de Gruyter Studies in Mathematics 23

Editors: Heinz Bauer · Jerry L. Kazdan · Eduard Zehnder

de Gruyter Studies in Mathematics

1 Riemannian Geometry, 2nd rev. ed., *Wilhelm P. A. Klingenberg*
2 Semimartingales, *Michel Métivier*
3 Holomorphic Functions of Several Variables, *Ludger Kaup and Burchard Kaup*
4 Spaces of Measures, *Corneliu Constantinescu*
5 Knots, *Gerhard Burde and Heiner Zieschang*
6 Ergodic Theorems, *Ulrich Krengel*
7 Mathematical Theory of Statistics, *Helmut Strasser*
8 Transformation Groups, *Tammo tom Dieck*
9 Gibbs Measures and Phase Transitions, *Hans-Otto Georgii*
10 Analyticity in Infinite Dimensional Spaces, *Michel Hervé*
11 Elementary Geometry in Hyperbolic Space, *Werner Fenchel*
12 Transcendental Numbers, *Andrei B. Shidlovskii*
13 Ordinary Differential Equations, *Herbert Amann*
14 Dirichlet Forms and Analysis on Wiener Space, *Nicolas Bouleau and Francis Hirsch*
15 Nevanlinna Theory and Complex Differential Equations, *Ilpo Laine*
16 Rational Iteration, *Norbert Steinmetz*
17 Korovkin-type Approximation Theory and its Applications, *Francesco Altomare and Michele Campiti*
18 Quantum Invariants of Knots and 3-Manifolds, *Vladimir G. Turaev*
19 Dirichlet Forms and Symmetric Markov Processes, *Masatoshi Fukushima, Yoichi Oshima, Masayoshi Takeda*
20 Harmonic Analysis of Probability Measures on Hypergroups, *Walter R. Bloom and Herbert Heyer*
21 Potential Theory on Infinite-Dimensional Abelian Groups, *Alexander Bendikov*
22 Methods of Noncommutative Analysis, *Vladimir E. Nazaikinskii, Victor E. Shatalov, Boris Yu. Sternin*

Heinz Bauer

Probability Theory

Translated from the German by Robert B. Burckel

Walter de Gruyter
Berlin · New York 1996

Author
Heinz Bauer
Mathematisches Institut
der Universität Erlangen-Nürnberg
Bismarckstraße 1a
D-91054 Erlangen, FRG

Translator
Robert B. Burckel
Department of Mathematics
Kansas State University
137 Cardwell Hall
Manhattan, Kansas 66506-2602
USA

1991 Mathematics Subject Classification: 60-01

♾ Printed on acid-free paper which falls within the guidelines of the ANSI to ensure permanence and durability.

Library of Congress Cataloging-in-Publication Data

Bauer, Heinz, 1928–
[Wahrscheinlichkeitstheorie. English]
Probability theory / Heinz Bauer ; translated from the German by Robert B. Burckel.
p. cm. – (De Gruyter studies in mathematics ; 23)
Includes bibliographical references and indexes.
ISBN 3-11-013935-9 (alk. paper)
1. Probabilities. I. Title. II. Series.
QA273.B266713 1996
519.2–dc20 95-39173
CIP

Die Deutsche Bibliothek – Cataloging-in-Publication Data

Bauer, Heinz:
Probability theory / Heinz Bauer. Transl. from the German by Robert B. Burckel. – Berlin ; New York : de Gruyter, 1996
(De Gruyter studies in mathematics ; 23)
Dt. Ausg. u.d.T.: Bauer, Heinz: Wahrscheinlichkeitstheorie
ISBN 3-11-013935-9
NE: GT

Printed in Germany.
Typesetting: Oldřich Ulrych, Prague, Czech Republic.
Printing: Gerike GmbH, Berlin.
Binding: Lüderitz & Bauer, Berlin. Cover design: Rudolf Hübler, Berlin.

Preface

"This book is intended to serve as a guide to the student of probability theory." That sentence stands at the beginning of the Preface to my book *Probability Theory and Elements of Measure Theory*, which was published in 1972 (by Holt, Rinehart and Winston, Inc., New York) and in 1981 in a second edition (by Academic Press Inc., London, Ltd.). Now 23 years later a new book is appearing, bearing a new title: *Probability Theory*.

The question naturally arises whether and to what extent the contents and the aims of this new book have changed. The second part of this question is easier to answer: The aim remains the same, to serve as a reliable guide to those studying probabiliy theory. The answer to the first part has several components: On one hand, the first part of the old book devoted to measure theory has been eliminated. Actually it has been extensively re-worked and was published in German (first edition 1990, second 1992), under the title *Mass- und Integrationstheorie*. An English translation is in preparation. All this in response to a wish expressed by many earlier readers. The part of the older book devoted to probability theory has been extensively re-written. A new conception seemed necessary in order to better orient the book toward contemporary developments. An introductury text can no longer claim to bring the reader to the absolute frontiers of research. The pace of the latter in probability theory has been far too rapid in the last two decades. A book like this must, however, open up for the reader the possibility of progressing further with minimal strain into the specialized literature. I kept this requirement constantly in mind while writing the book. The idea of a guide is also to be understood in this sense: The reader should be led to hike through basic terrain along well-secured paths, now and then even scrambling up to a particular prominence in order to get an overview of a region. After this he should be prepared to forge ahead into less-developed parts of the terrain, if need be with special guides or, if research drives him, to penetrate into wholly new territory on his own.

Among the most significant features of the book, a few should be emphasized here: Commensurate with its importance, martingale theory is gone into quite early and deeply. The law of the iterated logarithm is proved in a form which goes back to V. Strassen; this considerably sharpens the classical theorem of Ph. Hartman and A. Wintner. Two long chapters are devoted to the theory of stochastic processes. In particular, Brownian motion – nowadays a fundamental mathematical concept – and the Ornstein-Uhlenbeck process are discussed in great detail. Also the style of exposition has changed: Some redundency was consciously built in here and there. The decisive determinant of the value of a textbook is, however, the reliability and precision of what it promulgates. Reliability includes the demand that the text must be self-contained in the sense that proofs always

be complete: The reader must not be referred to exercises in order to reduce the length of a proof. So the only prerequisite for reading this book is a sufficient knowledge of measure and integration theory. Any standard textbook devoted to this subject will provide the reader with the necessary background and details.

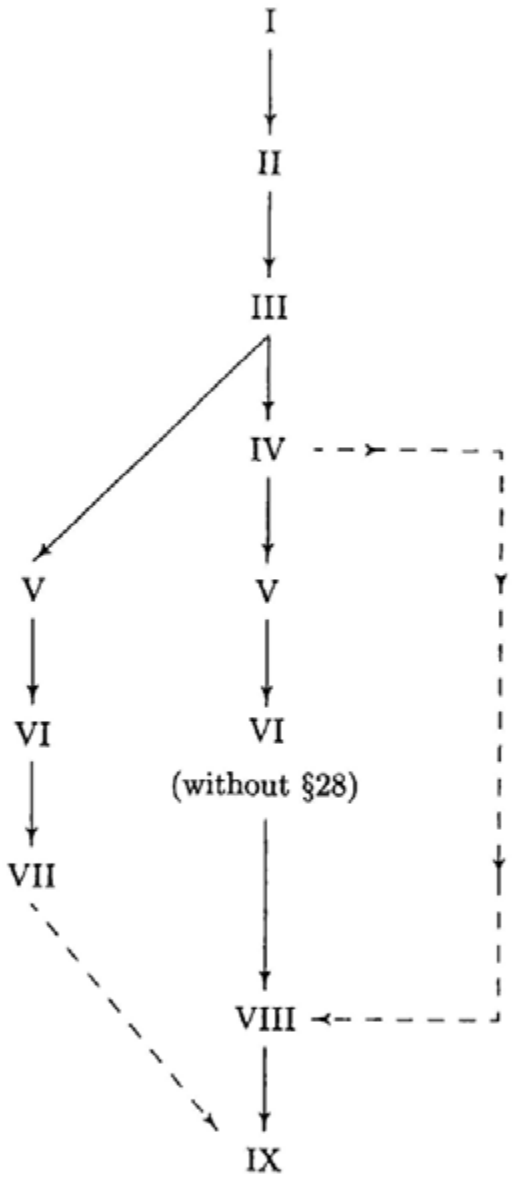
This book is much more than a pure translation of the German original (cf. BAUER [1991] in the Bibliography). It is in fact a revised and improved version of that book. A translator, in the strict sense of the word, could never do this job. This explains why I have to express my deep gratitude to my very special translator, to my American colleague Professor Robert B. Burckel from Kansas State University. He had gotten to know my book by reading its very first German edition. I owe our friendship to his early interest in it. He expended great energy, especially on this new book, using his extensive acquaintance with the literature to make many knowledgeable suggestions, pressing for greater clarity and giving intensive support in bringing this enterprise to a good conclusion.

I thank Dr. Oldřich Ulrych from Charles University, Prague, for his skill and patience in preparing the book manuscript in TEX for final processing. My family deserves thanks for their sacrifices, understanding and consideration. Finally, I thank my publisher Walter de Gruyter & Co. and, above all, Dr. Manfred Karbe for publishing my book both in English and in German.

Erlangen, May 1995 *Heinz Bauer*

Table of Contents

Interdependence of chapters



The sequence I–III, V–VII corresponds to a more classically oriented treatment of the theme "limit behavior of sums of independent random variables".

The dotted-line path from IV indicates that direct access from there to the theory of stochastic processes in chapter VIII (say, up to §43) is possible.

The dotted line from VII indicates that a knowledge of §31 (up to and including Theorem 31.1) from this chapter suffices for chapter IX.

Notation

$A := B$	A is defined by this equation
$\subset, \cup, \cap, \bigcup, \bigcap, \complement, \setminus$	set-theoretic symbols
$A \triangle B$	symmetric difference of sets $[:= (A \setminus B) \cup (B \setminus A)]$
$\mathscr{P}(\Omega)$	power set of a set Ω
$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$	set of all natural, whole, rational, real, and complex numbers, respectively (Note: $0 \notin \mathbb{N}$.)
$\mathbb{C}^*, \mathbb{R}^*$	$\mathbb{C} \setminus \{0\}, \mathbb{R} \setminus \{0\}$
$\mathbb{R}_+, \mathbb{Z}_+, \bar{\mathbb{R}}_+$	non-negative reals, integers, extended reals
$\bar{\mathbb{R}} := [-\infty, +\infty]$	extended number line
real function	function with values in $\mathbb{R}$
numerical function	function with values in $\bar{\mathbb{R}}$
$f \mid A_0$	restriction to $A_0 \subset A$ of the function $f : A \to B$
$(\Omega, \mathscr{A})$	measurable space, i.e., set Ω with σ-algebra $\mathscr{A}$ of subsets of Ω
$\Omega_0 \cap \mathscr{A}$	trace of the σ-algebra $\mathscr{A}$ on Ω_0
$(\Omega, \mathscr{A}, \mu)$	measure space = measurable space with measure μ on $\mathscr{A}$
$N_p(f)$	$\mathscr{L}^p$-seminorm $[:= (\int \lvert f\rvert^p \, d\mu)^{1/p}]$, $1 \le p < +\infty$
$\mathscr{L}^p(\mu)$	$\mathscr{L}^p$-space of all real-valued $\mathscr{A}$-measurable functions f on Ω with $N_p(f) < +\infty$ $[(\Omega, \mathscr{A}, \mu)$ a measure space], $1 \le p < +\infty$.
$\mathscr{B}(E)$	Borel σ-algebra in a topological space E (generated by its open subsets)
$\mathscr{B}^d = \mathscr{B}(\mathbb{R}^d)$	Borel σ-algebra in $\mathbb{R}^d$
$\bar{\mathscr{B}}^1 = \mathscr{B}(\bar{\mathbb{R}})$	Borel σ-algebra in $\bar{\mathbb{R}}$
λ^d	Lebesgue-Borel measure on $\mathscr{B}^d$ (abbreviation: L-B measure)
λ_C^d	restriction of λ^d to $C \cap \mathscr{B}^d$, $C \in \mathscr{B}^d$ (abbreviation: L-B measure on C)
$f^+ = f \vee 0 = \max\{f, 0\}$	positive part of the numerical function f
$f^- = (-f)^+ = -(f \wedge 0)$	negative part of f
$\lvert f\rvert = f \vee (-f) = f^+ + f^-$	absolute value of f
$f_n \uparrow f$	the sequence (f_n) of numerical functions on Ω is increasing and f is its upper envelope
$f_n \downarrow f$	$(-f_n) \uparrow (-f)$
$A_n \uparrow A$	$1_{A_n} \uparrow 1_A$ (:= indicator function of A)
$A_n \downarrow A$	$1_{A_n} \downarrow 1_A$

Introduction

Even chance is not unfathomable;
it has its regularity.
(NOVALIS, *Fragmente*)

Probability theory owes its existence to the desire to gain mathematical insights into processes governed by chance, and in particular to discover and investigate any laws at work in such processes. Thus the task of developing mathematical models for the study of experiments involving chance stands at the very beginning of the development of probability itself. Originally the primary interest was in the chance mechanisms in gambling games. The search for mathematical models and methods that would permit deeper insights into the ambient space we inhabit stands at the very beginning of the development of geometry. The question of a better understanding of the course of games of chance played a role in the history of probability analogous to that played by questions about land measurement in geometry. Geometry has long since abandoned the restricted line of development suggested by this initial problem and with more sophisticated concepts and methods — for example, differential geometric, algebraic-geometric, and topological — turned to new kinds of question. However, probability theory did not go through a corresponding metamorphosis until more recent times. In fact, as a mathematical theory whose concepts and construction satisfy the usual demands of rigor, it did not even exist before 1933, when A.N. Kolmogorov anchored it in analysis by using a general notion of measure and a theory of integration built on it. Since that time, it has not just gone beyond the original task of modelling chance phenomena. More importantly, in the course of its development, especially in recent years, it has erected bridges to other mathematical disciplines in which originally no connection to probability whatsoever was discernible. Among such disciplines, besides number theory and ergodic theory which came rather early into the scope of probability theory, are the theory of partial differential equations and differential geometry. Surprisingly, confirmation of old results, but above all new points of view and insights, are gained which, though formulatable without any knowledge of probability theory, are not always provable, much less fully transparent, without it.

In view of this multifaceted development, the challenge of writing a quite new book on probability theory is especially attractive. One must not neglect the classical repertory, but at the same time new developments must be given their due. The author tried to keep this obligation in mind.

This book presupposes a basic knowledge of measure and integration theory, as is to be found in many standard textbooks. How this is deployed in probability

theory is presented in the first chapter. The concepts of probability space, event, random variable, expected value, variance and distribution are the main features there. In particular, the reader will get some feeling for the intuition-enhancing special jargon of probability theory. The second chapter treats the concept of (stochastic) independence, which gives the theory its distinctive characteristic and raises it above the plane of general measure and integration theory.

The power of these ideas is first demonstrated in chapter III in studying the Law of Large Numbers. The form of this law discovered by N. Etemadi will be proved. One of its applications which is basic to the further developments concerns the limiting behavior of sums $S_n = X_1 + \ldots + X_n$, in which $X_1, X_2, \ldots$ is an independent sequence of identically distributed, integrable, real random variables. The Zero-One Laws of Borel, Kolmogorov, and Hewitt-Savage find a natural place in this chapter. Closely linked to this is the extensive chapter IV on martingales, which simultaneously accomplishes several tasks within our overall program: It makes available the concept of conditional expectation and that of a martingale, which is so indepensable to the modern theory. Among the many topics where it finds application, mention should be made of the Strong Law of Large Numbers and the whole broad area of stochastic processes. The investigation of the limiting behavior of sums S_n begun in chapter III is then continued in chapters V–VII, chapter V being devoted to preparing the necessary tools from Fourier analysis. This chapter also presupposes some knowledge of measure theory in Polish and in locally compact spaces. The Fourier-transform technique is brought fully to bear on the Central Limit Theorem in chapter VI, the last section of which is devoted to the Gauss measures and normal distributions in higher dimensional euclidean spaces. Chapter VII concludes the study of the sums S_n: Using ideas of A. deAcosta the Law of the Iterated Logarithm in V. Strassen's formulation is proved.

The vast and particularly current theme of stochastic processes (which, as already mentioned, is anticipated in chapter IV) gets treated in the last two chapters (VIII and IX). Chapter VIII deals with the construction of processes having pre-assigned finite-dimensional distributions and with the question of realizing these processes subject to certain requirements on their paths. Two extensive and important classes of processes are studied: Markov and Gauss processes. Special attention is given to Brownian motion as a stochastic process, the related Ornstein–Uhlenbeck process, and the Poisson process.

The last chapter, IX, is almost exclusively devoted to a deeper study of Brownian motion. First, the various connections with martingale theory are brought to the fore and the behavior of Brownian paths investigated. Then examples of stochastic integrals are discussed in order to get Brownian motion into the role of a stochastic "integrator" and to be able to discuss the Ornstein-Uhlenbeck process at the end. Finally, the strong Markov property of Brownian motion is treated and its significance illustrated with numerous examples of its applicability. In the course of this, partial differential equations enter the picture for the first time, notably the Laplace and the heat equation. The reader is perforce led into the questions of Stochastic Analysis. These and other questions will be

investigated just deeply enough to show the reader the newer directions of research and stimulate him to further study.

The exposition is accompanied with many examples. Their job is to make the principal results more understandable to the reader and to demarcate them from questions leading farther afield. The many exercises that accompany the text are designed to deepen it, as well as to provide the reader with the means of testing his understanding and assessing his prowess.

The book easily contains enough material for a two-semester, four-hour-per-week course. But it can be worked through in different ways, some of which are indicated in the schematic on page x.

In conclusion, a few words about the background in measure and integration theory expected of the reader. First of all, this comprises general measure theory and the integral built from it, including product measures (but only involving finitely many factors). In studying martingales some knowledge of equi-integrable sequences of functions will be supposed. The regularity properties of Borel measures come to the fore in chapter V and later in chapter VIII. From chapter V onward a few (but critical) uses of compactness properties of the weak topology are made. To make things a little easier for the reader, specific references (at first frequent but later only occasional) are given to the author's book *Maß- und Integrationstheorie* (= BAUER [1990] in the bibliography). These will be encoded by the letters MI, for example, in the form *MI, Theorem 21.4*. Moreover the notation used in MI will be employed here; the most important of these are assembled in the notation index on page xi.

Chapter I

Basic Concepts of the Theory

The goal of probability theory is to provide methods of describing and analyzing experiments with random outcomes. In particular, mathematical models for an adequate study of such experiments involving chance have to be developed. In such experiments we are interested in the observation of "events" or "random magnitudes", as well as the calculation of the "probability" with which such events occur, or the "expected values" of such magnitudes. Consequently, the first job of the theory is to find an appropriate framework in which to define and study these concepts.

The principal goal of this chapter is to set up these and other basic ideas of the theory. The notion of a probability space and the associated measure and integration theory prove to be fundamental for this undertaking.

§1. Probability spaces and the language of probability theory

1. First of all we'll be concerned with a *throw in a dice game.* The events to be studied are, at the most primitive level, those which are colloquially described as "a k was thrown" ($k = 1, 2, \ldots, 6$). These events — they'll later be called elementary events — correspond in a one-to-one manner with the natural numbers $1, 2, \ldots, 6$ used to name them. But more complicated events are also of interest: the number thrown is even, or it is odd, or it is not a 1. Obviously these events can be identified with sets formed from the numbers $1, 2, \ldots, 6$; namely with $\{2, 4, 6\}$, $\{1, 3, 5\}$ and $\{2, 3, 4, 5, 6\}$, respectively. In this way these events appear as subsets of the set $\Omega := \{1, 2, 3, 4, 5, 6\}$. Any outcome whatsoever of this experiment can be described via the relevant numbers thrown as a subset of Ω; and conversely, every subset of Ω can be interpreted as such an event. Among the subsets of Ω are, in particular, the empty set $\emptyset$ and the set Ω itself; they represent the impossible event (no number at all is showing after the throw) and the certain event (at least one of the numbers $1, 2, \ldots, 6$ is showing). In this way the set of all events or outcomes is identified with the power set $\mathscr{P}(\Omega)$ of Ω.

This identification has a noteworthy property. If A and B are subsets of Ω, that is, events, then it is intuitively clear what we are to understand by "A or B", "A and B", and "not A". By means of the identification described above these

events will be represented by the sets $A \cup B$, $A \cap B$ and $\complement A$, respectively. We can then speak of the algebra, even the σ-algebra $\mathscr{P}(\Omega)$ of events.

In our dice game we further speak of the probability of the event E or the probability of the occurrence of the event E. This probability is a real number $P(E)$ assigned to E. There is thus a function

$$P : \mathscr{P}(\Omega) \to \mathbb{R}$$

on hand. If n throws are made in succession, the event E will occur in, say, k of them and in the other $n - k$ of them the event "not E"$= \complement E$ will occur. It is "practically" certain that the quotient k/n, which measures the frequency of occurrence of the event E, will deviate from $|E| / |\Omega|$ by less than any prescribed $\varepsilon > 0$ if n is sufficiently large. Here $|E|$ denotes the number of elements in the set $E \subset \Omega$. This fact of life, known as the "law of large numbers", leads us to regard the quotient $|E| / |\Omega|$ as the probability of the event E, that is, leads us to define

$$P(E) := \frac{|E|}{|\Omega|} \qquad (E \in \mathscr{P}(\Omega)). \tag{1.1}$$

From (1.1) the following properties of the function P are immediate: $P \geq 0$, $P(\emptyset) = 0$, $P(\Omega) = 1$, $P(E) \in [0,1]$ and $P(E \cup F) = P(E) + P(F)$ for two "incompatible" or "disjoint" events E and F, that is, events E and F with $E \cap F = \emptyset$, whose simultaneous occurrence is impossible. Thus P is a *content* and, on account of the finiteness of Ω, even a *measure* on $\mathscr{P}(\Omega)$, with the "normalization" $P(\Omega) = 1$. Evidently, P is the only measure on $\mathscr{P}(\Omega)$ which is so normalized and for which the "elementary events" $\{k\}$, $k = 1, 2, \ldots, 6$, are all "equi-probable", that is, for which

$$P(\{1\}) = \ldots = P(\{6\}).$$

This requirement of equal probability for the elementary events corresponds to the demand that the die be "fair", not "loaded".

2. Had we discussed, instead of this experiment, another, say the tossing of a coin or three consecutive throws of the die, we would have formally reached the same end. In the first of these cases, Ω would be the set $\{H, T\}$ consisting of the two (distinct) symbols H (for "head") and T (for "tail"). In the second case Ω would comprise all ordered triples (k_1, k_2, k_3) from among the numbers $1, 2, \ldots, 6$. In both cases $\mathscr{P}(\Omega)$ is identified as the set of all events and the relevant probabilities are given via (1.1) as measures on $\mathscr{P}(\Omega)$ with the normalization $P(\Omega) = 1$.

In the examples considered up 'til now Ω was always a finite set and all subsets were interpreted as events. The next example will show that it can be otherwise when infinite Ω are involved.

3. A gun is being fired on a target range. Despite any possible preliminary practice, the point struck by the bullet can be regarded as random. Too many

hard-to-account-for factors influence it; for example, unsureness in the eye and hand of the shooter, mechanical malfunctions (or just minute irregularities) in the gun, air currents along the firing range, etc. The events in which we are interested are then of the following kind: The hit occurs in a prescribed subset E of the target, for example, in one of the twelve rings of a bullseye. This example already clearly shows how events can be identified in a very natural way with subsets E of Ω. The probability $P(E)$ of such an event E can be taken to be approximately proportional to the two-dimensional Lebesgue measure $\lambda^2(E)$ of E; thus we define

$$P(E) := \frac{\lambda^2(E)}{\lambda^2(\Omega)}. \tag{1.2}$$

In doing this we must, of course, require that E be a Borel subset (or at least a Lebesgue measurable subset) of Ω. Nevertheless, the structure of (1.2) is analogous to that of (1.1).

Whereas until now we were inclined to interpret all subsets of Ω as events, here the limited scope of the Lebesgue-Borel measure forces us to regard only certain subsets (that is, hits in such subsets) as events, namely the Borel subsets of Ω. This proves to be completely adequate for practical needs. It is of paramount importance that despite differences from the earlier examples, we have here formally the same situation: The set $\mathscr{A}$ of these (Borel) events is, just like $\mathscr{P}(\Omega)$, a σ-algebra in Ω, the σ-algebra of all Borel subsets of Ω. It appears here in the form of the trace on Ω of the σ-algebra $\mathscr{B}^2$ of all Borel subsets of $\mathbb{R}^2$, that is, in the notation introduced in MI, (1.4), $\mathscr{A} = \Omega \cap \mathscr{B}^2$. The function P defined via (1.2) is again a measure on this σ-algebra, satisfying the normalization $P(\Omega) = 1$. Once again we encounter a measure space $(\Omega, \mathscr{A}, P)$ with the same normalization on P.

In the axiomatic founding of geometry, algebra, topology and other areas of mathematics concepts like point and straight line, number, neighborhood, etc. are not defined intrinsically. Similarly it has turned out that for the construction of a theory of probability intrinsic definitions of concepts like "event" and "probability" are not necessary, and in fact, to avoid logical difficulties and to give the theory the broadest and easiest applicability, such definitions are not worth attempting. Just as in the other areas of mathematics mentioned, in probability theory everything comes down to the formal properties of the concepts. We owe to the Russian mathematician A.N. KOLMOGOROV (1903–1987) — cf. KOLMOGOROV [1933] — the idea that normalized measure spaces can serve for the construction of a theory of probability which meets all the customary demands of rigor in mathematics. Lebesgue's original investigations of the measure and integration concepts were purely geometrically motivated and geometric ideas were in the foreground during their development. It was É. BOREL (1871–1956) who first demonstrated the utility of this theory for solving probabilistic problems, in connection with the law of large numbers: In BOREL [1909] the connection

between probability and the countable additivity of measures is made for the first time. Kolmogorov's vision of founding probability theory on the concept of a normalized measure space has become the accepted orthodoxy.

The basic context for defining probability-theoretic concepts is therefore always a (generally arbitrary, given) normalized measure space. It is frequently designated $(\Omega, \mathscr{A}, P)$, where Ω is a set, $\mathscr{A}$ a σ-algebra in Ω and P a measure on $\mathscr{A}$ normalized by $P(\Omega) = 1$. Every measure normalized in this way is called a *probability measure*. Every measure space involving such a probability measure is called a *probability space*. These concepts come up as early as §6 in MI. Because measures are increasing functions, all probability measures satisfy

$$0 \leq P(A) \leq 1 \qquad (A \in \mathscr{A}). \tag{1.3}$$

The elements of the σ-algebra $\mathscr{A}$ are called *events*. By means of P a number $P(A)$ is assigned to every event $A \in \mathscr{A}$; it is called the *probability* of A, or of the occurrence of the event A. The points $\omega \in \Omega$ are called *elementary events*. An arbitrary set of such elementary events, that is, an arbitrary subset of Ω is occasionally interpreted as a "theoretically possible event". When this is done, the events encompassed by $\mathscr{A}$ are correspondingly interpreted as "observable events". At this point the "scope problem" of measure theory obtrudes upon us; this is the problem, discussed at the end of §8 in MI, of the size and invariance properties of the σ-algebra $\mathscr{A}$.

It should be mentioned that in most concrete examples of probability spaces $(\Omega, \mathscr{A}, P)$ the singleton sets $\{\omega\}$, $\omega \in \Omega$, are all indeed events in the σ-algebra $\mathscr{A}$. This property is, however, not part of the definition of a probability space and is not a consequence of it, as the reader can confirm with easy examples.

While the general concepts and theorems of probability theory deal with an arbitrary probability space $(\Omega, \mathscr{A}, P)$ — the "steering mechanism" of the probabilistic evolution — the treatment of any *concrete* problem of probability theory requires the choice of an *explicit* probability space. In the course of further developments we will repeatedly have to select probability spaces particularly suited to the problem at hand. An important example, basic for what follows, will now be discussed.

4. We consider a finite number n of experiments with random outcomes. Let the probability space $(\Omega_i, \mathscr{A}_i, P_i)$ serve as the mathematical description of the i^{th} experiment $\mathscr{E}_i$, $i = 1, 2, \ldots, n$. We are interested in a new random experiment $\mathscr{E}$ which consists of carrying out the experiments $\mathscr{E}_1, \ldots, \mathscr{E}_n$ one after another, or simultaneously but "without mutual influence". (For example, we could have $\mathscr{E}_1 = \ldots = \mathscr{E}_n$ and each $\mathscr{E}_i$ be the throwing of a fair die. Then the experiment $\mathscr{E}$ consists of n consecutive throws of one die or, what amounts to the same thing, one throw of n identical dice. We have considered the case $n = 3$ of this example in **2.**) The random outcomes of the experiment $\mathscr{E}$ can then be represented by points in the product set $\Omega_1 \times \ldots \times \Omega_n$, that is, by n-tuples $(\omega_1, \ldots, \omega_n)$ of $\omega_i \in \Omega_i$. If now

$A_i \in \mathscr{A}_i$ for $i = 1, \ldots, n$, we may be interested in the experiment $\mathscr{E}$ whether the sequence of events $A_1, A_2, \ldots, A_n$ occurs. One is then interested in that event in the experiment $\mathscr{E}$ which is represented by the subset $A_1 \times \ldots \times A_n$ of $\Omega_1 \times \ldots \times \Omega_n$. As no mutual influence is supposed to prevail among these events, one is inclined to regard the product $P_1(A_1) \cdot \ldots \cdot P_n(A_n)$ as the probability of the event $A_1 \times \ldots \times A_n$. Now the products $A_1 \times \ldots \times A_n$ with A_i running through $\mathscr{A}_i$ generate precisely the product σ-algebra

$$\mathscr{A}_1 \otimes \ldots \otimes \mathscr{A}_n$$

and according to Theorem 23.9 of MI the product measure $P_1 \otimes \ldots \otimes P_n$ is the only measure P which satisfies

$$P(A_1 \times \ldots \times A_n) = P_1(A_1) \cdot \ldots \cdot P_n(A_n)$$

for arbitrary $A_i \in \mathscr{A}_i$. It is obvious that P is a probability measure and that consequently

$$\bigotimes_{i=1}^{n} (\Omega_i, \mathscr{A}_i, P_i)$$

is a probability space adequate to the description of the experiment $\mathscr{E}$. For finite sets Ω_i and for $\mathscr{A}_i := \mathscr{P}(\Omega_i)$ one obviously gets $\mathscr{A}_1 \otimes \ldots \otimes \mathscr{A}_n = \mathscr{P}(\Omega_1 \times \ldots \times \Omega_n)$. And in the case of n throws of a die we come back to the probability space which was introduced *a priori* in **2** for $n = 3$.

The intuition behind the idea of an "event" has led to the introduction of special notations and locutions to supplement the purely set-theoretic ones. If $(\Omega, \mathscr{A}, P)$ is a probability space, then

$$\emptyset \quad \text{and} \quad \Omega$$

are also called the *impossible event* and the *sure event.* Correspondingly, events E with

$$P(E) = 0 \quad \text{or} \quad P(E) = 1$$

are called, respectively, *almost impossible* or *almost sure* (*almost certain*). Instead of (P-)almost everywhere, one says (P-)*almost surely* or *with probability one.* In turn, these are abbreviated to (P-)*a.s.* or *with prob.* 1.

One says that the event E *implies* or *entails* the event F in case

$$E \subset F.$$

The events E and F are called *disjoint* or *incompatible* when

$$E \cap F = \emptyset.$$

The events

$$E \cup F, \quad E \cap F, \quad E \setminus F = E \cap \complement F$$

are described by saying "*at least one of the events E and F occurs*", "*E and F both occur*", "*E occurs but F doesn't*". If $(E_n)_{n\in\mathbb{N}}$ is a sequence of events, then

$$\bigcup_{n\in\mathbb{N}} E_n\,, \quad \bigcap_{n\in\mathbb{N}} E_n$$

are referred to as the event that "*some E_n occurs*", respectively, "*E_n occurs for every n*".

Finally we set

$$\{E_n \text{ for almost all } n\} := \liminf_{n\to\infty} E_n \tag{1.4}$$

(where "almost all" means "with at most finitely many exceptions", equivalently "for all sufficiently large") and

$$\{E_n \text{ for infinitely many } n\} := \limsup_{n\to\infty} E_n \tag{1.5}$$

or in shorter form

$$\{E_n \text{ i.o.}\} := \{E_n \text{ for infinitely many } n\}\,. \tag{1.5'}$$

Here i.o. means "infinitely often". Correspondingly for probabilities we have

$$\begin{aligned} P\{E_n \text{ for some } n\} &= P(\bigcup_{n\in\mathbb{N}} E_n) \\ P\{E_n \text{ for all } n\} &= P(\bigcap_{n\in\mathbb{N}} E_n) \\ P\{E_n \text{ for almost all } n\} &= P(\liminf_{n\to\infty} E_n) \\ P\{E_n \text{ i.o.}\} &= P(\limsup_{n\to\infty} E_n)\,, \end{aligned}$$

in which the round parentheses in expressions like $P(\{\ldots\})$ are routinely dispensed with.

These and other self-evident notations will be employed whenever the situation warrants. They have the advantage of letting us formulate probabilistic assertions in particularly suggestive ways.

§2. Laplace experiments and conditional probabilities

The preceding discussion will now be illustrated with some examples. These typify the kinds of questions asked in elementary probability theory and they also come up in many practical problems. All of them have to do, directly or indirectly, with *Laplace experiments*. These are random experiments with only finitely many

possible outcomes, each equally likely. The associated mathematical model is a probability space $(\Omega, \mathscr{A}, P)$ in which Ω has $N \in \mathbb{N}$ elements, $\mathscr{A} = \mathscr{P}(\Omega)$ and P is the unique measure satisfying $P(\{\omega\}) = N^{-1}$ for every $\omega \in \Omega$. This probability space (which is uniquely determined save for the meaning of the points of Ω) is called the *Laplace probability space of order* N. Calculation of the probability of an event E involves combinatorial considerations. The point of departure is formula (1.1), according to which $P(E)$ is the ratio of the "number of cases favorable to E" to the "number N of possible cases".

1. An urn contains n balls of identical size and texture, of two colors, black and white, and well-mixed; say, b black and w white balls ($b + w = n$). A certain number $m \leq n$ of balls are randomly withdrawn and we ask about the probability that exactly $k \leq b$ of them are black. Two scenarios have to be distinguished:

(a) *Drawing without replacement.* The m balls are withdrawn one after the other and left outside the urn. If we index the balls with the natural numbers 1 through n, then Ω consists of all sequences $(a_1, \ldots, a_m)$ of m distinct numbers from $\{1, \ldots, n\}$ and so it contains

$$N := n(n-1) \cdot \ldots \cdot (n-m+1)$$

elements. Under the hypothesis that we are dealing with a Laplace experiment, that is, $(\Omega, \mathscr{A}, P)$ is the Laplace probability space of order N, we obtain the solution as follows. The event E being studied consists of all sequences $(a_1, \ldots, a_m)$ of the kind described in which black balls are indexed by exactly k of the numbers $a_1, \ldots, a_m$. We can associate k distinct numbers in $\binom{m}{k}$ ways with the available indices $1, \ldots, m$. k black balls can be drawn one at a time in exactly $b(b-1) \cdot \ldots \cdot (b-k+1)$ different ways. The remaining $m-k$ indices can be associated in $w(w-1) \cdot \ldots \cdot (w-m+k+1)$ further ways with numbers on white balls. The probability sought is therefore

$$\binom{m}{k} \frac{b(b-1) \cdot \ldots \cdot (b-k+1) w(w-1) \cdot \ldots \cdot (w-m+k+1)}{n(n-1) \cdot \ldots \cdot (n-m+1)} = \frac{\binom{b}{k}\binom{w}{m-k}}{\binom{n}{m}}.$$

This formula is actually valid for arbitrary $k \leq m$, since for $k > s$ it delivers the probability 0.

We can also interpret the problem as follows: The m balls are not drawn successively from the urn, but are taken *in one draw*. Then there are obviously $\binom{n}{m}$ possible and $\binom{b}{k}\binom{w}{m-k}$ favorable cases. We then arrive at the same probability.

Scenario **(a)** is also encountered, in a different form, in the following practical problem. Instead of an urn with balls, suppose we are discussing a day's production of a mass-produced article. We translate "black" as "defective" and "white"' as "non-defective". Then we have just computed the probability that in a random *sample* of size m from the day's production we find exactly k defectives. Of

course, this involves the number b of all defectives, which is in general unknown. Eliminating the unknown b involves the kind of problems studied in mathematical statistics.

(b) *Drawing with replacement.* Every ball drawn is immediately returned to the urn; after another mixing of the contents of the urn, the next ball is drawn at random. Here we obviously have the situation described under **4** in §1: The Laplace experiment "draw one ball" is repeated m times without mutual influence. The single experiment is described by the Laplace probability space $(\Omega_0, \mathscr{A}_0, P_0)$ of order n, the combined experiment by the product $(\Omega, \mathscr{A}, P)$ of m copies of $(\Omega_0, \mathscr{A}_0, P_0)$. Then $(\Omega, \mathscr{A}, P)$ is the Laplace probability space of order n^m. If A_i denotes the event that the i^{th} draw is a black ball, then

$$p := P_0(A_i) = \frac{b}{n} \quad \text{and} \quad P_0(\complement A_i) = 1 - p\,.$$

The event $A_{i_1 \ldots i_k} \in \mathscr{A}$ of drawing a black ball on the $i_1^{\text{st}}, \ldots, i_k^{\text{th}}$ draws $(1 \leq i_1 < \ldots < i_k \leq m)$ and a white one on the remaining $m - k$ draws is the product $B_1 \times \ldots \times B_n$, where $B_{i_\nu} := A_{i_\nu}$ $(\nu = 1, \ldots, k)$ and $B_j := \complement A_j$ for the remaining indices j. Consequently,

$$P(A_{i_1 \ldots i_k}) = p^k (1-p)^{m-k}$$

and on account of the additivity of P

$$\binom{m}{k} p^k (1-p)^{m-k}, \qquad \text{with } p = \frac{b}{n},$$

is the desired probability.

2. Each of m persons chooses a natural number from the set $\{1, \ldots, n\}$, at random and without knowledge of what anyone else does (with the same given n for all people). After all of them have made their choices, the results are reported. What is the probability that m different numbers were chosen?

The situation can obviously (always under the hypothesis that this is a Laplace experiment) also be described as follows: There are n balls in an urn, m balls are chosen at random with replacement. What is the probability that m different balls will be drawn? Since there are obviously

$$n(n-1) \cdot \ldots \cdot (n-m+1)$$

favorable cases, the answer is

$$\frac{n(n-1) \cdot \ldots \cdot (n-m+1)}{n^m}\,.$$

3. Two cards are chosen successively, without replacement, from a well-shuffled deck of 52 cards. What is the probability that (a) the second card drawn is an ace, (b) the second card is an ace if the first draw already turned up an ace?

To answer (a) we naturally use the Laplace probability space $(\Omega, \mathscr{A}, P)$ of order $51 \cdot 52$. The number of cases favorable to the described event is $4 \cdot 51$. Thus the probability sought is $1/13$. In case (b) we can use the Laplace probability space of order $4 \cdot 51$ and compute the probability sought as $\frac{4 \cdot 3}{4 \cdot 51} = \frac{1}{17}$. Although the same experiment was performed, we are using different probability spaces to answer the two questions. This seems impractical and leads us to the following method of solution using the same probability space: Let $A \in \mathscr{A}$ be the event whose probability was sought in (a) and let $B \in \mathscr{A}$ be the event that the first draw yields an ace. The answer to (b) then goes as follows: $P(B)$ is the number of possible cases and $P(A \cap B)$ is the number of cases favorable to the desired event; hence

$$\frac{P(A \cap B)}{P(B)}$$

is the probability sought.

More generally, we have the following situation: Suppose we are given an arbitrary probability space $(\Omega, \mathscr{A}, P)$ and an event $B \in \mathscr{A}$ with $P(B) > 0$. Then obviously

$$A \mapsto \frac{P(A \cap B)}{P(B)} \qquad (A \in \mathscr{A})$$

is again a probability measure P_B on $\mathscr{A}$. We have $P_B(B) = 1$ even though $P(B)$ is generally not equal to 1. In the passage from P to P_B, B has become an event of probability 1. Accordingly, for every $A \in \mathscr{A}$ we call $P_B(A)$ the *conditional probability of A given B*, or *under the hypothesis B*, and we write $P(A|B)$ for $P_B(A)$; that is, we define

$$P(A|B) := \frac{P(A \cap B)}{P(B)}. \tag{2.1}$$

Thus in the preceding problem (b) we were dealing with the computation of a conditional probability.

Equation (2.1) can be generalized immediately. Suppose we are given a (possibly finite) sequence $(B_n)_{n \in I}$ (where either $I = \{1, \ldots, n_0\}$ with $n_0 \in \mathbb{N}$, or $I = \mathbb{N}$) of pairwise disjoint events $B_n \in \mathscr{A}$ with $P(B_n) > 0$ for all n and $\Omega = \bigcup B_n$. Since $A = \bigcup (A \cap B_n)$, it then follows from the σ-additivity of P that

$$P(A) = \sum_{n \in I} P(A \cap B_n)$$

and hence with the help of (2.1)

$$P(A) = \sum_{n \in I} P(B_n) P(A|B_n) \qquad (A \in \mathscr{A}). \tag{2.2}$$

This is the *formula of total probability*. For $P(A) > 0$ we have

$$P(B_n|A) = \frac{P(B_n) P(A|B_n)}{\sum_{i \in I} P(B_i) P(A|B_i)} \qquad (n \in I), \tag{2.3}$$

since $P(B_n|A) = P(A \cap B_n)\,(P(A))^{-1}$. This is the *Bayes formula*, named after TH. BAYES (1702–1761). The usefulness of both formulas is illustrated by the following example:

4. Suppose we are given $N+1$ urns $U_0, \ldots, U_N$. In each urn U_n there are N identically-shaped balls, well mixed; n of them are black and $N-n$ are white $(n = 0, \ldots, N)$. Suppose we choose an urn at random and draw one ball from it. What is the probability that we draw a black ball?

Let B_n be the event that the ball was drawn from the n^{th} urn and A the event that a black ball was chosen. Under the assumption that all urns are equally probable, we have

$$P(B_n) = (N+1)^{-1} \quad \text{and} \quad P(A|B_n) = \frac{n}{N}\,.$$

Then by (2.2)

$$P(A) = \frac{1 + \ldots + N}{N(N+1)} = \frac{1}{2}\,.$$

We ask further: Assume that a black ball was drawn. What is the probability that it was drawn from the n^{th} urn? The answer is given by (2.3):

$$P(B_n|A) = \frac{2n}{N(N+1)}\,.$$

As expected, this probability increases proportionally with n.

Since we are obviously dealing with a Laplace experiment, the usual combinatorial considerations would also have led to the same results. But formulas (2.2) and (2.3) make our work easier. If we now imagine a random mechanism by which we choose the urn U_n with probability $\alpha_n > 0$ $(\alpha_0 + \ldots + \alpha_N = 1)$, then we have a Laplace experiment only if, as before, the α_n are all equal. Nevertheless, (2.2) and (2.3) lead to answers: We now have $P(B_n) = \alpha_n$ and again $P(A|B_n) = n/N$. Thus if we introduce

$$M := \alpha_1 + 2\alpha_2 + \ldots + N\alpha_N\,,$$

we obtain

$$P(A) = \frac{M}{N} \quad \text{and} \quad P(B_n|A) = \frac{n\alpha_n}{M}\,.$$

We have a Laplace experiment here only after having chosen an urn. We therefore speak of a *relay-experiment.*

The detailed specification of the probability space used is also very instructive. It is based on the following general considerations: Let Ω be an arbitrary nonempty set and $(\Omega_i)_{i=1,\ldots,m}$ a partition of Ω into pairwise disjoint sets $\Omega_i \neq \emptyset$. Suppose each of those sets is the carrier of a probability space $(\Omega_i, \mathscr{A}_i, P_i)$. The system $\mathscr{A}$ of all sets $A \subset \Omega$ with $A \cap \Omega_i \in \mathscr{A}_i$ for every $i = 1, \ldots, m$ is a σ-algebra in Ω consisting of all sets $A_1 \cup \ldots \cup A_m$ with $A_i \in \mathscr{A}_i$. $(\Omega, \mathscr{A})$ is called the *direct*

sum of the measurable spaces $(\Omega_i, \mathscr{A}_i)$. Furthermore, if $\alpha_1, \ldots, \alpha_m$ are positive real numbers which satisfy $\alpha_1 + \ldots + \alpha_m = 1$, then on account of (2.2)

$$A \mapsto \sum_{i=1}^{m} \alpha_i P_i(A \cap \Omega_i)$$

is the only probability measure P on $\mathscr{A}$ satisfying $P(\Omega_i) = \alpha_i$ and $P(A|\Omega_i) = P_i(A \cap \Omega_i)$ for every $i = 1, \ldots, m$.

In our example, $m = N + 1$, every $(\Omega_i, \mathscr{A}_i, P_i)$ is a Laplace probability space of order N with pairwise disjoint sets Ω_i, and $\Omega = \bigcup \Omega_i$. The probability space $(\Omega, \mathscr{A}, P)$ just constructed is then the mathematical model of the relay-experiment described earlier.

Exercises

1. A die is tossed n times. What is the probability that exactly at the n^{th} toss a 4 is thrown for the k^{th} time $(1 \leq k \leq n)$?

2. Suppose we are given three urns U_1, U_2, U_3 which contain, well mixed, identically-shaped balls, each either black, white or red. Assume that

U_1 contains 2 black, 3 white, 5 red balls,

U_2 contains 4 black, 2 white, 4 red balls,

U_3 contains 2 black, 5 white, 3 red balls.

(a) What is the probability of drawing from U_1 without replacement first a black, then a black, and then a red ball?

(b) What is the probability P_b (or P_w, or P_r) of drawing, after choosing at random one of the urns, a black (or white, or red) ball? Why is $P_b + P_w + P_r = 1$?

(c) What is the probability of drawing, after choosing at random one of the urns, 4 black balls successively without replacement? What is the probability that these 4 black balls come from urn U_2?

3. How great is the probability that 10 persons chosen at random have their birthdays in different months?

4. In an urn there are, well mixed, identically-shaped balls of r different colors; namely, $k_i > 0$ balls of color C_i $(i = 1, \ldots, r)$. Suppose that n balls are taken in one draw $(1 \leq n \leq k_1 + \ldots + k_r)$. What is the probability of obtaining exactly n_i balls of color C_i $(n_i \geq 0, n_1 + \ldots + n_r = n)$?

5. In an urn there are $b \geq 1$ black and $w \geq 1$ white balls, well mixed. We draw balls successively without replacement. What is the probability that in d draws at least k $(\leq b)$ black balls are obtained? (Here $d \in \{1, \ldots, b + w\}$.)

6. (*Pólya's urn model*) In an urn there are $b \geq 1$ black and $w \geq 1$ white balls, well mixed. A ball is drawn at random. It is replaced and, moreover, t balls of the color drawn are added. The next ball is drawn at random and the above procedure is repeated. Define $N := b + w$, and let n be a natural number. Prove: The probability that in n draws k black and $n - k$ white balls appear $(0 \leq k \leq n)$ equals

$$\binom{n}{k} \cdot \frac{b(b+t) \cdot \ldots \cdot (b+[k-1]t)w(w+t) \cdot \ldots \cdot (w+[n-k-1]t)}{N(N+t) \cdot \ldots \cdot (N+[n-1]t)}.$$

Is the situation described in **1(b)** a special case of this?

§3. Random variables: Distribution, expected value, variance, Jensen's inequality

In an experiment with random outcomes, we are often interested not only in the random outcome itself, but also in numbers and more general mathematical quantities determined by the random outcome of the experiment. Such quantities are called random quantities or random variables.

We might think, say, of the sum of the spots on three throws of a die or the distance of a hit from the center of a target in shooting. If $(\Omega, \mathscr{A}, P)$ is the probability space constructed in §1 for the mathematical description of those experiments, then with every elementary event $\omega \in \Omega$ we associate a real number $X(\omega)$; in this case the sum of spots on the die or the distance of a hit from the bullseye. In both examples it is evident that we are dealing with an $\mathscr{A}$-measurable mapping $X : \Omega \to \mathbb{R}$. Because of this measurability, the inverse image $X^{-1}(B)$ is an event for every Borel set $B \subset \mathbb{R}$. And then $P(X^{-1}(B))$ is interpreted as the probability that X takes a value in B.

Motivated by these and similar examples, we define

3.1 Definition. Let $(\Omega, \mathscr{A}, P)$ be a probability space and $(\Omega', \mathscr{A}')$ a measurable space. Then every $\mathscr{A}$-$\mathscr{A}'$-measurable mapping $X : \Omega \to \Omega'$ is called a *random variable (with values in Ω')* or a *$(\Omega', \mathscr{A}')$-random variable.*

In the cases $(\Omega', \mathscr{A}')$ is $(\mathbb{R}, \mathscr{B}^1)$ or $(\bar{\mathbb{R}}, \bar{\mathscr{B}}^1)$ or $(\mathbb{R}^d, \mathscr{B}^d)$, we also speak of *real* or *numerical* or *d-dimensional* (or $\mathbb{R}^d$-valued) *random variables.* For every event $A \in \mathscr{A}$, the indicator function 1_A is a real random variable. It is called the *indicator variable* of A. Instead of elementary functions, in probability theory we speak of *elementary random variables.* These are thus those of the form

$$X = \sum_{i=1}^{n} \alpha_i 1_{A_i}$$

with $n \in \mathbb{N}$, events $A_i \in \mathscr{A}$ and coefficients $\alpha_i \in \mathbb{R}_+$. If coefficients $\alpha_i \in \mathbb{R}$ are allowed, then we speak of *simple* random variables.

A mapping $X : \Omega \to \mathbb{R}^d$ is — e.g., according to Remark 2 in §22 of MI — a d-dimensional random variable if and only if each of its components is a real random variable. Random variables are customarily denoted by capital Latin letters, frequently X, Y, Z. In what follows, when we deal with events or random variables, we will always mean events or random variables from or on the same probability space $(\Omega, \mathscr{A}, P)$.

We return to the general case of an $(\Omega', \mathscr{A}')$-random variable X. In analogy with the customary notation of integration theory (cf. MI, §9) and with a view to the intuitive interpretation given in the introduction, we set (for $A' \in \mathscr{A}'$)

$$\{X \in A'\} := X^{-1}(A') \tag{3.1}$$

and, after dispensing with some parentheses,

$$P\{X \in A'\} := P(X^{-1}(A')) . \tag{3.2}$$

We call $\{X \in A'\}$ the event "*X lies in A'*" and $P\{X \in A'\}$ the probability of this event.

The mapping $A' \mapsto P\{X \in A'\}$ is nothing other than the image measure $X(P)$ of P under X (cf. MI, §7). Since $P\{X \in \Omega'\} = P(\Omega) = 1$, it is a probability measure on $\mathscr{A}'$.

3.2 Definition. Let X be a $(\Omega', \mathscr{A}')$-random variable on a probability space $(\Omega, \mathscr{A}, P)$. Then the image measure $X(P)$ is called the *distribution* (or the *probability law*) of X (with respect to the probability measure P). We use both the symbols $\mathrm{dist}(X)$ and P_X for it; that is, we set

$$\mathrm{dist}(X) := P_X := X(P) . \tag{3.3}$$

Should it be necessary to stress the role of P, we also write

$$P\text{-}\mathrm{dist}(X) := P_X . \tag{3.3'}$$

Thus the equality

$$P_X(A') = P\{X \in A'\} \qquad (A' \in \mathscr{A}') \tag{3.4}$$

holds as a matter of definition.

The probabilities $P\{X \in A'\}$ can consequently be computed by means of the distribution P_X. For this one does not need to know explicitly — and this is of special importance for the applications — the generally rather complicated probability space $(\Omega, \mathscr{A}, P)$ steering the random phenomena. Accordingly those concepts and properties of random variables which can be formulated in terms of their distributions play the most prominent role in probability theory. Such concepts and properties are occasionally called *probability-theoretic concepts* or *properties*.

The expected value of a real or numerical random variable X, to be defined now, is such a concept. Intuitively, to every elementary event $\omega \in \Omega$ corresponds the chance-determined value $X(\omega)$ of X. It's in the nature of experiments with random outcomes that the question of the "mean" or "expected" value of X should come up. Since $P(\Omega) = 1$, the integral $\int X\, dP$, whenever it exists, is a good candidate for this job.

3.3 Definition. Let X be a numerical random variable on a probability space $(\Omega, \mathscr{A}, P)$. If either $X \geq 0$ or X is P-integrable, we call

$$E(X) := E_P(X) := \int X\, dP \tag{3.5}$$

the *expected value* of X.

The properties of the expected value are, thus, those of the integral. In particular, the condition

$$E(|X|) < +\infty$$

is equivalent to the integrability of the numerical random variable X (cf. MI, Theorem 12.2), and for integrable X

$$|E(X)| \leq E(|X|)\,. \tag{3.6}$$

Definition (3.5) still makes sense if X is only quasi-integrable, that is, if only one of $E(X^+)$ and $E(X^-)$ is finite (cf. the Remark in §12 of MI). To indicate this state of affairs succinctly we say *the expected value $E(X)$ exists*. However, this extended version of Definition 3.3 will be needed only very seldom in the sequel.

We restrict ourselves in the further discussion to a *real* random variable X. The distribution P_X is a probability measure on $\mathscr{B}^1$. A general transformation theorem for image measures (cf. MI, §19) gives

$$E(f \circ X) = \int f\, dP_X\,, \tag{3.7}$$

also written as

$$E_P(f \circ X) = E_{P_X}(f) \tag{3.7'}$$

for every Borel measurable real function f on $\mathbb{R}$ which is either non-negative or P_X-integrable.

Thus if either $X \geq 0$ or X is integrable, and we take for f the function $x \mapsto x$ on $\mathbb{R}$, we get

$$E(X) = \int x P_X(dx)\,, \tag{3.8}$$

thereby recognizing that $E(X)$ is a probability-theoretic concept, that is, that it really depends only on the distribution. Likewise for the integrability of X: it is equivalent to the P_X-integrability of the function $x \mapsto x$ on $\mathbb{R}$.

We shall now reformulate the foregoing in a way that will prove particularly useful later.

3.4 Theorem. *Every real random variable X satisfies*

$$\sum_{n=1}^{\infty} P\{|X| \geq n\} \leq E(|X|) \leq 1 + \sum_{n=1}^{\infty} P\{|X| \geq n\}\,; \tag{3.9}$$

so the integrability of X is equivalent to the convergence of the series

$$\sum_{n=1}^{\infty} P\{|X| \geq n\}\,.$$

If X takes only values in $\mathbb{N}$, then

$$E(X) = \sum_{n=1}^{\infty} P\{X \geq n\}\,. \tag{3.10}$$

Proof. The integrability claim follows at once from (3.9) because integrability just means $E(|X|) < +\infty$.

In proving (3.9) we can assume that $X \geq 0$. The events

$$A_n := \{n \leq X < n+1\} \qquad (n = 0, 1, \ldots)$$

are pairwise disjoint and cover Ω. Consequently,

$$E(X) = \sum_{n=0}^{\infty} \int_{A_n} X\, dP\,. \tag{3.11}$$

From the definition of A_n the inequalities

$$nP(A_n) \leq \int_{A_n} X\, dP \leq (n+1)P(A_n) \qquad (n = 0, 1, \ldots)$$

are immediate and from them, on account of $\sum P(A_n) = P(\Omega) = 1$, follows

$$\sum_{n=1}^{\infty} nP(A_n) \leq E(X) \leq \sum_{n=0}^{\infty} (n+1)P(A_n) = 1 + \sum_{n=1}^{\infty} nP(A_n)\,. \tag{3.12}$$

If we set $B_n := \{X \geq n\}$ for $n = 0, 1, \ldots$, then $B_{n+1} \subset B_n$,

$$A_n = B_n \setminus B_{n+1}\,,$$

and for every $N \in \mathbb{N}$

$$\begin{aligned} \sum_{n=1}^{N} nP(A_n) &= \sum_{n=1}^{N} nP(B_n) - \sum_{n=1}^{N} nP(B_{n+1}) \\ &= \sum_{n=1}^{N} nP(B_n) - \sum_{n=1}^{N} (n-1)P(B_n) - NP(B_{N+1})\,, \end{aligned}$$

so that

$$\sum_{n=1}^{N} nP(A_n) + NP(B_{N+1}) = \sum_{n=1}^{N} P(B_n)\,. \tag{3.13}$$

Since

$$0 \leq NP(B_{N+1}) \leq (N+1)P(B_{N+1}) \leq \int_{B_{N+1}} X\, dP\,,$$

and its right-hand member converges to 0, thanks to $B_n \downarrow \emptyset$, whenever X is integrable, passage to the limit on N in (3.13) yields

$$\sum_{n=1}^{\infty} nP(A_n) = \sum_{n=1}^{\infty} P(B_n)\,, \tag{3.13'}$$

for integrable X; and (3.9) for such X then follows from (3.12). In case $E(X) = +\infty$, (3.12) informs us that the series $\sum_{n=1}^{\infty} nP(A_n)$ diverges. *A fortiori* then from (3.13) the series $\sum_{n=1}^{\infty} P(B_n)$ will diverge. This confirms (3.13') and the validity of (3.9) in this case. If X takes only values from $\mathbb{N}$, then $A_n = \{X = n\}$ and (3.11) reads

$$E(X) = \sum_{n=1}^{\infty} nP(A_n)\,.$$

Equality (3.10) is therefore affirmed in (3.13'). □

At this point let us agree upon a useful (that will be borne out later) alternative way of writing integrals of the form $\int_A X\,dP$.

Namely, if X is a real (or numerical) random variable which is non-negative or integrable, we will set

$$E(X;A) := \int_A X\,dP \qquad (A \in \mathscr{A}) \tag{3.14}$$

If A has a form like $\{\alpha \le X \le \beta\}$, $\{X \ne Y\}$, etc. we will further shorten this by dispensing with the curly brackets:

$$E(X;\alpha \le X \le \beta)\,,\ E(X;X \ne Y)\,,\ \text{etc.}$$

These conventions will reduce unnecessary notational headaches.

Besides the expected value of a real random variable X, the expected values $E(X^p)$ for exponents $p \in \mathbb{N}$ and $E(|X|^p)$ for real $p \ge 1$ play an important role; of course, for the first of these, integrability of X^p has to be hypothesized. $E(X^p)$ is called the *central* p^{th} *moment* and $E(|X|^p)$ the *absolute* p^{th} *moment*. The transformation theorem for image measures invoked at (3.7) shows that here, just as in the case $p = 1$, we have to do with a probability-theoretic concept. For random variables $X \in \mathscr{L}^p(P)$ and real numbers α there are also the p^{th} *moment* (resp., *absolute moment*) *centered at* α defined by

$$E((X-\alpha)^p)\,,\quad \text{resp.,}\quad E(|X-\alpha|^p)\,.$$

The case $p = 2$ is of special importance.

3.5 Definition. For every integrable real random variable X the expression

$$V(X) := E([X - E(X)]^2) \tag{3.15}$$

is called the *variance* of X. The expression

$$\sigma(X) := +\sqrt{V(X)} \in [0, +\infty] \tag{3.16}$$

is called the *standard deviation* (or the *dispersion*) of X. Often the notation $\sigma^2(X)$ or $\mathrm{Var}(X)$ is preferred to $V(X)$.

3.6 Theorem. *Let X be a real random variable on a probability space $(\Omega, \mathscr{A}, P)$. For integrable X we always have*

$$V(X) = E(X^2) - E(X)^2 = \int x^2 P_X(dx) - \left(\int x P_X(dx)\right)^2, \tag{3.17}$$

$$E(X)^2 \le E(X^2). \tag{3.17'}$$

If follows that X is square-integrable if and only if X is integrable and $V(X)$ is finite.

Proof. Suppose X is integrable. Then $\alpha := E(X)$ is a real number and this constant function lies in all $\mathscr{L}^p(P)$ spaces because P is a finite measure. Hence either both X and $X - \alpha$ belong to $\mathscr{L}^2(P)$ or neither does. If both do, the first equality in (3.17) is a consequence of (3.15) and linearity of the integral. If neither does, both sides of that equation are $+\infty$. In either case, (3.17') follows from the non-negativity of $V(X)$ and the finiteness of $E(X)$. The second equality in (3.17) comes from (3.7).

From (3.17) it is clear that X is square-integrable if X is integrable and $V(X) < +\infty$. On the other hand, if X is square-integrable, then $|X| \le 1 \vee |X|^2 \le 1 + |X|^2 \in \mathscr{L}^1(P)$. □

Remark. 1. From (3.17) it follows that X and $X - E(X)$ have the same variance; we have, namely

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

Real random variables Y with $E(Y) = 0$ are called *centered*. In this terminology the passage from X to $X - E(X)$ produces a centered random variable having the same variance as X, a process called *centering on the expected value*.

Via the variance, the well-known *Chebyshev Inequality* ((20.1) in MI) can be put in the form most often used in this subject:

$$P\{|X - E(X)| \ge \alpha\} \le \frac{1}{\alpha^2} V(X), \tag{3.18}$$

in which α is a positive real number, X an integrable real random variable. If it has finite variance, that is, if in addition $V(X) < +\infty$, then from (3.18) follows the assertion that the probability $P\{|X - E(X)| \geq \alpha\}$ of a random variable's deviating by at least α from its expected value tends to 0 as α tends to ∞.

We return to the inequalities (3.6) and (3.17′) for expected values. These turn out to be special cases of a general inequality, named after its discoverer the Danish mathematician J. JENSEN (1859–1925), which we plan to present next.

First we need to review some properties of convex functions (cf. VARBERG and ROBERTS [1973]). Let $I \subset \mathbb{R}$ be an arbitrary non-void interval. A real function q defined on I is called *convex* if it satisfies

$$q(\alpha x + (1-\alpha)y) \leq \alpha q(x) + (1-\alpha)q(y) \tag{3.19}$$

for all $x, y \in I$ and all $\alpha \in [0,1]$. When the function $-q$ is convex, then q is called *concave*. Of course, we need only demand (3.19) for $x < y$ and $0 < \alpha < 1$.

In basic analysis courses one mostly learns only that for differentiable functions q on I the convexity of q is equivalent to the derivative q' being an increasing function. The following considerations, which are of an analogous flavor, have to proceed without the differentiability assumptions if we want them to apply to functions like $x \mapsto |x|$, which are not differentiable (at $x = 0$) although (as seen by an easy application of the triangle inequality) convex on $\mathbb{R}$. More generally, for every $p \geq 1$ the function $x \mapsto |x|^p$ is convex on $\mathbb{R}$. For $p > 0$ this function is differentiable at 0 (and certainly elsewhere on $\mathbb{R}$) and its derivative is the function $x \mapsto \operatorname{sign}(x) p\,|x|^{p-1}$, which is increasing. Here $\operatorname{sign}(x)$ is $+1$, -1 or 0 according as $x > 0$, $x < 0$ or $x = 0$.

It is immediate from (3.19) that the upper envelope $q_1 \vee \ldots \vee q_n$ of finitely many functions $q_1, \ldots, q_n$ which are each convex on I is itself convex on I. Candidates for the q_i are, for example, the *affine* functions $x \mapsto \alpha_i + \beta_i x$ ($\alpha_i, \beta_i \in \mathbb{R}$). These are, evidently, precisely the functions which are *both* convex and concave on I.

The geometric significance of (3.19) is the following: For $x, y \in I$ with $x < y$ and $t \in [x, y]$ every point $(t, q(t))$ of the graph of q lies underneath the line segment joining the points $(x, q(x))$ and $(y, q(y))$. Consequently, the inequality

$$q(t) \leq q(x) + S(x,y)(t-x) \qquad (t \in [x,y]), \tag{3.19′}$$

for arbitrary $x, y \in I$ with $x < y$, is equivalent to (3.19). The notation here means the *difference quotient*

$$S(x,y) := \frac{q(x) - q(y)}{x - y}, \tag{3.20}$$

defined for every distinct pair $x, y \in I$. It measures the *slope* of the aforementioned line segment. Evidently, $S(x,y) = S(y,x)$.

A slight reformulation of (3.19) leads to an equivalent description of convex functions.

3.7 Lemma. *A real function q on I is convex if and only if for any three points $x < t < y$ in I the inequalities*

$$S(x,t) \leq S(x,y) \leq S(t,y) \tag{3.21}$$

are satisfied.

Proof. A natural parameterization of the segment T joining $(x, q(x))$ and $(y, q(y))$ is

$$\xi \mapsto (\xi, q(x) + S(x,y)(\xi - x)) \qquad (\xi \in [x,y]).$$

Therefore, if q is convex, the geometric observation made earlier tells us that

$$q(t) \leq q(x) + S(x,y)(t-x)\,,$$

whence $S(x,t) \leq S(x,y)$. Upon noting the equality

$$q(x) + S(x,y)(\xi - x) = q(y) + S(x,y)(\xi - y)$$

for $\xi \in [x,y]$, we obtain the second inequality claimed in an analogous way. For the converse direction of the lemma, write $t \in]x,y[$ in the form $t = \alpha x + (1-\alpha)y$ with $0 < \alpha < 1$. □

The consequence which is most important for us is:

3.8 Theorem. *A convex function q defined on an interval I has both left-hand and right-hand derivatives at each interior point of I and is consequently continuous on the interior $\mathring{I}$ of I. Letting $q'_+(x)$ denote the right-hand derivative of q at $x \in \mathring{I}$, yields an increasing function q'_+ on $\mathring{I}$, which moreover satisfies*

$$q(y) \geq q(x) + q'_+(x)(y-x) \qquad (x \in \mathring{I},\ y \in I). \tag{3.22}$$

Inequality (3.22) expresses the fact that the right-hand tangent to the graph of q at the point $(x, q(x))$ runs *beneath* the graph of q.

Proof. Let $x \in \mathring{I}$. For every $t_1, t_2 \in I$ with $x < t_1 < t_2$ (notice that there are such points in I) (3.21) furnishes the inequality $S(x,t_1) \leq S(x,t_2)$, so $t \mapsto S(x,t)$ is an increasing function on $]x, +\infty[\cap I$. There are also points $x_1 < x$ in I and another application of (3.21) shows that the number $S(x_1, x)$ is a lower bound for this function of t. The monotonicity then insures the existence of the (real-valued) right-hand derivative

$$q'_+(x) := \lim_{t \to x, t > x} S(x,t)\,.$$

From this of course follows the right-hand continuity of q at every $x \in \mathring{I}$. Analogous reasoning — or application of the preceding result to the convex function $x \mapsto q(-x)$ — confirms the left-hand differentiability of q on $\mathring{I}$ and therewith the continuity of q on $\mathring{I}$.

Given $x, y \in \mathring{I}$ with $x < y$, choose $t, u \in I$ with $x < t < y < u$ and apply (3.21) twice, to get

$$S(x,t) \leq S(x,y) \leq S(y,u)\,.$$

Let $t \downarrow x$, $u \downarrow y$ and these yield $q'_+(x) \leq q'_+(y)$, that is, the fact that q'_+ is an increasing function. To prove (3.22), let $x \in \mathring{I}$, $y \in I$. Consider first the case $x < y$ and choose $t \in]x, y[$. Then the inequality (3.22) follows from a passage to the limit $t \downarrow x$ in the left-hand inequality of (3.21). An analogous application of (3.21), this time with $y < x < t$, confirms (3.22) in case $y < x$. When $x = y$ (3.22) is trivially an equality. □

It should be noted that q'_+ at the left-hand endpoint of I may possibly exist only in the improper sense of being $-\infty$, and q'_- at the right-hand endpoint as $+\infty$. An example of both phenomena, which is moreover discontinuous at each endpoint is the indicator function of the doubleton $\{0,1\}$ on the interval $I := [0,1]$.

We are now in a position to acquire the generalization of inequalities (3.6) and (3.17′) alluded to earlier. These inequalities result from the choices $I = \mathbb{R}$, $q(x) := |x|$, and $q(x) := x^2$ in the theorem.

3.9 Theorem (Jensen's Inequality). *Let X be an integrable random variable on a probability space $(\Omega, \mathscr{A}, P)$ taking values in an open interval $I \subset \mathbb{R}$. Then the expected value $E(X)$ lies in I. For every convex function q on I the composite $q \circ X$ is a random variable. If it is integrable, then*

$$q(E(X)) \leq E(q \circ X)\,. \tag{3.23}$$

Proof. That $E(X)$ lies in I is essentially a consequence of the well-known fact that $E(Y) = 0$ for a random variable $Y \geq 0$ implies $Y = 0$ P-a.s. (cf. MI, 13.2): If $X(\omega) < \beta \in \mathbb{R}$ for all $\omega \in \Omega$, then $E(X) \leq \beta$. This entails $E(X) < \beta$ since, otherwise, $E(\beta - X) = 0$ would imply $\beta - X = 0$ P-a.s. which contradicts $\Omega = \{X < \beta\}$. Similarly, $\alpha < X(\omega)$ for all $\omega \in \Omega$ and some $\alpha \in \mathbb{R}$ implies $\alpha < E(X)$.

According to 3.8 q is continuous on $\mathring{I} = I$. Hence for U open $\subset \mathbb{R}$, $q^{-1}(U)$ is open and consequently $(q \circ X)^{-1}(U) = X^{-1}(q^{-1}(U)) \in \mathscr{A}$. That is, $q \circ X$ is $\mathscr{A}$-measurable, in other words, a random variable. The proof of Jensen's inequality itself rests on (3.22):

$$q(y) \geq q(x) + q'_+(x)(y - x) \qquad (x, y \in I). \tag{3.22′}$$

Since equality holds when $y = x$, we get

$$q(y) = \sup_{x \in I} \left[q(x) + q'_+(x)(y - x)\right] \qquad \text{for all } y \in I. \tag{3.24}$$

Taking y in (3.22′) to be $X(\omega)$ gives

$$q \circ X \geq q(x) + q'_+(x)(X - x)$$

and from this, in case $q \circ X$ is integrable

$$E(q \circ X) \geq q(x) + q'_+(x)(E(X) - x)\,,$$

holding for all $x \in I$. Since $E(X) \in I$, this inequality and the choice $y = E(X)$ in (3.24) yield

$$E(q \circ X) \geq \sup_{x \in I}[q(x) + q'_+(x)(E(X) - x)] = q(E(X))\,,$$

which is Jensen's inequality. □

As an immediate application we get the inequality

$$|E(X)|^p \leq E(|X|^p) \qquad (p \geq 1) \tag{3.25}$$

for every p^{th}-power integrable real random variable X; we have only to note that for such X, $|X| \leq 1 \vee |X|^p \leq 1 + |X|^p \in \mathscr{L}^1(P)$.

Remarks. 2. If we take an open interval I in $\mathbb{R}$ and the associated probability space $(I, I \cap \mathscr{B}^1, P)$ with the probability measure $P := \alpha\varepsilon_x + (1 - \alpha)\varepsilon_y$, $x, y \in I$, $\alpha \in [0, 1]$, and apply Jensen's inequality to the random variable X defined by $x \mapsto x$, we recover the inequality (3.19) which defines convexity. (Here ε_x is the Dirac measure at x.) Thus the convexity hypothesis in 3.9 is necessary as well as sufficient.

3. The proof of Jensen's inequality rests, in the final analysis, on the observation that a family of affine functions $(a_j)_{j \in J}$ on $\mathbb{R}$ exists satisfying

$$q(y) = \sup_{j \in J} a_j(y) \qquad \text{for all } y \in I. \tag{3.26}$$

In our case this was the family $(a_x)_{x \in I}$ given by

$$a_x(\xi) := q(x) + q'_+(x)(\xi - x) \qquad (\xi \in \mathbb{R}).$$

But there is always in fact a *sequence* $(a_j)_{j \in \mathbb{N}}$ which realizes (3.26). To see this, set $I_0 := I \cap \mathbb{Q}$. Then I_0 is countable and in place of (3.24) we have the (stronger) assertion

$$q(y) = \sup_{x \in I_0}[q(x) + q'_+(x)(y - x)] \tag{3.27}$$

for every $y \in I$. To prove this, select a sequence (x_n) in I_0 which converges to y. According to 3.8 q'_+ is an increasing function, consequently locally bounded; moreover, q is continuous on the open interval I. It therefore follows that

$$\lim_{n\to\infty} [q(x_n) + q'_+(x_n)(y - x_n)] = q(y)\,.$$

Equality (3.27) follows from this and (3.22′).

These remarks will be of service later in the proof of a Jensen inequality for *conditional* expectations (in Theorem 15.3).

Exercises

1. A $(\mathbb{R}^d, \mathscr{B}^d)$-random variable X on a probability space $(\Omega, \mathscr{A}, P)$ assumes only countable many values, ω'_i, $i \in I$ (a countable set). Show that

$$P_X = \sum_{i\in I} P\{X = \omega'_i\}\varepsilon_{\omega'_i}\,.$$

2. Consider the Laplace probability space $(\Omega, \mathscr{A}, P)$ from the **(b)** scenario in **1.** of §2; there $\Omega = \Omega_0 \times \ldots \times \Omega_0$ is the product of m copies of an n-element set Ω_0 (comprised, e.g., of colored balls). Let Ω_0 be divided into two disjoint pieces Ω_0^b and Ω_0^w (e.g., the black balls and the white balls). For each $\omega = (\omega_1, \ldots, \omega_m)$ let $X(\omega)$ denote the number of indices $i \in \{1, \ldots, m\}$ for which $\omega_i \in \Omega_0^b$. Determine the distribution of the random variable X.

3. Let $g : \mathbb{R} \to \mathbb{R}_+$ be an even, Borel measurable function which is increasing on $\mathbb{R}_+$ and strictly positive on $\mathbb{R} \setminus \{0\}$. Let X be a real random variable. Prove the following *generalization of the Chebyshev-Markov inequality* (cf. MI, (20.1)):

$$P\{|X| \geq \alpha\} \leq \frac{1}{g(\alpha)} A(g(X)) \qquad (\alpha > 0).$$

4. Let X be a random variable on $(\Omega, \mathscr{A}, P)$ with values in $\mathbb{N}$. Prove the equality

$$E(X) = \sum_{n\in\mathbb{N}} P\{X \geq n\}$$

(a) from scratch in an elementary way, and (b) by means of MI (23.10).

5. In the situation of Theorem 3.8, show that for any $x, y \in \mathring{I}$ with $x < y$ the inequalities

$$q'_-(x) \leq q'_+(x) \leq q'_-(y) \leq q'_+(y)$$

hold.

6. Use Jensen's inequality to deduce the following property of convex functions q on open intervals I of $\mathbb{R}$: For any finite number of points $x_1, \ldots, x_n \in I$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}_+$ with $\lambda_1 + \ldots + \lambda_n = 1$

$$q(\sum_{i=1}^n \lambda_i x_i) \leq \sum_{i=1}^n \lambda_i q(x_i)\,.$$

Does this hold if I is an arbitrary interval? What can be said if $n = \infty$ is allowed?

7. Prove that Theorem 3.9 is valid for arbitrary intervals $I \subset \mathbb{R}$, by analyzing the behavior of q at each endpoint of I. In fact, first show that q is *lower semicontinuous* on I, that is, $\{x \in I : q(x) > \alpha\}$ is relatively open in I for every $\alpha \in \mathbb{R}$.

§4. Special distributions and their properties

For a given measurable space $(\Omega', \mathscr{A}')$ all probability measures P' on $\mathscr{A}'$ appear as distributions of $(\Omega', \mathscr{A}')$-random variables when we allow arbitrary probability spaces $(\Omega, \mathscr{A}, P)$. We need only choose $\Omega = \Omega'$, $\mathscr{A} = \mathscr{A}'$, $P = P'$ and the identity mapping of Ω onto itself for X in order to obtain $P_X = P'$. This is why probability measures on a σ-algebra are often called (*probability*) *distributions.*

Below we shall discuss several important types of distributions for $(\mathbb{R}^d, \mathscr{B}^d)$-random variables, i.e., probability measures on $\mathscr{B}^d$ $(d \in \mathbb{N})$.

1. For every $x \in \mathbb{R}^d$ let ε_x denote the probability measure on $\mathscr{B}^d$ which is "point-mass 1 at x". Every random variable X having such a Dirac measure as distribution is said to be *singularly distributed* or *degenerate.* [In view of the introductory remarks of this section, "singular distribution" might mean "probability measure which is singular with respect to Lebesgue measure". The reader is alerted to this unfortunate terminological subtlety.] X is singularly distributed just exactly when it is almost surely constant.

It follows directly from the definition of variance, as well as from Chebyshev's inequality, that for an integrable real random variable X the condition $V(X) = 0$ is equivalent to the equality $X = E(X)$ holding almost everywhere. This in turn is equivalent to the statement

$$\operatorname{dist}(X) = \varepsilon_{E(X)},$$

in other words, that X is singularly distributed.

2. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^d$, $(\alpha_n)_{n\in\mathbb{N}}$ a sequence of non-negative real numbers with $\sum_{n\in\mathbb{N}} \alpha_n = 1$. Then

$$\mu := \sum_{n\in\mathbb{N}} \alpha_n \varepsilon_{x_n}$$

is a probability measure on $\mathscr{B}^d$. Every such distribution is called *discrete* and a random variable with such a distribution is said to be *discretely distributed.* In particular, every singular distribution ε_x is discrete.

Every discrete distribution on $\mathscr{B}^d$ is λ^d-*singular*, because it's supported on a countable subset $C \subset \mathbb{R}^d$ and for such sets $\lambda^d(C) = 0$ (cf. Definition 17.12 in MI).

In contrast to the λ^d-singular distributions on $\mathscr{B}^d$ are the λ^d-*continuous* ones.

3. Every λ^d-continuous probability measure μ on $\mathscr{B}^d$ is said to be *Lebesgue-continuous.* By the Radon-Nikodym theorem the Lebesgue-continuous probability measures are just the measures $\mu = f\lambda^d$ having Borel measurable densities $f \geq 0$ with $\int f\, d\lambda^d = 1$. This f, which is λ^d-almost everywhere uniquely determined by μ (cf. Theorem 17.11 in MI), is then called the *probability density* of μ.

For the real line ($d = 1$) we shall now discuss some special discrete and Lebesgue-continuous distributions.

4. Let p be a real number in $[0,1]$ and $q := 1 - p$. According to the binomial theorem

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1\,,$$

and therefore for every $n \in \mathbb{N}$

$$B(n;p) := \beta_n^p := \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \varepsilon_k \tag{4.1}$$

defines a discrete distribution on $\mathscr{B}^1$. Two notations are prevalent. The β and the B should suggest *binomial* or *Bernoulli distribution*, with the parameters n and p. Note that

$$B(n;0) = \varepsilon_0 \quad \text{and} \quad B(n;1) = \varepsilon_n \qquad (n \in \mathbb{N}). \tag{4.2}$$

For every β_n^p-distributed real random variable X, (3.7) and a few easy calculations yield

$$E(X) = \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} = np(p+q)^{n-1} = np\,; \tag{4.3}$$

$$E(X^2) = \sum_{k=1}^{n} k^2 \binom{n}{k} p^k q^{n-k} = np \sum_{k=1}^{n} k \binom{n-1}{k-1} p^{k-1} q^{n-k}$$

$$= np[(n-1)p + 1] = np(np+q) \tag{4.4}$$

and thus via (3.17)

$$V(X) = npq\,. \tag{4.5}$$

In particular, $V(X) = 0$ if and only if either $p = 0$ or $q = 1$, that is, if and only if we are in the situation (4.2).

In §2 part **1(b)** (m draws with replacement) we have already encountered the distribution $B(m;p)$. Let, namely, $(\Omega, \mathscr{A}, P)$ be the probability space considered there, the product of m copies of the Laplace probability space $(\Omega_0, \mathscr{A}_0, P_0)$. The elements of Ω_0 represent the b black and $n - b$ white balls in the urn. To every element $(\omega_1, \ldots, \omega_m) \in \Omega$, that is, to every series of m successive draws with replacement, we assign the number $X(\omega_1, \ldots, \omega_m)$ of black balls among $\{\omega_1, \ldots, \omega_m\}$. Then X is a real random variable on $(\Omega, \mathscr{A}, P)$ with the values $0, 1, \ldots, m$. According to the analysis carried out in §2, **1(b)** it has the distribution $B(m; b/n)$.

We will soon encounter further applications of the binomial distribution.

For practical applications of the binomial distribution a serious obstacle is the difficulty of calculating binomial coefficients, since buried in them is the rapidly

"exploding" factorial $n!$. Useful for coping with this difficulty is *Stirling's formula*, which says that for each $n \in \mathbb{N}$ there is a number $\theta(n) \in]0,1[$ such that

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{\theta(n)}{12n}}. \tag{4.6}$$

(Cf. STROMBERG [1981], p. 253.) Since

$$\lim_{n\to\infty} e^{\frac{\theta(n)}{12n}} = 1,$$

from (4.6) follows the asymptotical form

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \tag{4.6'}$$

of Stirling's formula. It is easy to derive the less precise form of Stirling's formula asserting that there is a positive constant α such that

$$n! = \alpha\sqrt{n}\left(\frac{n}{e}\right)^n e^{\frac{\theta(n)}{12n}} \tag{4.6''}$$

holds for some $\theta(n) \in]0,1[$ and every $n \in \mathbb{N}$ (cf. Exercise 2 below). A surprisingly simple probabilistic derivation of (4.6'), which also determines the constant $\alpha = \sqrt{2\pi}$, will be offered in §27 in connection with the Central Limit Theorem. (On the question of determining α compare also Exercise 3 below.)

5. If X is a $B(n;p)$-distributed random variable on a probability space $(\Omega, \mathscr{A}, P)$, then $X' := 2X - n$ defines another random variable on $(\Omega, \mathscr{A}, P)$. According to the rules for transforming image measures, $\mathrm{dist}(X')$ is the image of $\mathrm{dist}(X) = B(n;p)$ under the mapping $x \mapsto 2x - n$. Consequently, X' has the distribution

$$B_s(n;p) := \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \varepsilon_{-n+2k}. \tag{4.7}$$

Since only the values $0, 1, \ldots, n$ can be taken on with positive probability by X, the corresponding fact prevails for X' with respect to $-n, -n+2, \ldots, n-2, n$.

$B_s(n;p)$ is called the *symmetrical binomial distribution*. It will soon serve us well. Notice that for $n = 1$

$$B_s(1;p) = q\varepsilon_{-1} + p\varepsilon_1. \tag{4.7'}$$

6. The equality

$$e^\alpha = \sum_{k=0}^{\infty} \frac{\alpha^k}{k!}$$

shows that for every $\alpha \geq 0$

$$\pi_\alpha := \sum_{k=0}^{\infty} e^{-\alpha} \frac{\alpha^k}{k!} \varepsilon_k \tag{4.8}$$

defines a discrete probability measure on $\mathscr{B}^1$, which for $\alpha = 0$ coincides with ε_0. When $\alpha > 0$, we call π_α the *Poisson distribution* with parameter α. Again from (3.7) it follows that for every π_α-distributed random variable X

$$E(X) = \sum_{k=0}^{\infty} e^{-\alpha} \frac{\alpha^k}{k!} k = \alpha; \tag{4.9}$$

$$E(X^2) = \sum_{k=0}^{\infty} e^{-\alpha} \frac{\alpha^k}{k!} k^2 = \sum_{k=1}^{\infty} e^{-\alpha} \frac{\alpha^k}{(k-1)!}(k-1+1) = \alpha^2 + \alpha; \tag{4.10}$$

$$V(X) = \alpha. \tag{4.11}$$

The remaining examples deal with Lebesgue-continuous distributions.

7. From the well-known (cf. (16.10′) in MI) identity

$$(2\pi)^{-1/2} \int e^{-x^2/2}\, dx = 1$$

via a simple substitution it follows that for every $\alpha \in \mathbb{R}$ and every real $\sigma > 0$ the equation

$$g_{\alpha,\sigma^2}(x) := (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x-\alpha)^2}{2\sigma^2}} \tag{4.12}$$

defines a probability density on $\mathbb{R}$ and therefore

$$N(\alpha, \sigma^2) := \nu_{\alpha,\sigma^2} := g_{\alpha,\sigma^2} \lambda^1 \tag{4.13}$$

defines a probability measure on $\mathscr{B}^1$. It is called the *normal* or *Gauss distribution* on $\mathbb{R}$ with parameters α and σ^2. Here too, as with the binomial distribution, two names and two notations are prevalent. $N(0,1)$ is called the *standard* normal distribution.

The graph of the function g_{α,σ^2} is the familiar *bell-shaped* curve [so named by C.F. GAUSS (1777–1855)]. A simple analysis of the function g_{α,σ^2} reveals that it attains its maximum value of $(2\pi\sigma^2)^{-1/2}$ at the point $x = \alpha$ and only there, and that its only inflection points are $x = \alpha \pm \sigma$. The parameter σ^2 is therefore a measure of the breadth of the bell-shaped curve. Both parameters also have immediate probability-theoretic significance. Namely, any $N(\alpha, \sigma^2)$-distributed real random variable X satisfies

$$E(X) = \int x g_{\alpha,\sigma^2}(x)\, dx = \alpha; \tag{4.14}$$

$$E(X^2) = \int x^2 g_{\alpha,\sigma^2}(x)\, dx = \sigma^2 + \alpha^2; \tag{4.15}$$

$$V(X) = \sigma^2. \tag{4.16}$$

The derivation of (4.14) and (4.15) — from them (4.16) is immediate — is reduced via the linear transformation

$$x \mapsto T(x) := \sigma x + \alpha$$

to the case of the standard normal distribution $N(0,1)$. For

$$T(\nu_{0,1}) = \nu_{\alpha,\sigma^2} \tag{4.17}$$

and in particular (cf. MI, (24.10))

$$\nu_{\alpha,\sigma^2} = \varepsilon_\alpha * \nu_{0,\sigma^2}\,. \tag{4.17'}$$

But first we have to note that every $N(0,1)$-distributed (whence every $N(\alpha,\sigma^2)$-distributed) real random variable X is p^{th}-*power integrable* for every $p \geq 1$. For this we use the inequalities

$$\frac{1}{k!}t^k < e^t\,, \qquad t > 0, \quad k = 0,1,2,\ldots$$

which follow from a glance at the power series development of the exponential function. They imply that for every $p \geq 0$

$$|x|^p\, e^{-x^2/2} < 2^k k!\, |x|^{p-2k} \qquad (x \neq 0).$$

Since $x \mapsto |x|^\alpha$ is λ^1-integrable over $\mathbb{R} \setminus [-1,1]$ for each $\alpha < -1$, the λ^1-integrability over $\mathbb{R}$ of the function $x \mapsto |x|^p\, g_{0,1}(x)$ for each $p \geq 0$ follows upon choosing $k \in \mathbb{N}$ sufficiently large.

In particular, for $N(0,1)$-distributed random variables X the expected values

$$M_n := E(X^n) = \int x^n g_{0,1}(x)\, dx \tag{4.18}$$

exist for every integer $n \geq 0$. They are easy to compute. First of all

$$M_{2k-1} = 0 \qquad (k \in \mathbb{N}), \tag{4.19}$$

since the integrands involved then are odd functions. This covers (4.14), which is thus confirmed. For even $n \geq 2$ the integral M_n can be expressed in terms of M_{n-2}. Since all the integrands that intervene are continuous functions, all integrals also exist as absolutely convergent Riemann integrals. Hence we can employ integration by parts to get

$$\int x^n g_{0,1}(x)\, dx = [-x^{n-1} g_{0,1}(x)]_{-\infty}^{+\infty} + (n-1)\int x^{n-2} g_{0,1}(x)\, dx\,,$$

$g_{0,1}$ being obviously an anti-derivative of the function $x \mapsto -xg_{0,1}(x)$. What results is the recursion formula

$$M_{2k} = (2k-1)M_{2k-2} \qquad (k \in \mathbb{N}).$$

Since $M_0 = 1$, it follows that

$$M_{2k} = 1 \cdot 3 \cdot \ldots \cdot (2k-1) \qquad (k = 0, 1, 2, \ldots) \tag{4.20}$$

and in particular, (4.15) is confirmed.

For an $N(\alpha, \sigma^2)$-distributed random variable X the sequence (M_n) can be used to express

$$E(X^n) = \sum_{k=0}^{n} \binom{n}{k} \sigma^k \alpha^{n-k} M_k \qquad (n \in \mathbb{N}). \tag{4.21}$$

We have only to use the transformation $T(x) = \sigma x + \alpha$ again and note that

$$E(X^n) = \int x^n \nu_{\alpha,\sigma^2}(dx) = \int x^n T(\nu_{0,1})(dx) = \int (\sigma x + \alpha)^n \nu_{0,1}(dx)\,.$$

A probability estimate that is useful for many purposes is the following:

4.1 Lemma. *If X is a $N(0, \sigma^2)$-distributed random variable, then for all $\eta > 0$*

$$(2\pi)^{-1/2} \frac{\sigma\eta}{\sigma^2 + \eta^2} e^{-\eta^2/2\sigma^2} < P\{X \geq \eta\} < (2\pi)^{-1/2} \frac{\sigma}{\eta} e^{-\eta^2/2\sigma^2}\,. \tag{4.22}$$

Proof. It suffices to treat the case $\sigma = 1$. The general case follows upon replacing this $N(0,1)$-distributed random variable X with σX, which amounts to replacing η with η/σ. First, we have

$$P\{X \geq \eta\} = P_X([\eta, +\infty[) = \nu_{0,1}([\eta, +\infty[) = (2\pi)^{-1/2} \int_{\eta}^{\infty} e^{-x^2/2}\, dx\,.$$

The right-most inequality in (4.22) then follows from

$$\int_{\eta}^{\infty} e^{-x^2/2}\, dx < \int_{\eta}^{\infty} \frac{x}{\eta} e^{-x^2/2}\, dx = \frac{1}{\eta} \int_{\eta}^{\infty} x e^{-x^2/2}\, dx = \frac{1}{\eta} e^{-\eta^2/2}\,,$$

$-g_{0,1}$ being an anti-derivative of the last integrand. Via partial integration we get

$$\int_{\eta}^{\infty} x^{-2} e^{-x^2/2}\, dx = \frac{1}{\eta} e^{-\eta^2/2} - \int_{\eta}^{\infty} e^{-x^2/2}\, dx\,;$$

from which

$$\frac{1}{\eta}e^{-\eta^2/2} = \int_\eta^\infty (1+x^{-2})e^{-x^2/2}\,dx < \int_\eta^\infty (1+\eta^{-2})e^{-x^2/2}\,dx\,,$$

that is,

$$\frac{\eta}{1+\eta^2}e^{-\eta^2/2} < \int_\eta^\infty e^{-x^2/2}\,dx\,.$$

The left-most inequality in (4.22) follows from this and our first chain of equalities. □

8. The function

$$x \mapsto \frac{\alpha}{\pi}(\alpha^2+x^2)^{-1}$$

is also a probability density, denoted c_α, on $\mathbb{R}$, for every $\alpha > 0$, since

$$\int_{-\infty}^{+\infty} (1+x^2)^{-1}\,dx = \lim_{n\to\infty}[\arctan(x)]_{-n}^{+n} = \pi\,.$$

The probability measure

$$\gamma_\alpha := c_\alpha \lambda^1 \tag{4.23}$$

is called the (standard) *Cauchy distribution* with parameter $\alpha > 0$. It is easy to calculate directly that an expected value does not exist for any real random variable X which is γ_α-distributed. It suffices to note that for any $t \in \mathbb{R}_+$

$$\int_0^t \frac{x}{1+x^2}\,dx = \frac{1}{2}\log(1+t^2)$$

and consequently

$$E(X^+) = E(X^-) = \int_0^{+\infty} \frac{x}{1+x^2}\,dx = +\infty\,,$$

that is, X is not even quasi-integrable.

We close with an important class of Lebesgue-continuous probability measures on the σ-algebra $\mathscr{B}^2$ of Borel subsets of $\mathbb{R}^2$.

9. For real numbers $\varrho \in]-1,+1[$ and $\sigma > 0$ consider the continuous function $f : \mathbb{R}^2 \to]0,+\infty[$ defined by

$$f(x,y) := \frac{(1-\varrho^2)^{1/2}}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2}(x^2 - 2\varrho xy + y^2)\right]. \tag{4.25}$$

Evidently,

$$f(x,y) = \frac{(1-\varrho^2)^{1/2}}{2\pi\sigma^2} \exp\left(-\frac{1-\varrho^2}{2\sigma^2}y^2\right) \exp\left[-\frac{1}{2\sigma^2}(x-\varrho y)^2\right], \tag{4.25'}$$

from which follows (cf. **7** above)

$$\int f(x,y)\,dx = \left(\frac{1-\varrho^2}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1-\varrho^2}{2\sigma^2}y^2\right) \tag{4.26}$$

and therewith also $\int f\,d\lambda^2 = 1$. Consequently,

$$\mu_{\varrho,\sigma} := f\lambda^2 \tag{4.27}$$

defines a probability measure. It is called the 2-*dimensional standard normal distribution* with parameters $\varrho \in]-1,+1[$ and $\sigma > 0$.

We will study two- and multi-dimensional normal distributions in §30. (Cf., in particular, Example 2 of §30.)

Exercises

1. Show that a real random variable X is singularly distributed if and only if the probability $P\{X \le \alpha\}$ equals 0 or 1 for every real α.

2. The sequence of positive numbers

$$a_n := n!\left(\frac{e}{n}\right)^n n^{-1/2}$$

satisfies

$$\log\left(\frac{a_n}{a_{n+1}}\right) = \left(n+\frac{1}{2}\right)\log\left(\frac{n+1}{n}\right) - 1.$$

For $0 < x < 1$ the power series of the logarithm yields

$$\begin{aligned} 2x < \log\left(\frac{1+x}{1-x}\right) &= 2\sum_{n=0}^{\infty}\frac{x^{2n+1}}{2n+1} < 2x + \frac{2}{3}x^2\sum_{n=0}^{\infty}x^{2n+1} \\ &= 2x + \frac{x^2}{3}\left(\frac{1}{1-x} - \frac{1}{1+x}\right), \end{aligned}$$

and consequently for all $n \in \mathbb{N}$ (and $x := (2n+1)^{-1}$)

$$0 < \log\left(\frac{a_n}{a_{n+1}}\right) < \frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right).$$

Deduce that

$$0 < \log\left(\frac{a_n}{a_{n+k}}\right) < \frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+k}\right) \qquad (k, n \in \mathbb{N})$$

and go on to prove the existence of a positive number α to which the a_n decrease, and finally deduce (4.6″). (Cf. VAN DER WAERDEN [1936].)

3. With the help of Wallis' formula

$$\frac{\pi}{2} \prod_{n=1}^{\infty} \frac{4n^2}{4n^2-1} = \lim_{m\to\infty} \frac{2^{4m}(m!)^4}{((2m)!)^2(2m+1)}$$

(cf. STROMBERG [1981], p. 250) show that the constant α from Exercise 2 is $\sqrt{2\pi}$.

4. For an $N(0,\sigma^2)$-distributed random variable X derive from (4.22) the estimates

$$P\{X \geq \eta\} < (2/\pi)^{1/2} e^{-1} \frac{\sigma^3}{\eta^3} \quad \text{and} \quad (2\pi)^{-1/2} \frac{\sigma}{2\eta} e^{-\eta^2/2\sigma^2} < P\{X \geq \eta\},$$

valid for all $\eta > 0$ and all $\eta \geq \sigma > 0$, respectively. Show further that $P\{X \geq \eta\}$ is asymptotically equal to $(2\pi)^{-1/2}\sigma\eta^{-1}e^{-\eta^2/2\sigma^2}$ as $\eta \to +\infty$.
Hint: For all $x \in \mathbb{R}$, $xe^{1-x} \leq 1$.

5. Let b, m, n be integers such that $1 \leq b < n$ and $m \leq n$. Prove that:

(a) $$\eta_{n,b,m} := \binom{n}{m}^{-1} \sum_{k=0}^{m} \binom{b}{k}\binom{n-b}{m-k}\varepsilon_k$$

is a probability measure on $\mathscr{B}^1$. It is called the *hypergeometric distribution* with parameters n, b and m.
(b) $\eta_{n,b,m}$ is the distribution of a random variable on the probability space treated in §2, section **1(a)** (drawing without replacement). Compare this result with the probability-theoretical interpretation of the binomial distribution $B(n;p)$ given in **4.** of this section.

6. Let the real random variable X have the hypergeometric distribution $\eta_{n,b,m}$ of Exercise 5. Then

$$E(X) = mp \quad \text{and} \quad V(X) = \frac{n-m}{n-1} mqp,$$

where $p := \dfrac{b}{n}$ and $q := 1 - p$.

7. For every finite subset $\{p_1, \ldots, p_d\}$ of $\mathbb{R}_+$ with $p_1 + \ldots + p_d = 1$ and for every $n \in \mathbb{N}$ the sum

$$\sum \frac{n!}{n_1! \ldots n_d!} p_1^{n_1} \cdot \ldots \cdot p_d^{n_d} \varepsilon_{(n_1,\ldots,n_d)}$$

extended over all integers $n_1, \ldots, n_d \geq 0$ satisfying $n_1 + \ldots + n_d = n$, defines a discrete distribution on $\mathscr{B}^d$ — called a *multinomial distribution*.

§5. Convergence of random variables and distributions

From integration theory we know three different modes of convergence for sequences of measurable real functions. These also play an important role in probability theory. Let us recall — in the context of probability theory — these three convergence notions. To that end, $(\Omega, \mathscr{A}, P)$ will be a probability space, $(X_n)_{n\in\mathbb{N}}$ a sequence of real random variables, X a further real random variable, all defined on $(\Omega, \mathscr{A}, P)$.

Almost sure convergence: For the almost sure convergence of (X_n) to X each of the conditions

$$\lim_{n\to\infty} P\{\sup_{m\geq n} |X_m - X| > \varepsilon\} = 0 \qquad \text{(for all } \varepsilon > 0) \tag{5.1}$$

and

$$P(\limsup_{n\to\infty}\{|X_n - X| > \varepsilon\}) = 0 \qquad \text{(for all } \varepsilon > 0) \tag{5.2}$$

is necessary and sufficient (cf. MI, Lemma 20.6). The first of these can be reworded as

$$\lim_{n\to\infty} P\{|X_n - X| > \varepsilon \text{ for some } m \geq n\} = 0 \qquad \text{(for all } \varepsilon > 0) \tag{5.1'}$$

as well as

$$\lim_{n\to\infty} P\{|X_m - X| \leq \varepsilon \text{ for all } m \geq n\} = 1 \qquad \text{(for all } \varepsilon > 0). \tag{5.1''}$$

The notation for $P(\limsup E_n)$ introduced at the end of §1 permits (5.2) to be re-phrased in the suggestive forms

$$P\{|X_n - X| > \varepsilon \text{ for infinitely many } n\} = 0 \qquad \text{(for all } \varepsilon > 0) \tag{5.2}$$

$$P\{|X_n - X| > \varepsilon \text{ i.o.}\} = 0 \qquad \text{(for all } \varepsilon > 0) \tag{5.2'}$$

$$P\{|X_n - X| \leq \varepsilon \text{ for almost all } n\} = 1 \qquad \text{(for all } \varepsilon > 0). \tag{5.2''}$$

In (11.2) we will become acquainted with a useful sufficient condition for almost sure convergence.

$\mathscr{L}^p$-convergence: The sequence (X_n) is said to converge to X in p^{th} mean (or to be $\mathscr{L}^p$-convergent to X) when

$$\lim_{n\to\infty} E(|X_n - X|^p) = 0\,. \tag{5.3}$$

Here $1 \leq p < +\infty$. Because of the inequality (3.25), $\mathscr{L}^p$-convergence for such a p always entails $\mathscr{L}^1$-convergence, that is, *convergence in mean.*

Stochastic convergence: The sequence (X_n) is said to converge stochastically to X or to *converge to X in probability*, which we write $P\text{-}\lim_{n\to\infty} X_n = X$, if

$$\lim_{n\to\infty} P\{|X_n - X| \geq \varepsilon\} = 0 \qquad \text{(for all } \varepsilon > 0\text{);} \tag{5.4}$$

equivalently, if

$$\lim_{n\to\infty} P\{|X_n - X| > \varepsilon\} = 0 \qquad \text{(for all } \varepsilon > 0\text{).} \tag{5.4'}$$

Stochastic convergence of (X_n) to X follows from almost sure convergence, as well as from $\mathscr{L}^p$-convergence. (On this point compare Theorem 20.4 and Theorem 20.5 of MI.) Comparison of (5.1) and (5.4') constitutes a direct proof of the first of these implications. The second follows from the *Chebyshev-Markov inequality* (cf. MI, (20.1)), which can be written in the form

$$P\{|X_n - X| \geq \varepsilon\} \leq \varepsilon^{-p} E(|X_n - X|^p)\,. \tag{5.5}$$

Because of the significance of the distribution of a random variable for its probabilistic occurrence, the question suggests itself whether these three kinds of convergence of (X_n) to X imply any kind of convergence of the distributions P_{X_n} to the distribution P_X. Of course, we have to clarify what is meant by convergence of distributions. This new concept comes up in the next theorem. Let us denote by $C_b(\mathbb{R}^d)$ the vector space of all *bounded, continuous, real-valued* functions on $\mathbb{R}^d$; the case $d = 1$ of the number line will be our initial concern.

5.1 Theorem. *Suppose the sequence $(X_n)_{n\in\mathbb{N}}$ of real random variables on the probability space $(\Omega, \mathscr{A}, P)$ converges stochastically to a real random variable X on Ω. Then the corresponding sequence of distributions $(P_{X_n})_{n\in\mathbb{N}}$ converges weakly to the distribution P_X; that is,*

$$\lim_{n\to\infty} \int f\, dP_{X_n} = \int f\, dP_X \tag{5.6}$$

or equivalently (see (3.7))

$$\lim_{n\to\infty} E(f \circ X_n) = E(f \circ X) \tag{5.6'}$$

for every $f \in C_b(\mathbb{R})$.

If X is almost surely constant, so that P_X is a Dirac measure, then the converse implication holds.

Proof. First we will consider only $f \in C_b(\mathbb{R})$ which are uniformly continuous on $\mathbb{R}$. Thus to each $\varepsilon > 0$ corresponds a $\delta > 0$ such that

$$x', x'' \in \mathbb{R} \quad \& \quad |x' - x''| < \delta \quad \Rightarrow \quad |f(x') - f(x'')| < \varepsilon\,.$$

For the events $A_n := \{|X_n - X| \geq \delta\}$ $(n \in \mathbb{N})$ we then have the inequalities

$$\begin{aligned}\left|\int f\,dP_{X_n} - \int f\,dP_X\right| &= |E(f \circ X_n - f \circ X)| \leq E(|f \circ X_n - f \circ X|) \\ &= E(|f \circ X_n - f \circ X|\,; A_n) + E(|f \circ X_n - f \circ X|\,; \complement A_n) \\ &\leq 2\,\|f\|\,P(A_n) + \varepsilon P(\complement A) \leq 2\,\|f\|\,P(A_n) + \varepsilon\,,\end{aligned}$$

in which

$$\|f\| := \sup\{|f(x)| : x \in \mathbb{R}\}$$

denotes the *supremum norm* of f, and the trivial estimates

$$|f \circ X_n - f \circ X| \leq |f \circ X_n| + |f \circ X| \leq 2\,\|f\|\,,$$

as well as the notation (3.14) have come into play. From this chain of inequalities follows (5.6), because by definition of stochastic convergence $(P(A_n))$ is a null sequence.

Now let us deal with arbitrary $f \in C_b(\mathbb{R})$. The intervals $I_n := [-n, n]$ increase to $\mathbb{R}$, so $P_X(I_n) \uparrow 1$. For $\varepsilon > 0$ an $n_\varepsilon \in \mathbb{N}$ can be chosen so that

$$1 - P_X(I_{n_\varepsilon}) = P_X(\mathbb{R} \setminus I_{n_\varepsilon}) \leq \varepsilon\,.$$

Let u_ε be the function in $C_b(\mathbb{R})$ which equals 1 on I_{n_ε}, vanishes off $I_{n_\varepsilon+1}$ and is affine on $[n_\varepsilon, n_\varepsilon + 1]$ and on $[-n_\varepsilon - 1, -n_\varepsilon]$. If we set $f' := u_\varepsilon f$, then both functions u_ε and f' are uniformly continuous on I_{n_ε} and vanish identically in $\complement I_{n_\varepsilon}$, hence they are uniformly continuous on $\mathbb{R}$. From the first part of our proof we therefore know that

$$\lim_{n\to\infty} \int f'\,dP_{X_n} = \int f'\,dP_X \tag{5.7}$$

and

$$\lim_{n\to\infty} \int u_\varepsilon\,dP_{X_n} = \int u_\varepsilon\,dP_X \tag{5.8}$$

hence also, since P_{X_n} and P_X are each probability measures,

$$\lim_{n\to\infty} \int (1 - u_\varepsilon)\,dP_{X_n} = \int (1 - u_\varepsilon)\,dP_X\,. \tag{5.8'}$$

The rest of the proof now follows from a trivial consequence of the triangle inequality:

$$\begin{aligned}&\left|\textstyle\int f\,dP_{X_n} - \int f\,dP_X\right| \\ &\quad \leq \textstyle\int |f - f'|\,dP_{X_n} + \left|\int f'\,dP_{X_n} - \int f'\,dP_X\right| + \int |f' - f|\,dP_X\end{aligned} \tag{5.9}$$

via these considerations: The inequality $0 \leq 1 - u_\varepsilon \leq 1_{\mathbb{R}\setminus I_{n_\varepsilon}}$ implies that

$$\int (1 - u_\varepsilon)\, dP_X \leq P_X(\mathbb{R} \setminus I_{n_\varepsilon}) < \varepsilon\,,$$

so that on account of (5.8′)

$$\int (1 - u_\varepsilon)\, dP_{X_n} \leq \varepsilon$$

for almost all n, say for all $n \geq N_\varepsilon$. From this follows on the one hand

$$\int |f - f'|\, dP_X = \int |f|\,(1 - u_\varepsilon)\, dP_X \leq \|f\|\,\varepsilon\,,$$

and on the other

$$\int |f - f'|\, dP_{X_n} \leq \|f\| \int (1 - u_\varepsilon)\, dP_{X_n} \leq \|f\|\,\varepsilon$$

for all $n \geq N_\varepsilon$. Considering (5.7), it therefore indeed follows that the right-hand side of (5.9) is smaller than $2\,\|f\|\,\varepsilon + \varepsilon$ for all sufficiently large n, and this proves (5.6).

For the proof of the converse, let $X = \eta \in \mathbb{R}$ almost surely; thus $P_X = \varepsilon_\eta$. Given $\varepsilon > 0$, choose for the interval $I :=]\eta - \varepsilon, \eta + \varepsilon[$ a function (say, piecewise affine) $f \in C_b(\mathbb{R})$ which satisfies $f(\eta) = 1$ and $f \leq 1_I$. Then

$$\int f\, dP_{X_n} \leq P_{X_n}(I) = P\{X_n \in I\} \leq 1 \qquad (n \in \mathbb{N}).$$

By hypothesis $\int f\, dP_{X_n}$ tends to $f(\eta) = 1$. As $n \to \infty$ it therefore follows from the preceding inequality that

$$\lim_{n\to\infty} P\{X_n \in I\} = 1\,. \tag{5.10}$$

If we take account of the fact that $\{X_n \in I\} = \{|X_n - \eta| < \varepsilon\}$ and consequently

$$P\{|X_n - X| \geq \varepsilon\} = P\{|X_n - \eta| \geq \varepsilon\} = 1 - P\{X_n \in I\}\,,$$

then (5.10) just says that

$$\lim_{n\to\infty} P\{|X_n - X| \geq \varepsilon\} = 0\,,$$

holding for every $\varepsilon > 0$; that is, (X_n) converges stochastically to X. □

The convergence concept expressed by (5.6) will now be codified into a definition. This is to be found under more general hypotheses as Definition 30.7 in MI.

5.2 Definition. A sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability measures on $\mathscr{B}^d$ *converges weakly* to a probability measure μ on $\mathscr{B}^d$ if

$$\lim_{n\to\infty} \int f\, d\mu_n = \int f\, d\mu$$

for all functions $f \in C_b(\mathbb{R}^d)$. We write then

$$\lim_{n\to\infty} \mu_n = \mu\,. \tag{5.11}$$

If $X, X_1, X_2, \ldots$ are random variables on a probability space $(\Omega, \mathscr{A}, P)$ with values in $\mathbb{R}^d$ and the sequence $(P_{X_n})_{n\in\mathbb{N}}$ of their distributions converges weakly to the distribution P_X of X, or even more generally to a probability measure ν on $\mathscr{B}^d$, then the sequence (X_n) is said to *converge in distribution* to X, or to ν. [That X need not be unique is illustrated in the third example below.]

Remark. 1. Conditions (5.4) and (5.4′) make sense for $\mathbb{R}^d$-valued random variables X_n and X, if $|\ |$ is interpreted as the euclidean norm in $\mathbb{R}^d$. They define the concept of *stochastic convergence* for $\mathbb{R}^d$-valued random variables. Theorem 5.1, with $f \in C_b(\mathbb{R}^d)$, is valid for them too. The first paragraph of its proof needs no change. In the second, the sets I_n are defined as the balls $\{x \in \mathbb{R}^d : |x| \le n\}$, rather than intervals. Finally, the auxiliary function u_ε introduced there is to be replaced by the radial function constructed from it via $x \mapsto u_\varepsilon(|x|)$. Since (X_n) converges to X almost surely (resp., in $\mathscr{L}^p$-norm) if and only if the scalar sequence $(|X_n - X|)$ converges to 0 almost surely (resp., in $\mathscr{L}^p$-norm), it again follows from Theorems 20.4 and 20.5 of MI that these modes of convergence for $\mathbb{R}^d$-valued random variables imply stochastic convergence for them. These multidimensional generalizations will not be needed until the proof of Theorem 43.5.

Examples. 1. Let (x_n) be a sequence of real numbers. It converges to a real number x_0 if and only if the associated sequence (ε_{x_n}) of Dirac measures converges weakly to ε_{x_0}: From $\lim x_n = x_0$ follows of course $\lim f(x_n) = 0$ for every $f \in C_b(\mathbb{R})$, that is, the weak convergence of (ε_{x_n}) to ε_{x_0}. For the converse, consider for $\varepsilon > 0$ the function $f_\varepsilon \in C_b(\mathbb{R})$ defined by

$$f_\varepsilon(x) := \max\left(0, 1 - \varepsilon^{-1}\,|x - x_0|\right)\,.$$

Since $\{f_\varepsilon > 0\} =]x_0 - \varepsilon, x_0 + \varepsilon[$ and $\lim f_\varepsilon(x_n) = f_\varepsilon(x_0) = 1$, it must be that $x_n \in]x_0 - \varepsilon, x_0 + \varepsilon[$ for almost all n.

2. For every sequence (σ_n) of positive real numbers converging to 0,

$$\lim_{n\to\infty} N(0, \sigma_n^2) = \varepsilon_0\,.$$

Upon making the obvious extension of the concept of weak convergence to families (μ_t) of probability measures on $\mathscr{B}^1$ indexed by a real parameter t, we even have

$$\lim_{\sigma\to 0,\, \sigma>0} N(0, \sigma^2) = \varepsilon_0\,. \tag{5.12}$$

In fact, using the substitution $x = \sigma y$ ($\sigma > 0$), we get for each $f \in C_b(\mathbb{R})$

$$\int f\, d\nu_{0,\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} f(x) e^{-x^2/2\sigma^2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(\sigma y) e^{-y^2/2}\, dy\,.$$

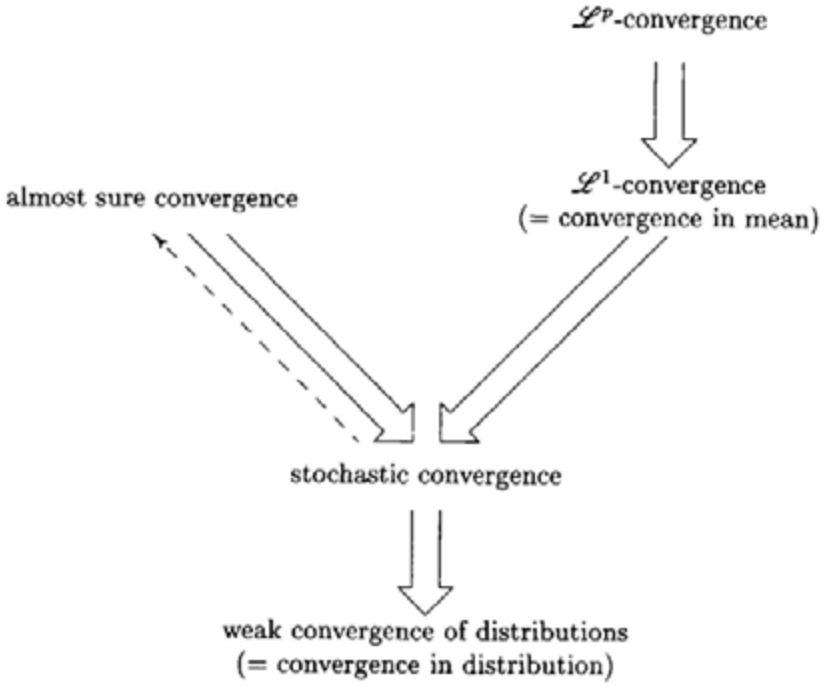
The integrand is majorized by the integrable function $y \mapsto \|f\|\, e^{-y^2/2}$, independent of σ. Therefore, from the Dominated Convergence Theorem, follows

$$\lim_{n\to\infty} \int f\, d\nu_{0,\sigma_n^2} = f(0)\,.$$

3. Consider the probability space $(\Omega, \mathscr{A}, P)$ with $\Omega := [0,1]$, $\mathscr{A} := [0,1] \cap \mathscr{B}^1$ and $P := \lambda^1_{[0,1]}$, and on it the constant sequence of random variables $X_n := 1_{[1/2,1]}$ together with $X := 1_{[0,1/2[}$. Evidently $P_{X_n} = P_X = \frac{1}{2}(\varepsilon_0 + \varepsilon_1)$, so the sequence (X_n) converges in distribution to X as well as to X_1. Since $|X_n - X| = |X_1 - X| = 1_{[0,1]}$, $X(\omega) = X_1(\omega)$ holds for no $\omega \in \Omega$ and $P\{|X_n - X| \geq 1\} = 1$ for every $n \in \mathbb{N}$, showing that generally stochastic convergence does not follow from convergence in distribution and that the latter does not almost surely determine its limit.

This shows that the supplemental hypothesis "X almost surely constant" imposed in the converse part of Theorem 5.1 cannot simply be dispensed with. (This phenomenon is further illustrated by Example 2 in §7.)

The diagram below (valid for $\mathbb{R}^d$-valued functions too) illustrates once more the relationships between the four modes of convergence discussed in this section.



In it the broken arrow is to remind us that from a stochastically convergent sequence a *subsequence* can always be extracted which converges almost surely (cf. MI, Theorem 20.7).

Remarks. 2. Every probability measure on $\mathbb{R}$ can be fully described by its distribution function (cf. MI, Theorem 6.6). In MI Theorem 30.13 and Exercise 7, §30 it is shown in terms of what convergence behavior of the distribution functions weak convergence can be described. In this connection see also Exercise 3 below.

3. The main part of Theorem 5.1 is more that a mere example illustrating the concept of weak convergence. It really goes to the probabilistic heart of the concept as the following converse shows: Suppose given a sequence (μ_n) of probability measures on $\mathscr{B}^1$ which converges weakly to a probability measure μ on $\mathscr{B}^1$. Then there always exists a probability space $(\Omega, \mathscr{A}, P)$ and real random variables X, X_n $(n \in \mathbb{N})$ on it such that $\mu = P_X$, $\mu_n = P_{X_n}$ (for all n) and the sequence (X_n) converges stochastically to X. It is even possible to arrange that $\Omega = [0,1]$, $\mathscr{A} = \Omega \cap \mathscr{B}^1$ and $P = \lambda_\Omega^1$. This is the content of a theorem of Skorokhod and Dudley. For a proof see p. 10 of SKOROKHOD [1965], where the result is treated in the generality of random variables taking values in a Polish space. A remarkable "global" version of this theorem is proven in FERNIQUE [1988].

Exercises

1. Proceeding from (5.2″), formulate and prove a Cauchy criterion for almost sure convergence of real random variables.

2. The sequence (X_n) of real random variables on the probability space $(\Omega, \mathscr{A}, P)$ satisfies:

$$P\{|X_n| > \varepsilon\} < \varepsilon \qquad \text{for almost all } n,$$

and for every $\varepsilon > 0$. Is this equivalent to its converging weakly to $X := 0$?

3. Let F_σ denote the distribution function of the normal distribution $N(0, \sigma^2)$ for $\sigma > 0$ and F_0 the distribution function of ε_0. Show that

$$\lim_{\sigma \to 0,\, \sigma > 0} F_\sigma(x) = F_0(x)$$

for all $x \neq 0$ but not for $x = 0$. (On this point compare Remark 2.)

4. For the family $(\pi_\alpha)_{\alpha > 0}$ of Poisson distributions on $\mathbb{R}$ show that

$$\lim_{\alpha \to 0,\, \alpha > 0} \pi_\alpha = \varepsilon_0$$

in the sense of weak convergence. Is there a probability measure μ on $\mathscr{B}^1$ such that $\pi_\alpha \to \mu$ as $\alpha \to +\infty$?

5. By analyzing the proof of Theorem 5.1 show that a sequence (μ_n) of probability measures on $\mathscr{B}^1$ converges weakly to a probability measure μ on $\mathscr{B}^1$ if and only if $\lim \int f \, d\mu_n = \int f \, d\mu$ for all bounded uniformly continuous real-valued functions f on $\mathbb{R}$. (Readers familiar with MI will already have encountered this phenomenon in Exercise 10, §30 there.)