$

where the sum is taken over a finite number of x's, $-\infty < x < \infty$. Then obviously $\xi(t)$ is a random process with independent increments $\xi(t) - \xi(s)$ on *disjoint* intervals $(s, t]$, as this is true for the Poisson processes; moreover,

$$\mathbf{E}\eta(t) = t\sum_x x\Delta F(x), \qquad \mathbf{D}\eta(t) = t\sum_x x^2\Delta F(x).$$

Taking in account the structure of the trajectories of a Poisson process, we see that $\xi(t)$, $t \geqslant 0$, is a piecewise constant function having a jump of the size x at the

moment τ when the corresponding Poisson process $\eta(t, \Delta x)$, $t \geqslant 0$, has a (unit) jump (see p. 92). If we consider the jump moments $\tau_0 < \tau_1 < \ldots$ alone, they appear exactly as in the *Poisson process*

$$\eta(t) = \sum_x \eta(t, \Delta x), \quad t \geqslant 0,$$

with the corresponding parameter

$$\lambda = \sum_x \Delta F(x)$$

(one can easily verify that a sum of *independent* Poisson processes is again a Poisson process, e.g. by applying characteristic functions (p. 73)). Although all jumps of the Poisson process $\eta(t)$, $t \geqslant 0$, are of fixed size (equal to 1), the jumps of the process $\xi(t)$, $t \geqslant 0$, have different sizes x, depending on which Poisson component $\eta(t, \Delta x)$, $t \geqslant 0$, jumps at this moment.

This preliminary discussion suggest a general construction of the process $\xi(t)$, $t \geqslant 0$, having jumps of any size x, $-\infty < x < \infty$, whose intensity is characterized by the corresponding measure $\mathrm{d}F(x)$.

Let us introduce a *Poisson stochastic measure* $\mathrm{d}\eta$ on

$$\mathbb{R}_+^2 = \{0 \leqslant t < \infty, \ -\infty < x < \infty\},$$

whose values $\Delta\eta$ are *independent* on *disjoint* sets $\Delta \subseteq \mathbb{R}_+^2$ and distributed according to the Poisson law with parameter $\Delta\lambda$ given by the corresponding values of the product measure

$$\mathrm{d}\lambda = \mathrm{d}t \times \mathrm{d}F(x)$$

on $\mathbb{R}_+^2$. We assume that

$$\int_{-\infty}^{\infty} x^2 \mathrm{d}F(x) < \infty, \tag{3.14}$$

so that $F(\mathrm{d}x)$ can be unbounded near the point $x = 0$. Introduce the stochastic measure

$$\mathrm{d}\eta_0 = \mathrm{d}\eta - \mathbf{E}\mathrm{d}\eta = \mathrm{d}\eta - \mathrm{d}\lambda$$

on $\mathbb{R}_+^2$, of the structure

$$\mathbf{E}\mathrm{d}\eta_0 = 0, \qquad \mathbf{E}\left|\mathrm{d}\eta_0\right|^2 = \mathrm{d}\lambda.$$

With condition (3.14), we can define

$$\xi_0(t) = \int_0^t \int_{-\infty}^{\infty} x\mathrm{d}\eta_0, \quad t > 0. \tag{3.15}$$

It is clear that $\xi_0(t)$, $t \geqslant 0$, is a random process *with independent increments*, since $\mathrm{d}\eta_0$ together with $\mathrm{d}\eta$ are stochastic measures with *independent* values on any disjoint time intervals $(s, t]$, in particular.

To analyse the stochastic integral representation (3.15), put

$$\eta(t, \Delta x) = \int_0^t \int_{\Delta} \mathrm{d}\eta, \quad t \geqslant 0,$$

and

$$\begin{aligned}\eta_0(t, \Delta x) = \int_0^t \int_{\Delta} \mathrm{d}\eta_0 &= \eta(t, \Delta x) - \mathbf{E}\eta(t, \Delta x) \\ &= \eta(t, \Delta x) - t \cdot \Delta F(x), \quad t \geqslant 0,\end{aligned}$$

where $\Delta \subseteq \mathbb{R}$ is an interval $-\infty < x < \infty$. For any fixed $t \geqslant 0$, $\eta(t, \Delta x)$ and $\eta_0(t, \Delta x)$ define stochastic measures on $\mathbb{R}$ which we denote by $\eta(t, \mathrm{d}x)$ and $\eta_0(t, \mathrm{d}x)$, respectively. Obviously,

$$\xi_0(t) = \int_{-\infty}^{\infty} x\, \mathrm{d}\eta_0(t, \mathrm{d}x), \quad t \geqslant 0. \tag{3.16}$$

Suppose, the measure $F(\mathrm{d}x)$, $-\infty < x < \infty$, is *finite*. Then, with condition (3.14), we can define

$$\lambda = \int_{-\infty}^{\infty} xF(\mathrm{d}x)$$

and represent the random process (3.16) as

$$\xi_0(t) = \xi(t) - \mathbf{E}\xi(t) = \xi(t) - \lambda t,$$

where

$$\xi(t) = \int_{-\infty}^{\infty} x\eta(t, \mathrm{d}x), \quad t \geqslant 0, \tag{3.17}$$

and

$$\xi(t) = \int_{-\infty}^{\infty} x\eta(t, \mathrm{d}x) = \lim_{n\to\infty} \sum_{x} x\eta(t, \Delta x)$$

is the limit of the corresponding approximating sums

$$\sum_{x} x\eta(t, \Delta x), \quad t \geqslant 0.$$

Note that any such sum, as a random process in $t \geqslant 0$, is exactly of the type discussed in the beginning.

The above characterization of the process $\xi_0(t)$, $t \geqslant 0$, as the limit of the corresponding 'jump-type' processes, can be extended to the general case, by putting

$$\xi_0(t) = \lim_{\varepsilon\to 0} \xi_\varepsilon(t),$$

where, for any $\varepsilon > 0$,

$$\xi_\varepsilon(t) = \int_{|x|\geqslant\varepsilon} x \, \mathrm{d}\eta_0(t, \mathrm{d}x), \quad t \geqslant 0,$$

is a process of the type (3.16), corresponding to the finite measure

$$\mathrm{d}F_\varepsilon(x) = \begin{cases} \mathrm{d}F(x), & |x| \geqslant \varepsilon, \\ 0, & |x| < \varepsilon. \end{cases}$$

Consequently, the process $\xi_0(t)$, $t \geqslant 0$, of (3.16) can be approximated by the sum of a jump-type process and a linear drift with the velocity

$$\lambda_\varepsilon = \int_{|x|\geqslant\varepsilon} x \, \mathrm{d}F(x),$$

which guarantees that the mean value of the process is zero. □

Let us return now to *continuous* processes with independent increments, which can be represented by the Brownian motion $\beta(t)$, $t \geqslant 0$, having zero mean and the diffusion cooefficient $\sigma^2 > 0$. It is clear that if $\beta(t)$ and $\xi_0(t)$ of (3.16) are independent and αt, $t \geqslant 0$, is a deterministic linear drift, then

$$\xi(t) = \alpha t + \beta(t) + \int_{-\infty}^{\infty} x\eta_0(t, \mathrm{d}x), \quad t \geqslant 0, \tag{3.18}$$

represent a certain class of *random processes with independent increments.*

Moreover, for any $0 < t_1 < t_2 < \ldots < t_n$ the (independent) increments

$$\xi(t_1) - \xi(0),\ \xi(t_2) - \xi(t_1), \ldots,\ \xi(t_n) - \xi(t_{n-1})$$

have exactly the same (joint) probability distribution as

$$\xi(t_1 + t) - \xi(t),\ \xi(t_2 + t) - \xi(t_1 + t), \ldots, \xi(t_n + 1) - \xi(t_{n-1} + t)$$

for any time shift $t \geqslant 0$; in short, the representation (3.18) gives us the process $\xi(t)$, $t \geqslant 0$, *with stationary increments.* □

What can be said about the probability distribution of increments of the process (3.18)? As $\xi(0) = 0$, it suffices to consider

$$\xi(t) = \xi(t) - \xi(0)$$

itself, thanks to the fact that the increments are *stationary*. We apply the method of characteristic functions (see pp. 72–82).

Consider first the jump-type component (3.17) alone, corresponding to a *finite* measure $\mathrm{d}F(x)$, $-\infty < x < \infty$. Let

$$\xi_n(t) = \sum x\eta(t, \Delta x), \quad n = 1, 2, \ldots,$$

be an approximating sequence of finite sums, with independent Poisson components $\eta(t, \Delta x)$,

$$\mathbf{E}\eta(t, \Delta x) = t \cdot \Delta F(x).$$

The characteristic function

$$f_n(u) = \mathbf{E}\mathrm{e}^{iu\xi_n(t)}, \quad -\infty < u < \infty,$$

is the product of the characteristic functions of the *independent* random variables $x\eta(t, \Delta x)$, which yields

$$\ln f_n(u) = t \sum_x \left(e^{iux} - 1\right) \Delta F(x).$$

Passing to the limit as $n \to \infty$ on the right side, we obtain

$$\ln f(u) = t \int_{-\infty}^{\infty} \left(e^{iux} - 1\right) dF(x), \quad -\infty < u < \infty,$$

with

$$f(u) = \lim_{n \to \infty} f_n(u)$$

being the limit characteristic function of the limit random variable

$$\xi(t) = \lim_{n \to \infty} \xi_n(t)$$

of (3.17). For its extension (3.18), with the deterministic component αt and a normally distributed $\beta(t)$, we obtain

$$\ln f(u) = t\left\{ i\alpha u - \frac{1}{2}\sigma^2 u + \int_{-\infty}^{\infty} \left(e^{iux} - 1 - iux\right) dF(x) \right\}, \tag{3.19}$$

$$-\infty < u < \infty.$$

Finally, we can consider the general scheme (3.18), with the 'jump-type' component (3.16); namely, using the finite approximation $dF_\varepsilon(x)$ of the measure $dF(x)$ as above, we get the same formula (3.19) as the limit of

$$\ln f_\varepsilon(t) = t\left\{ iau - \frac{1}{2}\sigma^2 u + \int_{|x| \geqslant \varepsilon} \left(e^{iux} - 1 - iux\right) dF(x), \quad -\infty < u < \infty, \right.$$

when $\varepsilon \to 0$, thanks to condition (3.14) which guarantees the existence of an integrable majorant for

$$e^{iux} - 1 - iux, \quad -\infty < u < \infty.$$

Let us formulate our result as follows.

THEOREM. *The stochastic integral representation* (3.16)–(3.18) *defines a random processes with independent stationary increments, the characteristic function of which is given by* (3.19).

4. Stochastic Differential Equations

4.1. STOCHASTIC DIFFERENTIALS

The formal expression

$$d\xi(t) = \alpha(t)\,dt + \beta(t)\,d\eta(t), \quad t > t_0,$$

called a *stochastic differential*, is just equivalent to the stochastic integral representation

$$\xi(t) - \xi(t_0) = \int_{t_0}^{t} \alpha(s)\,ds + \int_{0}^{t} \beta(s)\,d\eta(s), \quad t \geqslant t_0 \tag{4.1}$$

(in the $\mathcal{L}_2$-space, say). As we'll see below, equation (4.1) itself provides a certain characterization of the process.

For example, does the square $\xi(t) = \eta(t)^2$ of the Brownian motion $\eta(t)$, $t \geqslant t_0$, admit a stochastic differential? Actually, it does, namely

$$d\xi(t) = dt + 2\eta(t)\,d\eta(t)$$

(see p. 220), and this simple example shows that the problem of finding the stochastic differential $d\xi(t)$ for a given random process is not trivial. □

Let us consider a more general example. Suppose

$$\xi(t) = \int_{t_0}^{t} c(t, s)\,d\eta(s), \quad t \geqslant t_0,$$

where $c(t, s)$ is a deterministic function of $t \geqslant s \geqslant t_0$ having the continuous derivative $\frac{d}{dt}\,c(t, s)$. By interchanging the order of integration, we obtain

$$\int_{t_0}^{t} \left[\int_{t_0}^{u} \frac{d}{du} c(u, s)\,d\eta(s) \right] du = \int_{t_0}^{t} \left[\int_{s}^{t} \frac{d}{du} c(u, s)\,du \right] d\eta(s)$$

$$= \int_{t_0}^{t} \left[c(t, s) - c(s, s) \right] d\eta(s) = \xi(t) - \int_{t_0}^{t} c(s, s)\,d\eta(s).$$

Therefore,

$$d\xi(t) = \left[\int_{t_0}^{t} \frac{d}{dt} c(t, s)\,d\eta(s) \right] dt + c(t, t)\,d\eta(t). \tag{4.2}$$

We conclude this preliminary discussion with the suggestion to verify that a random process $\xi(t)$, $t \geqslant t_0$, admits the stochastic differential

$$d\xi(t) = \alpha(t)\,dt$$

with a *continuous* (in the square mean) $\alpha(t)$, $t \geqslant t_0$, if and only if it is *continuously* differentiable (in the square mean) and

$$\xi'(t) = \alpha(t) \tag{4.3}$$

(recall that we consider random functions with values in the $\mathcal{L}_2$-space only).

4.2. LINEAR STOCHASTIC DIFFERENTIAL EQUATIONS

We have in mind the system of stochastic differential equations

$$d\xi(t) = \xi'(t)\,dt, \ldots, d\xi^{(n-2)}(t) = \xi^{(n-1)}(t)\,dt,$$
$$d\xi^{(n-1)}(t) = \left[a_1(t)\xi^{(n-1)}(t) + \cdots + a_n(t)\xi(t)\right] dt + d\eta(t),$$

for a random process $\xi(t)$, $t \geqslant t_0$, and its square mean derivatives $\xi^{(k)}(t)$, $k \leqslant n-1$, which will be written in short as

$$d\xi^{(n-1)} - a_1\xi^{(n-1)}\,dt - \cdots - a_n\xi\,dt = d\eta, \quad t > t_0. \tag{4.4}$$

Here, the coefficients $a_k = a_k(t)$, $k = 1, \ldots, n$, are assumed to be continuous deterministic functions of $t \geqslant t_0$.

In the case $d\eta = 0$ the above equation becomes the ordinary homogeneous differential equation

$$\xi^{(n)} - a_1\xi^{(n-1)} - \cdots - a_n\xi = 0,$$

which has a unique solution

$$\xi(t) = \sum_{k=1}^{n-1} \xi_k w_k(t, t_0) \tag{4.5}$$

for any given

$$\xi(t_0) = \xi_0, \ldots, \xi^{(n-1)}(t_0) = \xi_{n-1}. \tag{4.6}$$

Here, $w_j(t, t_0)$ are the corresponding deterministic solutions with

$$w_k^{(j)}(t_0, t_0) = \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases} \quad j, k = 0, \ldots, n-1.$$

Obviously, for any given initial conditions (4.6), the solution of (4.4) is unique, since the difference of any two solutions satisfies the homogeneous equation with the *zero* initial condition. It is also clear that we have to look for the solution of (4.4) with the zero initial condition only, since the general solution is the sum of this particular solution and the solution of the homogeneous equation with a given initial data (4.6).

THEOREM. *The solution of the stochastic differential equation* (4.4) *with the zero initial conditions* (4.6) *is given by the formula*

$$\xi(t) = \int_{t_0}^{t} w(t, s)\, \mathrm{d}\eta(s), \quad t \geqslant t_0, \tag{4.7}$$

where the kernel $w(t, s)$, *as a function of* $t \geqslant s$, *satisfies the homogeneous differential equation*

$$w^{(n)} - a_1 w^{(n-1)} - \cdots - a_n w = 0, \quad t > s, \tag{4.8}$$

with the initial conditions

$$w(s, s) = 0, \ldots, w^{(n-2)}(s, s) = 0, \; w^{(n-1)}(s, s) = 1.$$

Proof. According to the general formula (4.2), the random function (4.7) has the stochastic differential of the form

$$\mathrm{d}\xi(t) = \left[\int_{t_0}^{t} w^{(1)}(t, s)\, \mathrm{d}\eta(s) \right] \mathrm{d}t + w(t, t)\, \mathrm{d}\eta(t)$$

where $w(t, t) = 0$. Hence, the derivative $\xi^{(1)}(t)$ (in the square mean) exists and is given by

$$\xi^{(1)}(t) = \int_{t_0}^{t} w^{(1)}(t, s)\, \mathrm{d}\eta(s), \quad t \geqslant t_0.$$

The existence of all $(n-1)$ derivatives

$$\xi^{(k)}(t) = \int_{t_0}^{t} w^{(k)}(t,s)\,\mathrm{d}\eta(s), \quad t \geqslant t_0, \quad k \leqslant n-1, \tag{4.9}$$

can be shown analogously. Using the general formula (4.2) for the $(n-1)$th derivative, we obtain

$$\mathrm{d}\xi^{(n-1)}(t) = \left[\int_{t_0}^{t} w^{(n)}(t,s)\,\mathrm{d}\eta(s)\right]\mathrm{d}t + w^{(n-1)}(t,t)\,\mathrm{d}\eta(t),$$

where

$$w^{(n)}(t,s) = a_1(t)w^{(n-1)}(t,s) + \cdots + a_n(t)w(t,s), \quad t > s,$$

and $w^{(n-1)}(t,t) = 1$. Together with (4.9), this proves equation (4.4) for the derivatives $\xi^{(k)}(t)$ and the theorem as well. □

Now, we can characterize the behaviour of a solution $\xi(t)$, $k \geqslant t_0$, of the linear stochastic differential equation (4.4), as follows: For $t \geqslant s$, and without the stochastic disturbance $\mathrm{d}\eta(t)$ on the right side of (4.4), the trajectory of the process is

$$x(t) = \sum_{k=0}^{n-1} \xi^{(k)}(s)w_k(t,s), \quad t \geqslant s,$$

which is a deterministic function except that it depends on the initial random variables $\xi^{(k)}(s)$, $k = 0,\ldots,n-1$; $w_k(t,s)$, $t > s$, is the solution of the ordinary homogeneous differential equation (4.8) with initial conditions

$$w_k^{(j)}(s,s) = \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases} \quad k,j = 0,\ldots,n-1.$$

If, for $t \geqslant s$, the term $\mathrm{d}\eta(t)$ is present on the right side of (4.4), the deviation of the process from the trajectory $x(t)$ is given by

$$\xi(t) - x(t) = \int_{s}^{t} w(t,u)\,\mathrm{d}\eta(u) - \sum_{k=0}^{n-1} \xi^{(k)}(s)w_k(t,s), \quad t \geqslant s, \tag{4.10}$$

which is *uncorrelated* with $x(t)$, since $d\eta(t)$, $t > s$, is *uncorrelated* with $\xi^{(k)}(s)$, $k = 0, \ldots, n-1$, given by

$$\xi^{(k)}(s) = \int_{t_0}^{s} w^{(k)}(s,u)\, d\eta(u),$$

see (4.9). □

Let us consider in more detail the first order differential equation

$$d\xi(t) = a(t)\xi(t)\, dt + d\eta(t), \quad t \geqslant t_0, \tag{4.11}$$

with the initial condition $\xi(t_0) = 0$, say. We assume that

$$\mathbf{E}d\eta(t) = 0, \qquad \mathbf{E}|d\eta(t)|^2 = \sigma^2\, dt.$$

The general formula (4.10) gives, in particular, the correlation function as

$$B(t,s) = \mathbf{E}\xi(t)\xi(s) = B(s,s)w(t,s), \quad t \geqslant s,$$

where

$$\frac{d}{dt} w(t,s) = a(t)w(t,s), \quad t > s,$$
$$w(s,s) = 1.$$

Hence $B(t,s)$, $t \geqslant s$, is the unique solution of the homogeneous differential equation

$$\frac{d}{dt} B(t,s) = a(t)B(t,s), \quad t > s, \tag{4.12}$$

with a given $B(s,s)$ at $t = s$. According to the stochastic integral representation (4.7), the variance

$$B(t,t) = \mathbf{E}\xi(t)^2 = \sigma^2 \int_{t_0}^{t} w(t,s)^2\, ds, \quad t \geqslant t_0,$$

satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} B(t,t) = 2a(t)B(t,t) + \sigma^2, \quad t \geqslant t_0. \tag{4.13}$$

Indeed,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} B(t,t) &= \sigma^2 w(t,t)^2 + 2\sigma^2 \int_{t_0}^{t} w(t,s)\left[\frac{\mathrm{d}}{\mathrm{d}t} w(t,s)\right] \mathrm{d}s \\ &= \sigma^2 + 2\sigma^2 a(t) \int_{t_0}^{t} w(t,s)^2 \, \mathrm{d}s \\ &= \sigma^2 + 2a(t)B(t,t).\end{aligned}$$

4.3. LINEAR DIFFERENTIAL EQUATIONS WITH CONSTANT COEFFICIENTS

Let us consider the stochastic differential equation (4.4) with constant coefficients. Suppose that all roots of the characteristic polynomial

$$P(z) = z^n - a_1 z^{n-1} - \cdots - a_{n-1} z - a_n$$

lie in the left half-plane $\mathrm{Re}(z) < 0$ of the complex parameter z; then, the corresponding equation (4.4) will be called stable. The kernel

$$w(t,s) = w(t-s), \quad t \geqslant s,$$

in (4.7) can be obtained by solving the differential equation

$$w^{(n)}(t) - a_1 w^{(n-1)}(t) - \cdots - a_n w(t) = 0$$

with the initial conditions

$$w(0) = 0, \ldots, w^{(n-2)}(0) = 0, \; w^{(n-1)}(0) = 1$$

(cf. (4.8)). Under the stability condition of the polynomial $P(z)$, the function $w(t)$ exponentially decreases with $t \to \infty$; setting $w(t) = 0$ for $t < 0$, we have

$$\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda t} w(t) \, \mathrm{d}t = \frac{1}{P(i\lambda)}, \quad -\infty < \lambda < \infty.$$

This formula can be easily obtained by partial integration of

$$\int_0^\infty e^{-i\lambda t}\left[w^{(n)}(t) - a_1 w^{(n-1)}(t) - \cdots - a_n w(t)\right] dt = 0.$$

Consequently,

$$w(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\lambda t}\frac{1}{P(i\lambda)}\,d\lambda, \quad -\infty < t < \infty. \tag{4.14}$$

Let us consider the random process

$$\xi(t) = \int_{t_0}^{t} w(t-s)\eta(ds), \quad t \geqslant t_0,$$

of (4.7), with the stochastic measure of the following structure:

$$\mathbf{E}\,d\eta(t) = 0, \qquad \mathbf{E}|\eta(dt)|^2 = \sigma^2\,dt. \tag{4.15}$$

Let us discuss long-time behaviour of the process $\xi(t)$ when $t - t_0 \to \infty$. Formally, it is convenient to assume that $t_0 \to -\infty$ (we suppose that the stochastic measure $\eta(dt)$ is defined on the whole real axis $-\infty < t < \infty$). Putting $t_0 = -\infty$, we obtain the random processes

$$\xi^*(t) = \int_{-\infty}^{\infty} w(t-s)\eta(ds) = \int_{-\infty}^{t} w(t-s)\eta(ds), \quad -\infty < t < \infty, \tag{4.16}$$

with $\mathbf{E}\xi^*(t) = 0$ and the *correlation function*

$$\begin{aligned}\mathbf{E}\xi^*(t)\overline{\xi^*(s)} &= \sigma^2\int_{-\infty}^{\infty} w(t-u)\overline{w(s-u)}\,du \\ &= \sigma^2\int_{-\infty}^{\infty} w(t-s+u)\overline{w(u)}\,du = B(t-s), \quad -\infty < s,t < \infty,\end{aligned}$$

which depends only on the difference $t-s$; i.e., $\xi^*(t)$ is *stationary* in the wide sense. Comparing $\xi(t)$ and $\xi^*(t)$, we easily obtain

$$\begin{aligned}\mathbf{E}\left|\xi(t) - \xi^*(t)\right|^2 &= \mathbf{E}\left|\int_{-\infty}^{t_0} w(t-s)\eta(ds)\right|^2 \\ &= \sigma^2\int_{-\infty}^{t_0}\left|w(t-s)\right|^2 ds = \sigma^2\int_{t-t_0}^{\infty}\left|w(u)\right|^2 du \longrightarrow 0\end{aligned}$$

as $t - t_0 \to \infty$, i.e.,

$$\xi(t) \to \xi^*(t), \quad t_0 \to -\infty,$$

in the square mean. Of course, the same result is true for a *general* solution $\xi(t)$, $t \geqslant t_0$, of the stochastic differential equation (4.4), since for any $k = 0, \ldots, n-1$

$$w_k(t, t_0) = w_k(t - t_0) \longrightarrow 0, \quad t_0 \to -\infty,$$

in the representation (4.5) of a *general* solution of the homogeneous equation. Thus, we get the following result.

THEOREM. *A general solution* $\xi(t)$, $t \geqslant t_0$, *of the stable stochastic differential equation* (4.4) *converges as* $t_0 \to -\infty$ *to the stationary process* $\xi^*(t)$, $-\infty < t < \infty$, *of* (4.16):

$$\xi(t) \to \xi^*(t), \quad -\infty < t < \infty.$$

Set

$$f(\lambda) = \frac{\sigma^2}{2\pi |P(i\lambda)|^2}, \quad -\infty < \lambda < \infty. \tag{4.17}$$

We can write the correlation function of the stationary process (4.16) in the following form:

$$\begin{aligned} B(t) &= \mathbf{E}\xi^*(t+s)\overline{\xi^*(s)} \\ &= \sigma^2 \int_{-\infty}^{\infty} w(t+s)\overline{w(s)}\,\mathrm{d}s = \int_{-\infty}^{\infty} \mathrm{e}^{i\lambda t} f(\lambda)\,\mathrm{d}\lambda, \quad -\infty < t < \infty. \end{aligned}$$

Hence

$$F(\mathrm{d}\lambda) = f(\lambda)\,\mathrm{d}\lambda$$

is the *spectral measure* of $\xi^*(t)$ ($f(\lambda)$ itself is called the *spectral density*). Indeed, by taking the inverse Fourier transform, one has

$$\begin{aligned} \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda t} B(t)\,\mathrm{d}t &= \frac{\sigma^2}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda t}\left[\int_{-\infty}^{\infty} w(t+s)\overline{w(s)}\,\mathrm{d}s\right]\mathrm{d}t \\ &= \frac{\sigma^2}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{i\lambda s}\left[\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda(t+s)} w(t+s)\,\mathrm{d}t\right]\overline{w(s)}\,\mathrm{d}s \\ &= \frac{\sigma^2}{2\pi}\left[\frac{1}{P(i\lambda)}\right]\overline{\int_{-\infty}^{\infty} \mathrm{e}^{-i\lambda s} w(s)\,\mathrm{d}s} = \frac{\sigma^2}{2\pi}\frac{1}{|P(i\lambda)|^2}. \end{aligned}$$

EXAMPLE (*Stochastic oscillations of a heavy pendulum*). The motion of a free pendulum is described by the second order differential equation

$$w''(t) + 2hw'(t) + a^2 w(t) = 0,$$

where $h > 0$ is a *small* parameter characterizing friction, while $a^2 > 0$ characterizes the frequency λ_0, $\lambda_0^2 = a^2 - h^2$, of the damped oscillations of the heavy pendulum. Let us imagine that the pendulum is a part of some ship equipment, and is subject to *high* frequency chaotic (*random*) oscillations caused by the rough sea.

The corresponding motion of the pendulum *forced* by a stochastic term $d\eta(t)$ with the structure (4.15) can be described by a *stationary* random process $\xi^*(t)$ with the *spectral density*

$$f(\lambda) = \frac{\sigma^2}{2\pi[(\lambda^2 - \lambda_0^2)^2 + 4h^2\lambda^2]}$$

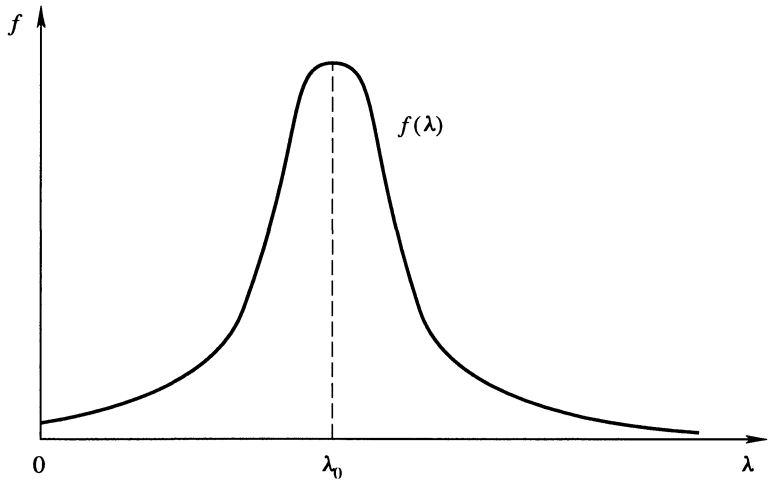

Fig. 19.

obtained from the general formula (4.17). The spectral density $f(\lambda)$, $0 \leqslant \lambda < \infty$, is concentrated near the point $\lambda = \lambda_0$ (see Figure 19), which shows that the most powerful harmonics of the *forced* motion correspond to frequencies λ close to the frequency λ_0 of the *free* oscillations:

$$\lambda \approx \lambda_0$$

(see p. 225 for the general spectral representation of a stationary process). This result is quite different from what we know about *forced* oscillations in the *deterministic* theory.

4.4. THE KALMAN–BUCY FILTER

We consider a random process $\xi(t)$, $t \geqslant t_0$, satisfying the stochastic differential equation

$$\mathrm{d}\xi(t) = \theta(t)\,\mathrm{d}t + \mathrm{d}\eta(t) \tag{4.18}$$

and the initial condition $\xi(t_0) = 0$, where $\theta(t)$ is a square-mean continuous random function. We interpret $\theta(t)$, $t \geqslant t_0$, as a 'signal', which is observed in the additive (random) noise given by a stochastic measure $\mathrm{d}\eta(t)$ of the structure

$$\mathbf{E}\,\mathrm{d}\eta(t) = 0, \qquad \mathbf{E}|\mathrm{d}\eta|^2 = \mathrm{d}t.$$

We suppose that $\mathrm{d}\eta(t)$, $t \geqslant t_0$, does not depend on $\theta(t)$, $t \geqslant t_0$ (or is just *uncorrelated* with it).

The problem is to estimate $\theta(t)$ given $\xi(s)$, $t_0 \leqslant s \leqslant t$. Below, we discuss the fundamental result due to Kalman and Bucy, which gives the solution to the above problem for random functions $\theta(t)$, $t \geqslant t_0$, satisfying the linear stochastic differential equation

$$\mathrm{d}\theta(t) = a(t)\theta(t)\,\mathrm{d}t + \mathrm{d}\eta_0(t) \tag{4.19}$$

with the initial condition $\theta(t_0) = 0$, where $a(t)$ is a nonrandom continuous function and $\mathrm{d}\eta_0(t)$, $t \geqslant t_0$, is a stochastic measure of the same structure as $\mathrm{d}\eta$ and independent of $\mathrm{d}\eta(t)$, $t \geqslant t_0$.

We shall consider linear estimators of $\theta(t)$ given by linear combinations of the 'observed' variables $\xi(s)$, $t_0 \leqslant s \leqslant t$, or by their limits in the square mean.

Obviously, each linear combination of the variables $\xi(s)$, $t_0 \leqslant s \leqslant t$ (for example, a linear combination of the variables $\xi(t_0), \xi(t_1), \ldots, \xi(t_n)$, where $t_0 < t_1 < \cdots < t_n$) can be written as the stochastic integral

$$\eta = \sum_{k=1}^{n} c_k\left[\xi(t_k) - \xi(t_{k-1})\right] = \int_{t_0}^{t} c(s)\,\mathrm{d}\xi(s)$$

with the corresponding piecewise constant function

$$c(s) = c_k, \quad t_{k-1} < s \leqslant t_k.$$

The limits of such linear combinations are the estimators which can be represented as stochastic integrals

$$\eta = \int_{t_0}^{t} c(s)\,\mathrm{d}\xi(s) = \int_{t_0}^{t} c(s)\theta(s)\,\mathrm{d}s + \int_{t_0}^{t} c(s)\,\mathrm{d}\eta(s), \tag{4.20}$$

where $c(s)$, $t_0 \leqslant s \leqslant t$, is an arbitrary function such that the last two integrals exist. Their sum will serve us as the definition of the stochastic integral with respect to $\mathrm{d}\xi(t)$, which we shall use from now on. (For example, the linear estimator (4.20) is well-defined for any continuous function $c(s)$, $t_0 \leqslant s \leqslant t$.)

Note that

$$\eta_1 = \int_{t_0}^{t} c_1(s)\theta(s)\,\mathrm{d}s, \qquad \eta_2 = \int_{t_0}^{t} c_2(s)\,\mathrm{d}\eta(s)$$

are uncorrelated (orthogonal) random variables in the $\mathcal{L}_2$-space and

$$\|\eta_2\|^2 = \mathbf{E}|\eta_2|^2 = \int_{t_0}^{t} |c_2(s)|^2\,\mathrm{d}s.$$

Among all estimators (4.20), we shall look for the optimal one

$$\widehat{\theta}(t) = \int_{t_0}^{t} c(t,s)\,\mathrm{d}\xi(s) \tag{4.21}$$

corresponding to a weight function $c(s) = c(t,s)$, $t_0 \leqslant s \leqslant t$, by applying the orthogonality condition

$$\mathbf{E}\big[\theta(t) - \widehat{\theta}(t)\big]^2 = 0$$

in the $\mathcal{L}_2$-space, which guarantees

$$\|\theta(t) - \widehat{\theta}(t)\|^2 = \min_{\eta} \|\theta(t) - \eta\|^2.$$

By expressing the orthogonality condition directly in terms of the weight function $c(t,s)$ in (4.21), one obtains the following

LEMMA. *Suppose that the function $c(t, s)$ is continuous in $t \geqslant s \geqslant t_0$ and satisfies the integral equation*

$$c(t, s) = B(t, s) - \int_{t_0}^{t} c(t, u)B(u, s)\,du, \quad t \geqslant s, \tag{4.22}$$

where $B(t, s) = \mathbf{E}\theta(t)\theta(s)$, $t, s \geqslant t_0$, is the correlation function of the random process $\theta(t)$, $t \geqslant t_0$. Then $c(t, s)$ is the weight function of the optimal estimator (4.21), *and*

$$c(t, t) = \mathbf{E}[\theta(t) - \widehat{\theta}(t)]^2 \tag{4.23}$$

is the corresponding quadratic error.

Proof. In fact, (4.22) implies the orthogonality condition, since, for arbitrary variables of the form (4.20), we have that

$$\begin{aligned}
&\mathbf{E}[\theta(t) - \widehat{\theta}(t)]\eta \\
&= \mathbf{E}\left[\theta(t) - \int_{t_0}^{t} c(t, u)\theta(u)\,du - \int_{t_0}^{t} c(t, u)\,d\eta(u)\right] \\
&\quad \times \left[\int_{t_0}^{t} c(s)\theta(s)\,ds + \int_{t_0}^{t} c(s)\,d\eta(s)\right] \\
&= \int_{t_0}^{t} c(s)\left[B(t, s) - \int_{t_0}^{t} c(t, u)B(u, s)\,du - c(t, s)\right] ds = 0.
\end{aligned}$$

Furthermore,

$$\mathbf{E}[\theta(t) - \widehat{\theta}(t)]^2 = \mathbf{E}\theta(t)[\theta(t) - \widehat{\theta}(t)] - \mathbf{E}\widehat{\theta}(t)[\theta(t) - \widehat{\theta}(t)],$$

where

$$\begin{aligned}
&\mathbf{E}\theta(t)[\theta(t) - \widehat{\theta}(t)] \\
&= \mathbf{E}\theta(t)^2 - \mathbf{E}\left[\theta(t) \cdot \int_{t_0}^{t} c(t, u)\theta(u)\,du\right] - \mathbf{E}\left[\theta(t) \cdot \int_{t_0}^{t} c(t, u)\,d\eta(u)\right] \\
&= B(t, t) - \int_{t_0}^{t} c(t, u)B(u, t)\,du = c(t, t)
\end{aligned}$$

according to (4.22) with $s = t$; and

$$\mathbf{E}\widehat{\theta}(t)\big[\theta(t) - \widehat{\theta}(t)\big] = 0$$

by the orthogonality condition with $\eta = \widehat{\theta}(t)$. Hence, the lemma is proved. □

As $\theta(t)$, $t \geqslant t_0$, satisfies the stochastic differential equation (4.19) and the initial condition $\theta(t_0) = 0$, it can be written as

$$\theta(t) = \int_{t_0}^{t} w_0(t, s)\, \mathrm{d}\eta_0(s), \quad t \geqslant t_0, \tag{4.24}$$

where $w_0(t, s)$, $t \geqslant s$, is the soliuton of the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} w_0(t, s) = a(t) w_0(t, s), \quad t > s,$$

with the initial condition $w_0(s, s) = 1$. Moreover, we know that the correlation function $B(t, s) = \mathbf{E}\theta(t)\theta(s)$ satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} B(t, s) = a(t) B(t, s), \quad t > s,$$

see p. 241.

Our aim is to find a function $c(t, s)$ which is continuous together with its derivative $\frac{\mathrm{d}}{\mathrm{d}t} c(t, s)$, for all parameters $t \geqslant s \geqslant t_0$, and which satisfies the integral equation (4.22).

By differentiating equation (4.22) with respect to t, we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}t} c(t, s) = a(t) B(t, s) - c(t, t) B(t, s) - \int_{t_0}^{t} \frac{\mathrm{d}}{\mathrm{d}t} c(t, u) B(u, s)\, \mathrm{d}u.$$

Assume there exists a function $x(t)$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t} c(t, s) = x(t) c(t, s), \quad t > s.$$

Then from (4.22) one has

$$\begin{aligned} x(t) B(t, s) &= x(t) \left[c(t, s) + \int_{t_0}^{t} c(t, u) B(u, s)\, \mathrm{d}u \right] \\ &= \big[a(t) - c(t, t)\big] B(t, s). \end{aligned}$$

Obviously, $x(t) = a(t) - b(t)$, where

$$b(t) = c(t, t).$$

Having obtained this unexpected result, it is natural to look for the function $c(t, s)$, $t \geqslant s$, as the solution of the differential equation of the form

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} c(t, s) = \big[a(t) - b(t)\big] c(t, s), \quad t > s, \\ &c(s, s) = b(s), \quad s \geqslant t_0. \end{aligned} \tag{4.25}$$

□

Take a continuous function $b(t)$, $t \geqslant t_0$, then the solution $c(t, s)$ of the linear differential equation (4.25) and its derivative $\frac{\mathrm{d}}{\mathrm{d}t} c(t, s)$ are jointly continuous in $t \geqslant s \geqslant t_0$. If we take this solution $c(t, s)$ as the weight function in (4.21), then, by expressing the stochastic differential $\mathrm{d}\widehat{\theta}(t)$ in the form (4.2):

$$\mathrm{d}\widehat{\theta}(t) = \left[\int_{t_0}^{t} \frac{\mathrm{d}}{\mathrm{d}t} c(t, s)\, \mathrm{d}\xi(s)\right] \mathrm{d}t + c(t, t)\, \mathrm{d}\xi(t),$$

and using (4.25), we obtain

$$\mathrm{d}\widehat{\theta}(t) = \big[a(t) - b(t)\big]\widehat{\theta}(t)\, \mathrm{d}t + b(t)\big[\widehat{\theta}(t)\, \mathrm{d}t + \mathrm{d}\eta(t)\big].$$

According to (4.19), the difference $\Delta(t) = \theta(t) - \widehat{\theta}(t)$ satisfies the linear stochastic differential equation

$$\begin{aligned} &\mathrm{d}\Delta(t) = \mathrm{d}\theta(t) - \mathrm{d}\widehat{\theta}(t) \\ &\big[a(t) - b(t)\big]\Delta(t)\, \mathrm{d}t + \big[\mathrm{d}\eta_0(t) - b(t)\, \mathrm{d}\eta(t)\big] \end{aligned} \tag{4.26}$$

with the initial condition $\Delta(t_0) = 0$, whose solution can be written as

$$\Delta(t) = \int_{t_0}^{t} w(t, s)\, \mathrm{d}\eta_0(s) - \int_{t_0}^{t} w(t, s) b(s)\, \mathrm{d}\eta(s).$$

Here, the weight function $w(t,s)$, $t \geqslant s$, is the solution of the ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} w(t,s) = \big[a(t) - b(t)\big] w(t,s), \quad t > s,$$
$$w(s,s) = 1,$$

see p. 241. By comparing this equation to (4.25), we see that

$$c(t,s) = w(t,s)b(s), \quad t \geqslant s. \tag{4.27}$$

□

As $c(t,s)$ defines the optimal estimator (4.21), the corresponding function $b(t) = c(t,t)$ is given by

$$b(t) = \mathbf{E}\Delta(t)^2, \quad t \geqslant t_0,$$

see (4.23). From (4.13) and (4.26) it follows that the variance $b(t)$ is the solution of the following *Ricatti equation*

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} b(t) = 2a(t)b(t) - b(t)^2 + 1, \quad t > t_0, \\ &b(t_0) = 0. \end{aligned} \tag{4.28}$$

Let $b(t)$, $t \geqslant t_0$, be the solution of (4.28). From (4.19), (4.26) we obtain for $\widehat{\theta}(t) = \theta(t) - \Delta(t)$ the stochastic differential equation

$$\mathrm{d}\widehat{\theta}(t) = \big[a(t) - b(t)\big]\widehat{\theta}(t)\,\mathrm{d}t + b(t)\,\mathrm{d}\xi(t), \quad t > t_0, \tag{4.29}$$

whose solution, with $\widehat{\theta}(t_0) = 0$, can be written as

$$\begin{aligned} \widehat{\theta}(t) &= \int_{t_0}^{t} c(t,s)\,\mathrm{d}\xi(s) \\ &= \int_{t_0}^{t} c(t,s)\theta(s)\,\mathrm{d}s + \int_{t_0}^{t} c(t,s)\,\mathrm{d}\eta(s). \end{aligned} \tag{4.30}$$

Let us show that the weight function $c(t, s)$ satisfies the integral equation (4.22). By applying (4.24), (4.26) and the differential equation (4.28) for $b(t) = \mathbf{E}\Delta(t)^2$, we easily obtain that the function

$$f(t) = \mathbf{E}\widehat{\theta}(t)\big[\theta(t) - \widehat{\theta}(t)\big] = \mathbf{E}\big[\theta(t) - \Delta(t)\big]\Delta(t)$$
$$= \int_{t_0}^{t} w_0(t, s)w(t, s)\,\mathrm{d}s - b(t)$$

satisfies the homogeneous differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} f(t) = \big[2a(t) - b(t)\big]f(t),$$
$$f(t_0) = 0.$$

Hence $f(t) \equiv 0$. From (4.30) we get

$$b(t) = \mathbf{E}\theta(t)\big[\theta(t) - \widehat{\theta}(t)\big] - \mathbf{E}\widehat{\theta}(t)\big[\theta(t) - \widehat{\theta}(t)\big]$$
$$= \mathbf{E}\theta(t)\big[\theta(t) - \widehat{\theta}(t)\big] = B(t, t) - \int_{t_0}^{t} c(t, s)B(s, t)\,\mathrm{d}s, \quad t \geqslant t_0.$$

Together with (4.25), for the function

$$c_0(t, s) = c(t, s) + \int_{t_0}^{t} c(t, u)B(u, s)\,\mathrm{d}u - B(t, s), \quad t \geqslant s,$$

we obtain the homogeneous equation

$$\frac{\mathrm{d}}{\mathrm{d}t} c_0(t, s) = \big[a(t) - b(t)\big]c_0(t, s), \quad t > s,$$
$$c_0(s, s) = 0.$$

Hence $c_0(t, s) \equiv 0$, which proves (4.22).

The above discussion can be summarized in the following

THEOREM. *The optimal estimator* $\widehat{\theta}(t)$ *of* $\theta(t)$ *is given by the stochastic integral* (4.21), *with the weight function* $c(t, s)$, $t \geqslant s$, *together with the function* $b(t) = \mathbf{E}[\theta(t) - \widehat{\theta}(t)]^2$, *satisfying the system* (4.25), (4.28) *of differential equations. The optimal estimator* $\widehat{\theta}(t)$, $t \geqslant t_0$, *can be obtained by solving the stochastic differential equation* (4.29).

SUBJECT INDEX

Other *Mathematics and Its Applications* titles of interest:

P.M. Alberti and A. Uhlmann: *Stochasticity and Partial Order. Doubly Stochastic Maps and Unitary Mixing.* 1982, 128 pp. ISBN 90-277-1350-2

A.V. Skorohod: *Random Linear Operators.* 1983, 216 pp. ISBN 90-277-1669-2

I.M. Stancu-Minasian: *Stochastic Programming with Multiple Objective Functions.* 1985, 352 pp. ISBN 90-277-1714-1

L. Arnold and P. Kotelenez (eds.): *Stochastic Space-Time Models and Limit Theorems.* 1985, 280 pp. ISBN 90-277-2038-X

Y. Ben-Haim: *The Assay of Spatially Random Material.* 1985, 336 pp. ISBN 90-277-2066-5

A. Pazman: *Foundations of Optimum Experimental Design.* 1986, 248 pp. ISBN 90-277-1865-2

P. Kree and C. Soize: *Mathematics of Random Phenomena. Random Vibrations of Mechanical Structures.* 1986, 456 pp. ISBN 90-277-2355-9

Y. Sakamoto, M. Ishiguro and G. Kitagawa: *Akaike Information Criterion Statistics.* 1986, 312 pp. ISBN 90-277-2253-6

G.J. Szekely: *Paradoxes in Probability Theory and Mathematical Statistics.* 1987, 264 pp. ISBN 90-277-1899-7

O.I. Aven, E.G. Coffman (Jr.) and Y.A. Kogan: *Stochastic Analysis of Computer Storage.* 1987, 264 pp. ISBN 90-277-2515-2

N.N. Vakhania, V.I. Tarieladze and S.A. Chobanyan: *Probability Distributions on Banach Spaces.* 1987, 512 pp. ISBN 90-277-2496-2

A.V. Skorohod: *Stochastic Equations for Complex Systems.* 1987, 196 pp. ISBN 90-277-2408-3

S. Albeverio, Ph. Blanchard, M. Hazewinkel and L. Streit (eds.): *Stochastic Processes in Physics and Engineering.* 1988, 430 pp. ISBN 90-277-2659-0

A. Liemant, K. Matthes and A. Wakolbinger: *Equilibrium Distributions of Branching Processes.* 1988, 240 pp. ISBN 90-277-2774-0

G. Adomian: *Nonlinear Stochastic Systems Theory and Applications to Physics.* 1988, 244 pp. ISBN 90-277-2525-X

J. Stoyanov, O. Mirazchiiski, Z. Ignatov and M. Tanushev: *Exercise Manual in Probability Theory.* 1988, 368 pp. ISBN 90-277-2687-6

E.A. Nadaraya: *Nonparametric Estimation of Probability Densities and Regression Curves.* 1988, 224 pp. ISBN 90-277-2757-0

H. Akaike and T. Nakagawa: *Statistical Analysis and Control of Dynamic Systems.* 1998, 224 pp. ISBN 90-277-2786-4

Other *Mathematics and Its Applications* titles of interest:

A.V. Ivanov and N.N. Leonenko: *Statistical Analysis of Random Fields.* 1989, 256 pp. ISBN 90-277-2800-3

V. Paulauskas and A. Rackauskas: *Approximation Theory in the Central Limit Theorem. Exact Results in Banach Spaces.* 1989, 176 pp. ISBN 90-277-2825-9

R.Sh. Liptser and A.N. Shiryayev: *Theory of Martingales.* 1989, 808 pp. ISBN 0-7923-0395-4

S.M. Ermakov, V.V. Nekrutkin and A.S. Sipin: *Random Processes for Classical Equations of Mathematical Physics.* 1989, 304 pp. ISBN 0-7923-0036-X

G. Constantin and I. Istratescu: *Elements of Probabilistic Analysis and Applications.* 1989, 488 pp. ISBN 90-277-2838-0

S. Albeverio, Ph. Blanchard and D. Testard (eds.): *Stochastics, Algebra and Analysis in Classical and Quantum Dynamics.* 1990, 264 pp. ISBN 0-7923-0637-6

Ya.I. Belopolskaya and Yu.L. Dalecky: *Stochastic Equations and Differential Geometry.* 1990, 288 pp. ISBN 90-277-2807-0

A.V. Gheorghe: *Decision Processes in Dynamic Probabilistic Systems.* 1990, 372 pp. ISBN 0-7923-0544-2

V.L. Girko: *Theory of Random Determinants.* 1990, 702 pp. ISBN 0-7923-0233-8

S. Albeverio, PH. Blanchard and L. Streit: *Stochastic Processes and their Applications in Mathematics and Physics.* 1990, 416 pp. ISBN 0-9023-0894-8

B.L. Rozovskii: *Stochastic Evolution Systems. Linear Theory and Applications to Non-linear Filtering.* 1990, 330 pp. ISBN 0-7923-0037-8

A.D. Wentzell: *Limit Theorems on Large Deviations for Markov Stochastic Process.* 1990, 192 pp. ISBN 0-7923-0143-9

K. Sobczyk: *Stochastic Differential Equations. Applications in Physics, Engineering and Mechanics.* 1991, 410 pp. ISBN 0-7923-0339-3

G. Dallaglio, S. Kotz and G. Salinetti: *Distributions with Given Marginals.* 1991, 300 pp. ISBN 0-7923-1156-6

A.V. Skorohod: *Random Processes with Independent Increments.* 1991, 280 pp. ISBN 0-7923-0340-7

L. Saulis and V.A. Statulevicius: *Limit Theorems for Large Deviations.* 1991, 232 pp. ISBN 0-7923-1475-1

A.N. Shiryaev (ed.): *Selected Works of A.N. Kolmogorov, Vol. 2: Probability Theory and Mathematical Statistics.* 1992, 598 pp. ISBN 90-277-2795-X

Yu.I. Neimark and P.S. Landa: *Stochastic and Chaotic Oscillations.* 1992, 502 pp. ISBN 0-7923-1530-8

Other *Mathematics and Its Applications* titles of interest:

Y. Sakamoto: *Categorical Data Analysis by AIC.* 1992, 260 pp.
ISBN 0-7923-1429-8

Lin Zhengyan and Lu Zhuarong: *Strong Limit Theorems.* 1992, 200 pp.
ISBN 0-7923-1798-0

J. Galambos and I. Katai (eds.): *Probability Theory and Applications.* 1992, 350 pp. ISBN 0-7923-1922-2

N. Bellomo, Z. Brzezniak and L.M. de Socio: *Nonlinear Stochastic Evolution Problems in Applied Sciences.* 1992, 220 pp. ISBN 0-7923-2042-5

A.K. Gupta and T. Varga: *Elliptically Contoured Models in Statistics.* 1993, 328 pp. ISBN 0-7923-2115-4

B.E. Brodsky and B.S. Darkhovsky: *Nonparametric Methods in Change-Point Problems.* 1993, 210 pp. ISBN 0-7923-2122-7

V.G. Voinov and M.S. Nikulin: *Unbiased Estimators and Their Applications. Volume 1: Univariate Case.* 1993, 522 pp. ISBN 0-7923-2382-3

V.S. Koroljuk and Yu.V. Borovskich: *Theory of* U-*Statistics.* 1993, 552 pp.
ISBN 0-7923-2608-3

A.P. Godbole and S.G. Papastavridis (eds.): *Runs and Patterns in Probability: Selected Papers.* 1994, 358 pp. ISBN 0-7923-2834-5

Yu. Kutoyants: *Identification of Dynamical Systems with Small Noise.* 1994, 298 pp. ISBN 0-7923-3053-6

M.A. Lifshits: *Gaussian Random Functions.* 1995, 346 pp. ISBN 0-7923-3385-3

M.M. Rao: *Stochastic Processes: General Theory.* 1995, 635 pp.
ISBN 0-7923-3725-5

Yu.A. Rozanov: *Probability Theory, Random Processes and Mathematical Statistics.* 1995, 267 pp. ISBN 0-7923-3764-6