De Gruyter Textbook

Hans-Otto Georgii · Stochastics

Hans-Otto Georgii

Stochastics

Introduction to Probability and Statistics

2nd Revised and Extended Edition

Translated by
Marcel Ortgiese, Ellen Baake and the Author

De Gruyter

Prof. Dr. Hans-Otto Georgii
Mathematical Institute
Ludwig-Maximilians-Universität Munich
Theresienstr. 39
80333 Munich
Germany

Mathematics Subject Classification 2010: Primary: 60-01; Secondary: 62-01

ISBN 978-3-11-029254-1
e-ISBN 978-3-11-029360-9

Library of Congress Cataloging-in-Publication Data

A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Typesetting: Dimler & Albroscheit Partnerschaft, Müncheberg
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen
∞ Printed on acid-free paper

Printed in Germany

www.degruyter.com

Preface

Chance – or what appears to us as such – is ubiquitous. Not only in the games of chance such as lottery or roulette where risk is played with, but also in substantial parts of everyday life. Every time an insurance company uses the claim frequencies to calculate the future premium, or a fund manager the most recent stock charts to rearrange his portfolio, every time cars are jamming at a traffic node or data packages at an internet router, every time an infection spreads out or a bacterium turns into a resistant mutant, every time pollutant concentrations are measured or political decisions are based on opinion polls – in all such cases a considerable amount of randomness comes into play, and there is a need to analyse the random situation and to reach at rational conclusions in spite of the inherent uncertainty. Precisely this is the objective of the field of stochastics, the 'mathematics of chance'. Stochastics is therefore a highly applied science, which tries to solve concrete demands of many disciplines. At the same time, it is genuine mathematics – with sound systematics, clear-cut concepts, deep theorems and sometimes surprising cross-connections. This interplay between applicability on the one side and mathematical precision and elegance on the other makes up the specific appeal of stochastics, and a variety of natural questions determines its lively and broad development.

This book offers an introduction to the typical way of thinking, as well as the basic methods and results of stochastics. It grew out of a two-semester course which I gave repeatedly at the University of Munich. It is addressed to students of mathematics in the second year, and also to scientists and computer scientists who intend not only to apply stochastics, but also to understand its mathematical side. The two parts of stochastics – probability theory and statistics – are presented in two separate parts of the book because of their own scopes and methods, but are united under the same cover on purpose. For, the statistics is built on the concepts and methods of probability theory, whereas the latter needs the former as a bridge to reality. In the choice of the material I confined myself deliberately to the central subjects belonging to the standard curriculum of the corresponding mathematical courses. (It is thus unavoidable that some readers will miss their favourite subjects, e.g., the resampling methods of statistics.) The standard themes, however, are discussed with all necessary details. Rather than starting with discrete models I preferred to present (and motivate) the general measure theoretic framework right from the beginning, and some theoretical issues are also treated later as the case arises. In general, however, the measure theoretic apparatus is confined to what is absolutely necessary, and the emphasis is on the development of a stochastic intuition.

This text comprises a little more material than can be presented in a four-hour course over two semesters. The reader may thus want to make a selection. Several possibilities present themselves. For a first overview, the reader may concentrate on concepts, theorems, and examples and skip all proofs. In particular, this is a practicable route for non-mathematicians. A deeper understanding, of course, requires the study of a representative selection of proofs. On the other hand, a reader mainly interested in the theory and familiar with some applications may skip a portion of examples. For a short route through Part I leading directly to statistics, one can restrict oneself to the essentials of the first chapters up to Section 3.4, as well as Sections 4.1 and 4.3, and 5.1 and 5.2. The core of Part II consists of Sections 7.1 – 7.5, 8.1 – 8.2, 9.2, Chapter 10, as well as 11.2 and 12.1. Depending on the specific interests, it will facilitate orientation to browse through some portions of the text and return to them later when needed. A list of notational conventions can be found on page 395.

At the end of each chapter there is a collection of problems offering applications, additions, or supplements to the text. Their difficulty varies, but is not indicated because the reader should follow only his or her interests. The main point is to try for oneself. Nevertheless, this second English edition now provides draft solutions of selected problems, marked with ^{S}. These should be used for self-testing, rather than lulling the reader's willingness to tackle the problems independently.

As every textbook, this one grew out of more sources than can possibly be identified. Much inspiration came from the classical German texts of Ulrich Krengel [38] and Klaus Krickeberg and Herbert Ziezold [39], which had strong influence on the introductory courses in stochastics all over Germany. I also got many impulses from my Munich stochastics colleagues Peter Gänßler and Helmut Pruscha as well as all those responsible for the problem classes of my lectures during more than two decades: Peter Imkeller, Andreas Schief, Franz Strobl, Karin Münch-Berndl, Klaus Ziegler, Bernhard Emmer, and Stefan Adams. I am very grateful to all of them.

The English translation of the German original would not have appeared without the assistance of two further colleagues: Marcel Ortgiese accepted to lay the foundation by preparing an initial English version, so that I could concentrate on details and cosmetic changes. Ellen Baake took pleasure in giving a final polish to the English and suggesting numerous clarifications. I gratefully acknowledge their help.

Munich, June 2012 *Hans-Otto Georgii*

Contents

Preface v
Mathematics and Chance 1

I Probability Theory

1 Principles of Modelling Chance 7
1.1 Probability Spaces 7
1.2 Properties and Construction of Probability Measures 14
1.3 Random Variables 20
Problems 24

2 Stochastic Standard Models 27
2.1 The Uniform Distributions 27
2.2 Urn Models with Replacement 30
2.3 Urn Models without Replacement 35
2.4 The Poisson Distribution 39
2.5 Waiting Time Distributions 40
2.6 The Normal Distributions 46
Problems 48

3 Conditional Probabilities and Independence 51
3.1 Conditional Probabilities 51
3.2 Multi-Stage Models 57
3.3 Independence 64
3.4 Existence of Independent Random Variables, Product Measures 70
3.5 The Poisson Process 75
3.6 Simulation Methods 79
3.7 Tail Events 83
Problems 86

4 Expectation and Variance 92
4.1 The Expectation 92
4.2 Waiting Time Paradox and Fair Price of an Option 100
4.3 Variance and Covariance 107
4.4 Generating Functions 110
Problems 114
5 The Law of Large Numbers and the Central Limit Theorem 119
5.1 The Law of Large Numbers 119
5.2 Normal Approximation of Binomial Distributions 131
5.3 The Central Limit Theorem 138
5.4 Normal versus Poisson Approximation 143
Problems 146
6 Markov Chains 151
6.1 The Markov Property 151
6.2 Absorption Probabilities 155
6.3 Asymptotic Stationarity 159
6.4 Recurrence 171
Problems 181

II Statistics

7 Estimation 191
7.1 The Approach of Statistics 191
7.2 Facing the Choice 195
7.3 The Maximum Likelihood Principle 199
7.4 Bias and Mean Squared Error 205
7.5 Best Estimators 207
7.6 Consistent Estimators 214
7.7 Bayes Estimators 218
Problems 222
8 Confidence Regions 227
8.1 Definition and Construction 227
8.2 Confidence Intervals in the Binomial Model 233
8.3 Order Intervals 239
Problems 243

9 Around the Normal Distributions 246
9.1 The Multivariate Normal Distributions 246
9.2 The χ^2-, F- and t-Distributions 249
Problems 256
10 Hypothesis Testing 260
10.1 Decision Problems 260
10.2 Neyman–Pearson Tests 265
10.3 Most Powerful One-Sided Tests 271
10.4 Parameter Tests in the Gaussian Product Model 274
Problems 284
11 Asymptotic Tests and Rank Tests 289
11.1 Normal Approximation of Multinomial Distributions 289
11.2 The Chi-Square Test of Goodness of Fit 296
11.3 The Chi-Square Test of Independence 303
11.4 Order and Rank Tests 309
Problems 320
12 Regression Models and Analysis of Variance 325
12.1 Simple Linear Regression 325
12.2 The Linear Model 329
12.3 The Gaussian Linear Model 334
12.4 Analysis of Variance 342
Problems 351

Solutions 357
Tables 385
References 391
List of Notation 395
Index 399

Mathematics and Chance

What is stochastics[1]? In dictionaries of classical Greek one finds

στόχος	(stóchos)	goal, aim, guess, conjecture
στοχαστικός	(stochastikós)	skilful in aiming at, able to hit
στοχάζομαι	(stocházomai)	I aim at, guess at, infer, explore

Its current usage is captured by the sentence

Stochastics is the science of the rules of chance.

At first sight, this statement seems to be a contradiction in itself, since, in everyday life, one speaks of chance when no rules are apparent. A little thought, however, reveals that chance does indeed follow some rules. For example, if you flip a coin very often, you will have no doubt that heads will come up approximately half of the time. Obviously, this is a law of chance, and is generally accepted as such. Nevertheless, it is a widespread belief that such laws are too vague to be made precise, let alone to be formalised mathematically. Intriguingly, the opposite is true: Mathematics offers an exact language even for such seemingly disordered phenomena; a language that allows to state, and prove, a variety of laws obeyed by chance. The experience mentioned above, namely that heads shows up in about one half of a large number of coin flips, thus turns into a mathematical theorem, the law of large numbers. Stochastics is the part of mathematics that develops an appropriate formalism for describing the principles of randomness, for detecting rules where they seem to be absent, and for using them. This book presents the basic ideas and central results.

*

But what is chance? This is a philosophical question, and a satisfactory answer is still missing. Whether 'god plays dice with the universe' or, in fact, he does not (as was postulated by Albert Einstein in his famous dictum), whether randomness is only fictitious and merely a consequence of our incomplete knowledge or, on the contrary, is inherent to nature – these questions are still unresolved.

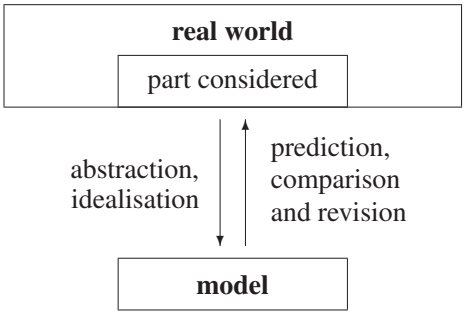
[1] So far, the noun 'stochastics' is not a well-established English word, in contrast to the adjective 'stochastic'. But since there is a need for a term comprising both probability theory and statistics, its use is spreading, and is expected to become standard, as it did in other languages. So we use it in this book.

As a matter of fact, however, we may refrain from trying to understand the nature of 'chance per se', by good reasons. We will never be able to investigate the universe as a whole. Rather, we will always have to focus on quite a special, restricted phenomenon, and we will restrict attention to the nature of this small part of reality. Even if the phenomenon considered could be explained in parts by its general circumstances (as we could, for instance, anticipate the number shown by a dice if we only knew in sufficient detail how it is thrown) – even then, it is much more practicable, and better adapted to our human perspective, if we adopt the viewpoint that the phenomenon is governed by chance. This kind of chance then comprises both, an indeterminacy possibly inherent to nature, and our (perhaps unavoidable) ignorance of the determining factors.

*

How does mathematics then come into play? As soon as a definite part of the real world is selected for investigation, one can try and collect all its relevant aspects into a mathematical model. Typically, this is done

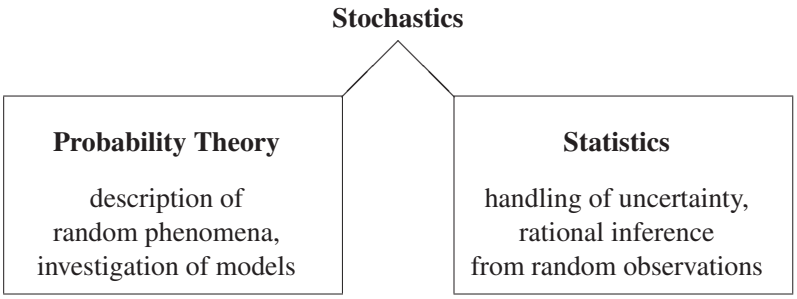
- ▷ by *abstracting* from 'dirty' details that might distract from the essentials of the problem at hand, that is, by 'smoothing away' all features that seem irrelevant; and, on the other hand,
- ▷ by *mathematical idealisation*, i.e., by widening the scope using mental or formal limiting procedures that allow for a clear-cut description of the relevant phenomena.



The model thus obtained can then be investigated mathematically, and the resulting predictions have to be checked against reality. If necessary, the model must be revised. In general, finding an appropriate model is a delicate process. It requires much flair and falls beyond mathematics. There are, however, some basic principles and mathematical structures; these will be discussed in this text.

*

Stochastics is composed of two equal parts: probability theory and statistics. The objective of probability theory is the description and investigation of specific random phenomena. Statistics seeks for methods of drawing rational conclusions from uncertain random observations. This, of course, requires and builds on the models of probability theory. The other way round, probability theory needs the validation obtained by comparing model and reality, which is made possible by statistics. Part I of this text offers an introduction to the basic concepts and results of probability theory. Part II then presents an introduction into theory and methods of mathematical statistics.



Part I

Probability Theory

Chapter 1

Principles of Modelling Chance

This chapter develops the fundamentals of modelling chance. The primary questions read: How can one describe a specific random phenomenon mathematically? What are the general properties of the stochastic model so obtained? How can one extract a particular aspect from a given model? In answering these questions, we will be led to the fundamental notions of 'probability space' and 'random variable'. We will also be faced with a few technical questions, which might be somewhat unpleasant to deal with at the beginning, but will enable us to concentrate on the principal ideas later on.

1.1 Probability Spaces

To build a mathematical model for a specific scenario of chance, one proceeds in three stages as follows.

1.1.1 Determining the Sample Space

If one aims at describing the effects of chance, the first question is: What can happen in the given situation? And which part of this is relevant? The possibilities that seem natural to distinguish are then collected into a set Ω. This is best understood by examples.

(1.1) Example. *Rolling a dice once.* If we roll a dice on a table, it can stop in infinitely many positions. We are not interested in its exact final position, and even less in the precise hand movement when throwing the dice, but only in the number that is showing. The interesting outcomes are thus captured by the set $\Omega = \{1, \dots, 6\}$. With the restriction to this Ω, we fade out the irrelevant part of reality.

(1.2) Example. *Rolling a dice several times.* If the dice is thrown n times and we are interested in the sequence of numbers shown, then the relevant outcomes are those in the product space $\Omega = \{1, \dots, 6\}^n$; for $\omega = (\omega_1, \dots, \omega_n) \in \Omega$ and $1 \le i \le n$, ω_i represents the number showing at the ith throw.

On the other hand, if we are not interested in the exact sequence of numbers, but only in how often each number appears, it is more natural to choose

$$\widehat{\Omega} = \Big\{(k_1, \dots, k_6) \in \mathbb{Z}_+^6 : \sum_{a=1}^{6} k_a = n\Big\}$$

as the set of all possible outcomes. Here $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ is the set of all non-negative integers, and k_a stands for the number of throws the dice shows a.

(1.3) Example. *Tossing a coin infinitely often.* When we toss a coin n times, the appropriate set of outcomes is $\Omega = \{0, 1\}^n$, in analogy with the previous example (provided we are interested in the sequence of outcomes). If we decide to flip the coin once more, do we have to consider a new Ω ? This would not be very practical; thus our model should not be limited to a fixed number of tosses. Moreover, we are especially interested in patterns that only appear for large n, that is in the limit as $n \to \infty$. Therefore it is often convenient to choose an idealised model, which admits an infinite number of tosses. (As an analogy, think of the mathematically natural transition from finite to infinite decimal fractions.) The set of all possible outcomes is then

$$\Omega = \{0, 1\}^{\mathbb{N}} = \left\{\omega = (\omega_i)_{i \in \mathbb{N}} : \omega_i \in \{0, 1\}\right\},$$

the set of all infinite sequences of zeros and ones.

As the examples demonstrate, the first step in the process of setting up a model for a random phenomenon is to decide which possibilities we would like to distinguish and observe, and which idealising assumptions might be useful. Considering this, one determines a set Ω of relevant outcomes. This Ω is called the *set of outcomes* or the *sample space*.

1.1.2 Determining a σ-Algebra of Events

In general we are not interested in the detailed outcome of a random experiment, but only in the occurrence of an *event* that consists of a certain selection of outcomes. Such events correspond to subsets of Ω.

(1.4) Example. *Events as sets of outcomes.* The event 'In n coin flips, heads shows at least k times' corresponds to the subset

$$A = \left\{\omega = (\omega_1, \dots, \omega_n) \in \Omega : \sum_{i=1}^{n} \omega_i \geq k\right\}$$

of the sample space $\Omega = \{0, 1\}^n$.

Our aim is to set up a system $\mathscr{F}$ of events, such that we can consistently assign to each event $A \in \mathscr{F}$ a probability $P(A)$ for A to occur.

Why so cautious? Why not assign a probability to *every* subset of Ω, in other words, why not simply take $\mathscr{F}$ to be the power set $\mathscr{P}(\Omega)$ (i.e., the set of *all* subsets of Ω)? Indeed, this is perfectly possible as long as Ω is countable. However, it is impossible in general, as the following 'no-go theorem' shows.

(1.5) Theorem. The power set is too large, Vitali 1905. *Let* $\Omega = \{0,1\}^{\mathbb{N}}$ *be the sample space for infinite coin tossing. Then there is* no *mapping* $P : \mathscr{P}(\Omega) \to [0,1]$ *with the properties*

(N) Normalisation. $P(\Omega) = 1$.

(A) σ-Additivity. *If* $A_1, A_2, \ldots \subset \Omega$ *are pairwise disjoint, then*

$$P\Big(\bigcup_{i\geq 1} A_i\Big) = \sum_{i\geq 1} P(A_i)\,.$$

(The probabilities of countably many incompatible events can be added up.)

(I) Flip invariance. *For all* $A \subset \Omega$ *and* $n \geq 1$ *one has* $P(T_n A) = P(A)$*; here*

$$T_n : \omega = (\omega_1, \omega_2, \ldots) \to (\omega_1, \ldots, \omega_{n-1}, 1-\omega_n, \omega_{n+1}, \ldots)$$

is the mapping from Ω *onto itself that inverts the result of the* n*th toss, and* $T_n A = \{T_n(\omega) : \omega \in A\}$ *is the image of* A *under* T_n*. (This expresses the fairness of the coin and the independence of the tosses.)*

At first reading, only the result is important, and its proof may be skipped.

Proof. We define an equivalence relation $\sim$ on Ω as follows. Say $\omega \sim \omega'$ if and only if $\omega_n = \omega'_n$ for all sufficiently large n. By the axiom of choice, there is a set $A \subset \Omega$ that contains exactly one element from each equivalence class.

Let $\mathscr{S} = \{S \subset \mathbb{N} : |S| < \infty\}$ be the set of all finite subsets of $\mathbb{N}$. As $\mathscr{S}$ is the union of the countably many finite sets $\{S \subset \mathbb{N} : \max S = m\}$ for $m \in \mathbb{N}$, it is countable. For $S = \{n_1, \ldots, n_k\} \in \mathscr{S}$ let $T_S := \prod_{n\in S} T_n = T_{n_1} \circ \cdots \circ T_{n_k}$ be the flip at all times contained in S. Then we have:

- $\Omega = \bigcup_{S\in\mathscr{S}} T_S A$, since for every $\omega \in \Omega$ there exists an $\omega' \in A$ such that $\omega \sim \omega'$, and thus an $S \in \mathscr{S}$ with $\omega = T_S\omega' \in T_S A$.
- The sets $(T_S A)_{S\in\mathscr{S}}$ are pairwise disjoint. For, suppose that $T_S A \cap T_{S'} A \neq \varnothing$ for some $S, S' \in \mathscr{S}$. Then there exist $\omega, \omega' \in A$ such that $T_S\omega = T_{S'}\omega'$ and so $\omega \sim T_S\omega = T_{S'}\omega' \sim \omega'$. By the choice of A, this means that $\omega = \omega'$ and thus $S = S'$.

Applying successively the properties (N), (A) and (I) of P we thus find

$$1 = P(\Omega) = \sum_{S\in\mathscr{S}} P(T_S A) = \sum_{S\in\mathscr{S}} P(A)\,.$$

This is impossible, since the infinite sum of the same number is either 0 or ∞. $\diamond$

How to proceed after this negative result? We have to insist on the properties (N), (A) and (I), since (N) and (A) are indispensable and elementary (just finite additivity is not sufficient, as we will see shortly), and (I) is characteristic of the coin tossing model. But the above proof has shown that the problems only arise for rather unusual, 'abnormal' sets $A \subset \Omega$. A natural way out is therefore to restrict the definition of

probabilities to an appropriate *subsystem* $\mathscr{F} \subset \mathscr{P}(\Omega)$, which excludes the 'abnormal' sets. Fortunately, it turns out that this suffices both in theory and in practice. In particular, we will see in Example (3.29) that a function P satisfying (N), (A) and (I) can indeed be defined on a suitable $\mathscr{F}$ that is large enough for all practical purposes.

What are sensible properties we should ask of the system $\mathscr{F}$? The minimal requirements are apparently those in the following

Definition. Suppose $\Omega \neq \varnothing$. A system $\mathscr{F} \subset \mathscr{P}(\Omega)$ satisfying

(a) $\Omega \in \mathscr{F}$ ($\mathscr{F}$ contains the 'certain event')

(b) $A \in \mathscr{F} \;\Rightarrow\; A^c := \Omega \setminus A \in \mathscr{F}$ ($\mathscr{F}$ allows the logical negation)

(c) $A_1, A_2, \ldots \in \mathscr{F} \;\Rightarrow\; \bigcup_{i\geq 1} A_i \in \mathscr{F}$ ($\mathscr{F}$ allows the countable logical 'or')

is called a *σ-algebra* (or *σ-field*) on Ω. The pair $(\Omega, \mathscr{F})$ is then called an *event space* or a *measurable space*.

These three properties can be combined to obtain some further properties of a σ-algebra $\mathscr{F}$. By (a) and (b), the 'impossible event' $\varnothing$ belongs to $\mathscr{F}$. Together with (c) this gives, for $A, B \in \mathscr{F}$, that $A \cup B = A \cup B \cup \varnothing \cup \cdots \in \mathscr{F}$, $A \cap B = (A^c \cup B^c)^c \in \mathscr{F}$, and $A \setminus B = A \cap B^c \in \mathscr{F}$. Similarly, the countable intersection of sets in $\mathscr{F}$ also belongs to $\mathscr{F}$.

The σ in the name σ-algebra has become convention as a reminder of the fact that, in (c), we consider countably infinite (instead of only finite) unions (σ like 'sums'). Finite unions are not sufficient, because we also need to consider so-called tail events, such as 'coin falls heads for infinitely many tosses' or 'the relative frequency of heads tends to $1/2$, if the number of tosses tends to ∞'. Such events can not be captured by finite unions, but require countably infinite unions (and intersections).

At this point, let us pause for a moment to distinguish between the three set-theoretic levels we are moving on; see Figure 1.1. The base level consists of the set Ω containing all outcomes ω. Above this there is the event level $\mathscr{P}(\Omega)$; its *elements* are *subsets* of the base level Ω. This structure repeats itself once more: σ-algebras are *subsets* of $\mathscr{P}(\Omega)$, so they are *elements* of the top level $\mathscr{P}(\mathscr{P}(\Omega))$.

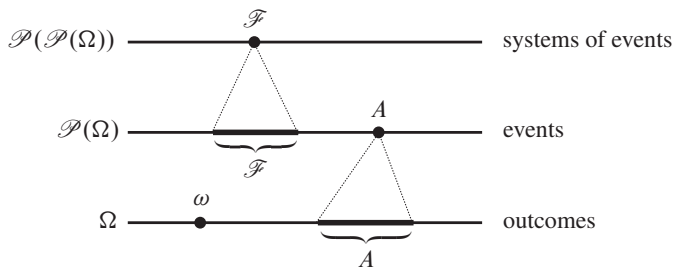


Figure 1.1. The three conceptual levels of stochastics.

How can one determine a σ-algebra in Ω ? One starts from a system $\mathscr{G}$ of 'good', that is, especially simple or natural sets, whose probability one can easily guess or determine. Then this system is enlarged just as much as necessary to obtain a σ-algebra. More precisely, the following construction principle can be used.

(1.6) Remark and Definition. *Generating σ-algebras.* If $\Omega \neq \varnothing$ and $\mathscr{G} \subset \mathscr{P}(\Omega)$ is arbitrary, then there is a unique smallest σ-algebra $\mathscr{F} = \sigma(\mathscr{G})$ on Ω such that $\mathscr{F} \supset \mathscr{G}$. This $\mathscr{F}$ is called the *σ-algebra generated by $\mathscr{G}$*, and $\mathscr{G}$ is called a *generator of $\mathscr{F}$*.

Proof. Let Σ be the system of all σ-algebras $\mathscr{A}$ in Ω satisfying $\mathscr{A} \supset \mathscr{G}$. (So Σ is a subset of the top level in Figure 1.1.) Σ is non-empty, since $\mathscr{P}(\Omega) \in \Sigma$. Hence, we can set $\mathscr{F} := \bigcap_{\mathscr{A} \in \Sigma} \mathscr{A}$. As each $\mathscr{A} \in \Sigma$ is a σ-algebra, so is $\mathscr{F}$, as is easily checked by spelling out the defining properties (a) to (c). $\mathscr{F}$ thus belongs to Σ, and is obviously its unique smallest element. This proves the claim. $\diamond$

Here are three standard examples of this construction.

(1.7) Example. *The power set.* Suppose Ω is countable and $\mathscr{G} = \big\{\{\omega\} : \omega \in \Omega\big\}$ the system containing the singleton sets of Ω. Then, $\sigma(\mathscr{G}) = \mathscr{P}(\Omega)$. Indeed, since every $A \in \mathscr{P}(\Omega)$ is countable, it follows from axiom (c) that $A = \bigcup_{\omega \in A} \{\omega\} \in \sigma(\mathscr{G})$.

(1.8) Example and Definition. *The Borel σ-algebra.* Let $\Omega = \mathbb{R}^n$ and

$$\mathscr{G} = \Big\{ \prod_{i=1}^{n} [a_i, b_i] : a_i < b_i,\ a_i, b_i \in \mathbb{Q} \Big\}$$

be the system consisting of all compact rectangular boxes in $\mathbb{R}^n$ with rational vertices and edges parallel to the axes. In honour of Émile Borel (1871–1956), the system $\mathscr{B}^n := \sigma(\mathscr{G})$ is called the *Borel σ-algebra* on $\mathbb{R}^n$, and every $A \in \mathscr{B}^n$ a *Borel set*; in the case $n = 1$, we simply write $\mathscr{B}$ instead of $\mathscr{B}^1$. The Borel σ-algebra is much larger than one might expect at first sight. Namely, we have:

(a) Every open set $A \subset \mathbb{R}^n$ is Borel. To see this, it is sufficient to note that every $\omega \in A$ has a neighbourhood $Q \in \mathscr{G}$ with $Q \subset A$, so that $A = \bigcup_{Q \in \mathscr{G},\, Q \subset A} Q$, a union of countably many sets. Our claim thus follows from property (c) of a σ-algebra.

(b) Every closed $A \subset \mathbb{R}^n$ is Borel, since A^c is open and hence by (a) Borel.

(c) It is impossible to describe $\mathscr{B}^n$ constructively. It does not only consist of countable unions of boxes and their complements; rather, the procedure of taking complements and countable unions has to be repeated as many times as there are countable ordinal numbers, hence uncountably often; see [31, p. 139] or [6, pp. 24, 29]. But this does not cause problems. It suffices to know that $\mathscr{B}^n$ is large enough to contain all sets in $\mathbb{R}^n$ that may occur in practice, but still smaller

than $\mathscr{P}(\mathbb{R})$. In fact, the existence of non-Borel sets follows from Theorem (1.5) and the proof of Theorem (3.12).

We will also need the following facts:

(d) Besides the system $\mathscr{G}$ of compact intervals, the Borel σ-algebra $\mathscr{B} = \mathscr{B}^1$ on $\mathbb{R}$ also admits the generator

$$\mathscr{G}' = \big\{]-\infty, c] : c \in \mathbb{R}\big\},$$

the system of all left-infinite closed half-lines. For, assertion (b) shows that $\mathscr{G}' \subset \mathscr{B}$ and thus, by the minimality of $\sigma(\mathscr{G}')$, $\sigma(\mathscr{G}') \subset \mathscr{B}$. Conversely, $\sigma(\mathscr{G}')$ contains all half-open intervals $]a, b] =]-\infty, b] \setminus]-\infty, a]$, and so all compact intervals $[a, b] = \bigcap_{n\geq 1}]a - \frac{1}{n}, b]$, hence also the σ-algebra $\mathscr{B}$ generated by them.

Likewise, $\mathscr{B}$ can also be generated by the left-infinite open half-lines, and also by the open or closed right-infinite half-lines.

(e) For $\varnothing \neq \Omega \subset \mathbb{R}^n$, the system $\mathscr{B}^n_\Omega = \{A \cap \Omega : A \in \mathscr{B}^n\}$ is a σ-algebra on Ω; it is called the *Borel σ-algebra on Ω*.

(1.9) Example and Definition. *Product σ-algebra.* Let Ω be the Cartesian product of arbitrary sets E_i, i.e., $\Omega = \prod_{i\in I} E_i$ for an index set $I \neq \varnothing$. Let $\mathscr{E}_i$ be a σ-algebra on E_i, $X_i : \Omega \to E_i$ the projection mapping onto the ith coordinate, and $\mathscr{G} = \{X_i^{-1} A_i : i \in I,\ A_i \in \mathscr{E}_i\}$ the system of all sets in Ω which are specified by an event in a single coordinate. Then, $\bigotimes_{i\in I} \mathscr{E}_i := \sigma(\mathscr{G})$ is called the *product σ-algebra* of the $\mathscr{E}_i$ on Ω. If $E_i = E$ and $\mathscr{E}_i = \mathscr{E}$ for all i, we write $\mathscr{E}^{\otimes I}$ instead of $\bigotimes_{i\in I} \mathscr{E}_i$. For example, the Borel σ-algebra $\mathscr{B}^n$ on $\mathbb{R}^n$ is exactly the n-fold product σ-algebra of the Borel σ-algebra $\mathscr{B} = \mathscr{B}^1$ on $\mathbb{R}$, meaning that $\mathscr{B}^n = \mathscr{B}^{\otimes n}$; cf. Problem 1.3.

The second step in the process of building a model can now be summarised as follows. Theorem (1.5) forces us to introduce a σ-algebra $\mathscr{F}$ of events in Ω. Fortunately, in most cases the choice of $\mathscr{F}$ is canonical. In this book, only the following three *standard cases* will appear.

- ▷ *Discrete case:* If Ω is at most countable, one can set $\mathscr{F} = \mathscr{P}(\Omega)$.
- ▷ *Real case:* For $\Omega \subset \mathbb{R}^n$, the natural choice is $\mathscr{F} = \mathscr{B}^n_\Omega$.
- ▷ *Product case:* If $\Omega = \prod_{i\in I} E_i$ and every E_i is equipped with a σ-algebra $\mathscr{E}_i$, one takes $\mathscr{F} = \bigotimes_{i\in I} \mathscr{E}_i$.

Once a σ-algebra $\mathscr{F}$ is fixed, every $A \in \mathscr{F}$ is called an *event* or a *measurable set*.

1.1.3 Assigning Probabilities to Events

The decisive point in the process of building a stochastic model is the next step: For each $A \in \mathscr{F}$ we need to define a value $P(A) \in [0, 1]$ that indicates the probability of A. Sensibly, this should be done so that the following holds.

(N) Normalisation: $P(\Omega) = 1$.

(A) σ-Additivity: For pairwise disjoint events $A_1, A_2, \ldots \in \mathscr{F}$ one has

$$P\Big(\bigcup_{i\geq 1} A_i\Big) = \sum_{i\geq 1} P(A_i)\,.$$

(Pairwise disjoint means that $A_i \cap A_j = \varnothing$ for $i \neq j$.)

Definition. Let $(\Omega, \mathscr{F})$ be an event space. A function $P : \mathscr{F} \to [0, 1]$ satisfying the properties (N) and (A) is called a *probability measure* or a *probability distribution*, in short a *distribution* (or, a little old-fashioned, a *probability law*) on $(\Omega, \mathscr{F})$. Then, the triple $(\Omega, \mathscr{F}, P)$ is called a *probability space*.

Properties (N) and (A), together with the non-negativity of the probability measure, are also known as *Kolmogorov's axioms*, since it was Andrej N. Kolmogorov (1903–1987) who in 1933 emphasised the significance of measures for the mathematical foundation of probability theory, and thus gave a decisive input for the development of modern probability theory.

Let us summarise: To describe a particular scenario of chance mathematically, one has to choose an appropriate probability space. Typically, the most delicate point is the choice of the probability measure P, since this contains all the relevant information about the kind of randomness. In Chapter 2, as well as in many examples later on, we will show how this can be done. At this point let us only mention the elementary, but degenerate, example of a probability measure that describes a situation without randomness.

(1.10) Example and Definition. *Deterministic case.* If $(\Omega, \mathscr{F})$ is an arbitrary event space and $\xi \in \Omega$, then

$$\delta_\xi(A) = \begin{cases} 1 & \text{if } \xi \in A\,, \\ 0 & \text{otherwise} \end{cases}$$

defines a probability measure δ_ξ on $(\Omega, \mathscr{F})$. It describes an experiment with the certain outcome ξ and is called the *Dirac distribution* or the *unit mass* at the point ξ.

We close this section with some remarks on the

Interpretation of probability measures. The concept of a probability space does not give an answer to the philosophical question what probability really is. The following are common answers:

(a) *The naive interpretation.* 'Nature' is uncertain about what it is doing, and $P(A)$ represents the degree of certainty of its decision to let A happen.

(b) *The frequency interpretation.* $P(A)$ is the relative frequency with which A occurs under some specified conditions.

(c) *The subjective interpretation.* $P(A)$ is the degree of certainty with which I would be willing to bet on the occurrence of A according to my personal evaluation of the situation.

(The interpretations (a) and (c) are dual concepts, the uncertainty moves from nature to the observer.)

In general, we cannot say which interpretation is to be preferred, since this depends on the nature of the problem at hand. If a random experiment can be repeated independently, the interpretations (a) and (b) may seem most natural. The probabilities of rain specified by weather forecasts are obviously based on (b), and so are the probabilities used in the insurance industry. The question that was asked before March 23 in 2001, namely 'What is the probability that humans will be injured by the crash of the space station *Mir*?', used the subjective interpretation (c), since it dealt with a singular event. A comprehensive and very stimulating historical and philosophical discussion of the notion of probability can be found in Gigerenzer et al. [23].

Fortunately, the validity of the mathematical statements about a probability model does not depend on its interpretation. The value of mathematics is not limited by the narrowness of human thought. This, however, should not to be misunderstood to mean that mathematics can take place in an 'ivory tower'. Stochastics thrives on the interaction with the real world.

1.2 Properties and Construction of Probability Measures

What are the implications of the assumption (A) of σ-additivity? We start with some elementary consequences.

(1.11) Theorem. Probability rules. *Every probability measure P on an event space $(\Omega, \mathscr{F})$ has the following properties, for arbitrary events $A, B, A_1, A_2, \ldots \in \mathscr{F}$.*

(a) $P(\varnothing) = 0$.

(b) Finite additivity. $P(A \cup B) + P(A \cap B) = P(A) + P(B)$, *and so in particular* $P(A) + P(A^c) = 1$.

(c) Monotonicity. *If $A \subset B$ then $P(A) \leq P(B)$.*

(d) σ-Subadditivity. $P\left(\bigcup_{i\geq 1} A_i\right) \leq \sum_{i\geq 1} P(A_i)$.

(e) σ-Continuity. *If either $A_n \uparrow A$ or $A_n \downarrow A$ (i.e., the A_n are either increasing with union A, or decreasing with intersection A), then $P(A_n) \underset{n\to\infty}{\longrightarrow} P(A)$.*

Proof. (a) Since the empty set is disjoint from itself, the sequence $\varnothing, \varnothing, \ldots$ consists of pairwise disjoint events. In this extreme case, σ-additivity (A) thus gives

$$P(\varnothing) = P(\varnothing \cup \varnothing \cup \cdots) = \sum_{i=1}^{\infty} P(\varnothing)\,.$$

But this is only possible when $P(\varnothing) = 0$.

(b) Suppose first that A and B are disjoint. Since property (A) requires an infinite sequence, we append the empty set infinitely often to the sets A and B. Hence we obtain from (A) and statement (a)

$$P(A \cup B) = P(A \cup B \cup \varnothing \cup \varnothing \cup \cdots) = P(A) + P(B) + 0 + 0 + \cdots\,.$$

So the probability is additive when an event is split into finitely many disjoint parts. In the general case, it thus follows that

$$\begin{aligned} P(A \cup B) + P(A \cap B) &= P(A \setminus B) + P(B \setminus A) + 2P(A \cap B) \\ &= P(A) + P(B)\,. \end{aligned}$$

The second assertion follows from the normalisation axiom (N) by taking $B = A^c$.

(c) For $B \supset A$ we conclude from (b) that $P(B) = P(A) + P(B \setminus A) \geq P(A)$ because probabilities are non-negative.

(d) Any union $\bigcup_{i\geq 1} A_i$ can actually be represented as a union of disjoint sets, by removing from A_i the part of A_i that is already contained in a 'previous' A_j. This procedure is known as the 'first entrance trick'. So we can write, using assumption (A) and statement (c),

$$P\Big(\bigcup_{i\geq 1} A_i\Big) = P\Big(\bigcup_{i\geq 1}\big(A_i \setminus \bigcup_{j<i} A_j\big)\Big) = \sum_{i\geq 1} P\big(A_i \setminus \bigcup_{j<i} A_j\big) \leq \sum_{i\geq 1} P(A_i)\,.$$

(e) If $A_n \uparrow A$, the σ-additivity (A) and the finite additivity (b) give, with $A_0 := \varnothing$,

$$\begin{aligned} P(A) &= P\Big(\bigcup_{i\geq 1}(A_i \setminus A_{i-1})\Big) = \sum_{i\geq 1} P(A_i \setminus A_{i-1}) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i \setminus A_{i-1}) = \lim_{n\to\infty} P(A_n)\,. \end{aligned}$$

The case $A_n \downarrow A$ follows by taking complements and using (b). $\diamond$

A less obvious, but equally important consequence of σ-additivity is the fact that each probability measure is already determined by its restriction to a suitable generator of the σ-algebra.

(1.12) Theorem. Uniqueness theorem. *Let $(\Omega, \mathscr{F}, P)$ be a probability space, and suppose that $\mathscr{F} = \sigma(\mathscr{G})$ for a generator $\mathscr{G} \subset \mathscr{P}(\Omega)$. If $\mathscr{G}$ is intersection-stable, in the sense that $A, B \in \mathscr{G}$ implies $A \cap B \in \mathscr{G}$, then P is uniquely determined by its restriction $P|_{\mathscr{G}}$ to $\mathscr{G}$.*

Although we will apply the uniqueness theorem repeatedly, its proof should be skipped at first reading, since this kind of reasoning will not be used later on.

Proof. Let Q be an arbitrary probability measure on $(\Omega, \mathscr{F})$ such that $P|_{\mathscr{G}} = Q|_{\mathscr{G}}$. The system $\mathscr{D} = \{A \in \mathscr{F} : P(A) = Q(A)\}$ then exhibits the following properties.

(a) $\Omega \in \mathscr{D}$.

(b) If $A, B \in \mathscr{D}$ and $A \subset B$, then $B \setminus A \in \mathscr{D}$.

(c) If $A_1, A_2, \ldots \in \mathscr{D}$ are pairwise disjoint, then $\bigcup_{i \geq 1} A_i \in \mathscr{D}$.

Indeed, (a) follows from (N), (c) from (A), and (b) is immediate because $P(B \setminus A) = P(B) - P(A)$ for $A \subset B$. A system $\mathscr{D}$ satisfying (a) to (c) is called a *Dynkin system* (after the Russian mathematician Eugene B. Dynkin, *1924). By assumption we have $\mathscr{D} \supset \mathscr{G}$. Thus $\mathscr{D}$ also contains the Dynkin system $d(\mathscr{G})$ generated by $\mathscr{G}$. As in Remark (1.6), $d(\mathscr{G})$ is defined to be the smallest Dynkin system containing $\mathscr{G}$; the existence of such a smallest Dynkin system is proved in exactly the same way as indicated there. The following lemma will show that $d(\mathscr{G}) = \sigma(\mathscr{G}) = \mathscr{F}$. As a consequence, we have that $\mathscr{D} = \mathscr{F}$ and hence $P = Q$. $\diamond$

To complete the proof, we need the following lemma.

(1.13) Lemma. Generated Dynkin system. *For an intersection-stable system $\mathscr{G}$, the identity $d(\mathscr{G}) = \sigma(\mathscr{G})$ holds.*

Proof. Since $\sigma(\mathscr{G})$ is a σ-algebra, it is also a Dynkin system, and since $d(\mathscr{G})$ is minimal, we have $\sigma(\mathscr{G}) \supset d(\mathscr{G})$. Conversely, we will show that $d(\mathscr{G})$ is a σ-algebra. For, this implies that $\sigma(\mathscr{G}) \subset d(\mathscr{G})$ by the minimality of $\sigma(\mathscr{G})$.

Step 1. $d(\mathscr{G})$ is intersection-stable. Indeed, $\mathscr{D}_1 := \{A \subset \Omega : A \cap B \in d(\mathscr{G})$ for all $B \in \mathscr{G}\}$ is obviously a Dynkin system, and since $\mathscr{G}$ is intersection-stable, we have $\mathscr{D}_1 \supset \mathscr{G}$. By the minimality of $d(\mathscr{G})$, it then follows that $\mathscr{D}_1 \supset d(\mathscr{G})$, i.e., we have $A \cap B \in d(\mathscr{G})$ for all $A \in d(\mathscr{G})$ and $B \in \mathscr{G}$.

Similarly, $\mathscr{D}_2 := \{A \subset \Omega : A \cap B \in d(\mathscr{G})$ for all $B \in d(\mathscr{G})\}$ is also a Dynkin system, and, by the above, $\mathscr{D}_2 \supset \mathscr{G}$. Hence, we also have $\mathscr{D}_2 \supset d(\mathscr{G})$, i.e., $A \cap B \in d(\mathscr{G})$ for all $A, B \in d(\mathscr{G})$.

Step 2. $d(\mathscr{G})$ is a σ-algebra. For, let $A_1, A_2, \ldots \in d(\mathscr{G})$. By Step 1, the sets

$$B_i := A_i \setminus \bigcup_{j<i} A_j = A_i \cap \bigcap_{j<i} \Omega \setminus A_j$$

then also belong to $d(\mathscr{G})$ and are pairwise disjoint. Hence $\bigcup_{i \geq 1} A_i = \bigcup_{i \geq 1} B_i \in d(\mathscr{G})$. $\diamond$

Our next question is: How can one construct a probability measure on a σ-algebra? In view of the uniqueness theorem, we can reformulate this question as follows: Under which conditions can we extend a function P defined on a suitable system $\mathscr{G}$ to a probability measure defined on the generated σ-algebra $\sigma(\mathscr{G})$?

A satisfactory answer is given by a theorem from measure theory, namely Carathéodory's extension theorem, see for example [4, 6, 12, 15, 16]; here we will not discuss this further. However, to guarantee the existence of non-trivial probability measures on non-discrete sample spaces, we will have to take the existence of the Lebesgue integral for granted. We will make use of the following

(1.14) Fact. *Lebesgue integral.* For every function $f : \mathbb{R}^n \to [0, \infty]$ that satisfies the measurability criterion

$$\{x \in \mathbb{R}^n : f(x) \le c\} \in \mathscr{B}^n \quad \text{for all } c > 0 \tag{1.15}$$

(which will be discussed further in Example (1.26)), one can define the *Lebesgue integral* $\int f(x)\,dx \in [0, \infty]$ in such a way that the following holds.

(a) For every Riemann-integrable function f, $\int f(x)\,dx$ coincides with the Riemann integral of f.

(b) For every sequence $f_1, f_2, \ldots$ of non-negative measurable functions as above, we have

$$\int \sum_{n\ge1} f_n(x)\,dx = \sum_{n\ge1} \int f_n(x)\,dx\,.$$

A proof of these statements can be found in numerous textbooks on analysis, such as [56, 57]. As property (a) shows, in concrete computations it often suffices to know the Riemann integral. However, the Riemann integral does not satisfy the σ-additivity property (b), which is essential for our purposes; it is equivalent to the monotone convergence theorem, see also Theorem (4.11c) later on.

In particular, the Lebesgue integral yields a sensible notion of volume for Borel sets in $\mathbb{R}^n$. Namely, let

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A\,, \\ 0 & \text{otherwise} \end{cases} \tag{1.16}$$

denote the *indicator function* of a set A. The integral over $A \in \mathscr{B}^n$ is then defined as

$$\int_A f(x)\,dx := \int 1_A(x) f(x)\,dx\,.$$

In the special case $f \equiv 1$, property (1.14b) then gives us the following result.

(1.17) Remark and Definition. *Lebesgue measure.* The mapping $\lambda^n : \mathscr{B}^n \to [0, \infty]$ that assigns to each $A \in \mathscr{B}^n$ its n-dimensional volume

$$\lambda^n(A) := \int 1_A(x)\, dx$$

satisfies the σ-additivity property (A), and we have $\lambda^n(\varnothing) = 0$. Consequently, λ^n is a 'measure' on $(\mathbb{R}^n, \mathscr{B}^n)$. It is called the ($n$-dimensional) *Lebesgue measure on* $\mathbb{R}^n$. For $\Omega \in \mathscr{B}^n$, the restriction λ^n_Ω of λ^n to $\mathscr{B}^n_\Omega$ is called the *Lebesgue measure on* Ω.

We will see repeatedly that the existence of many interesting probability measures can be deduced from the existence of the Lebesgue measure. Here, we will use the Lebesgue measure for constructing probability measures on $\mathbb{R}^n$ (or subsets thereof) by means of density functions, a procedure which is obvious for discrete spaces. See the illustration in Figure 1.2.

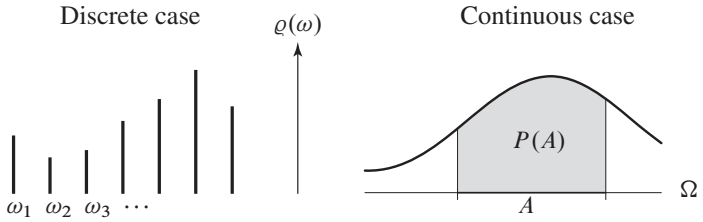


Figure 1.2. On the left: bar chart of a discrete density. On the right: Lebesgue density; its integral over an event A yields the probability $P(A)$.

(1.18) Theorem. Construction of probability measures via densities.

(a) Discrete case: *For countable* Ω*, the relations*

$$P(A) = \sum_{\omega \in A} \rho(\omega) \text{ for } A \in \mathscr{P}(\Omega), \quad \rho(\omega) = P(\{\omega\}) \text{ for } \omega \in \Omega$$

establish a one-to-one correspondence between the set of all probability measures P *on* $(\Omega, \mathscr{P}(\Omega))$ *and the set of all sequences* $\rho = (\rho(\omega))_{\omega \in \Omega}$ *in* $[0, 1]$ *such that* $\sum_{\omega \in \Omega} \rho(\omega) = 1$.

(b) Continuous case: *If* $\Omega \subset \mathbb{R}^n$ *is Borel, then every function* $\rho : \Omega \to [0, \infty[$ *satisfying the properties*

(i) $\{x \in \Omega : \rho(x) \le c\} \in \mathscr{B}^n_\Omega$ *for all* $c > 0$ (cf. (1.15)),

(ii) $\int_\Omega \rho(x)\, dx = 1$

determines a unique probability measure P on $(\Omega, \mathscr{B}_\Omega^n)$ via

$$P(A) = \int_A \rho(x)\, dx \quad \text{for } A \in \mathscr{B}_\Omega^n$$

(but not every probability measure on $(\Omega, \mathscr{B}_\Omega^n)$ is of this form).

Proof. The discrete case is obvious. In the continuous case, the claim follows immediately from Fact (1.14b), since $1_{\bigcup_{i\geq 1} A_i} = \sum_{i\geq 1} 1_{A_i}$ when the A_i are pairwise disjoint. $\diamond$

Definition. A sequence or function ρ as in Theorem (1.18) above is called a *density* (of P) or, more explicitly (to emphasise normalisation), a *probability density (function)*, often abbreviated as *pdf*. If a distinction between the discrete and continuous case is required, a sequence $\rho = (\rho(\omega))_{\omega\in\Omega}$ as in case (a) is called a *discrete density*, and a function $\rho : \Omega \to [0, \infty[$ as in case (b) a *Lebesgue density*.

A basic class of probability measures defined by densities is given by the uniform distributions, which are defined as follows and will be discussed in more detail in Section 2.1.

(1.19) Example and Definition. *The uniform distributions.* If Ω is finite, the probability measure having the constant discrete density $\rho(\omega) = 1/|\Omega|$ (so that all $\omega \in \Omega$ occur with the same probability) is called the (discrete) *uniform distribution* on Ω and is denoted by $\mathcal{U}_\Omega$.

Likewise, if $\Omega \subset \mathbb{R}^n$ is a Borel set with volume $0 < \lambda^n(\Omega) < \infty$, the probability measure on $(\Omega, \mathscr{B}_\Omega)$ with the constant Lebesgue density $\rho(x) = 1/\lambda^n(\Omega)$ is called the (continuous) *uniform distribution* on Ω; it is also denoted by $\mathcal{U}_\Omega$.

Let us conclude this section with some comments. In contrast to the discrete case, there is no one-to-one correspondence between probability measures and their densities in the continuous case. On the one hand, most probability measures on $\mathbb{R}^n$ fail to have a Lebesgue density, but many of the most common probability measures do. On the other hand, any two probability densities that differ only on a set of vanishing Lebesgue measure determine the same probability measure. For example, we have $\mathcal{U}_{]0,1[} = \mathcal{U}_{[0,1]}$.

Next we note that probability measures on Borel subsets of $\mathbb{R}^n$ can also be viewed as probability measures on all of $\mathbb{R}^n$. Specifically, let $\Omega \subset \mathbb{R}^n$ be a Borel set and P a probability measure on $(\Omega, \mathscr{B}_\Omega^n)$ with Lebesgue density ρ. The measure P can then be identified with the probability measure $\bar{P}$ on $(\mathbb{R}^n, \mathscr{B}^n)$ with density $\bar{\rho}$, where $\bar{\rho}(x) = \rho(x)$ for $x \in \Omega$ and $\bar{\rho}(x) = 0$ otherwise. Indeed, we have $\bar{P}(\mathbb{R}^n \setminus \Omega) = 0$, and $\bar{P}$ and P coincide on $\mathscr{B}_\Omega^n$. We will often carry out this identification without mentioning it explicitly. There is also an analogue in the discrete case: If $\Omega \subset \mathbb{R}^n$ is countable and P a probability measure on $(\Omega, \mathscr{P}(\Omega))$ with discrete density ρ,

we can identify P with the probability measure $\sum_{\omega\in\Omega} \rho(\omega)\,\delta_\omega$, which is defined on $(\mathbb{R}^n, \mathscr{B}^n)$, or in fact even on $(\mathbb{R}^n, \mathscr{P}(\mathbb{R}^n))$; here δ_ω is the Dirac measure introduced in (1.10).

Finally, it is possible to combine discrete and continuous probability measures. For example,

$$P(A) = \tfrac{1}{3}\,\delta_{-1/2}(A) + \tfrac{2}{3}\,\mathcal{U}_{]0,1/2[}(A)\,, \quad A \in \mathscr{B}, \tag{1.20}$$

defines a probability measure on $(\mathbb{R}, \mathscr{B})$, which for two thirds is 'blurred uniformly' over the interval $]0, 1/2[$ and assigns the extra probability $1/3$ to the point $-1/2$.

1.3 Random Variables

Let us return for a moment to the first step of setting up a model, as described in Section 1.1.1. The choice of the sample space Ω is not unique, but depends on how many details of the random phenomenon should be included into the model, and is therefore a matter of the appropriate *observation depth*.

(1.21) Example. *Tossing a coin n times*. On the one hand, one can record the result of every single toss; then $\Omega = \{0, 1\}^n$ is the appropriate sample space. Alternatively, one may restrict attention to the number of tosses when heads is showing. Then, the natural sample space is $\Omega' = \{0, 1, \ldots, n\}$. The second case corresponds to a lower observation depth. The process of reducing the observation depth can be described by the mapping $X : \Omega \to \Omega'$ that assigns to each $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ the sum $\sum_{i=1}^n \omega_i \in \Omega'$, which indicates the 'number of successes'.

The example shows: The transition from a given event space $(\Omega, \mathscr{F})$ to a coarser model $(\Omega', \mathscr{F}')$ providing less information is captured by a mapping from the detailed to the coarser sample space, i.e., from Ω to Ω'. In the general case, such a mapping should satisfy the requirement

$$A' \in \mathscr{F}' \;\Rightarrow\; X^{-1}A' \in \mathscr{F}, \tag{1.22}$$

that is, all events on the coarse level can be traced back to events on the detailed level via the preimage mapping X^{-1}. The situation is visualised in Figure 1.3.

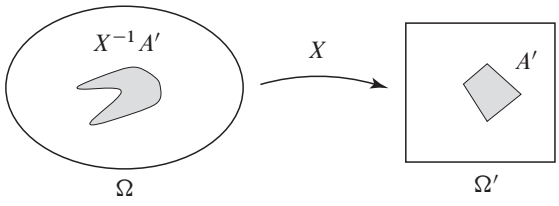


Figure 1.3. For a random variable, the preimage of an event in Ω' is an event in Ω.

Definition. Let $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$ be two event spaces. Then every mapping $X : \Omega \to \Omega'$ satisfying property (1.22) is called a *random variable from* $(\Omega, \mathscr{F})$ *to* $(\Omega', \mathscr{F}')$, or a *random element of* Ω'. Alternatively (in the terminology of measure theory), X is said to be *measurable* relative to $\mathscr{F}$ and $\mathscr{F}'$.

Due to (1.22), preimages will occur frequently in the following. In stochastics, it is common to use the suggestive notation

$$\{X \in A'\} := \{\omega \in \Omega : X(\omega) \in A'\} = X^{-1}A' . \tag{1.23}$$

Let us note first that condition (1.22) holds automatically in the discrete case.

(1.24) Example. *Random variables on discrete spaces.* If $\mathscr{F} = \mathscr{P}(\Omega)$, then *every* mapping $X : \Omega \to \Omega'$ is a random variable.

In the general case, the following criterion is crucial.

(1.25) Remark. *Measurability criterion.* In the set-up of the previous definition, suppose that $\mathscr{F}'$ is generated by a system $\mathscr{G}'$, in that $\mathscr{F}' = \sigma(\mathscr{G}')$. Then $X : \Omega \to \Omega'$ is already a random variable when $X^{-1}A' \in \mathscr{F}$ for all $A' \in \mathscr{G}'$ only.

Proof. The system $\mathscr{A}' := \{A' \subset \Omega' : X^{-1}A' \in \mathscr{F}\}$ is a σ-algebra, which by assumption contains $\mathscr{G}'$. Since $\mathscr{F}'$ is by definition the smallest such σ-algebra, we also have $\mathscr{A}' \supset \mathscr{F}'$, which means that X satisfies condition (1.22). $\diamond$

(1.26) Example. *Real random variables.* Let $(\Omega', \mathscr{F}') = (\mathbb{R}, \mathscr{B})$. For a real function $X : \Omega \to \mathbb{R}$ to be a random variable, it is sufficient that all sets of the form

$$\{X \le c\} := X^{-1}\,]-\infty, c\,]$$

belong to $\mathscr{F}$. (Alternatively, one can replace '$\le$' by '$<$', '$\ge$' or '$>$'.) This follows immediately from Remark (1.25) and Fact (1.8d).

It is often convenient to consider so-called extended real functions taking values in $\bar{\mathbb{R}} = [-\infty, \infty]$. $\bar{\mathbb{R}}$ is equipped with the σ-algebra generated by the intervals $[-\infty, c]$, $c \in \mathbb{R}$. (Think about how this relates to the Borel σ-algebra on $\mathbb{R}$.) Consequently, an extended real function $X : \Omega \to \bar{\mathbb{R}}$ is a random variable if and only if $\{X \le c\} \in \mathscr{F}$ for all $c \in \mathbb{R}$.

(1.27) Example. *Continuous functions.* Let $\Omega \subset \mathbb{R}^n$ and $\mathscr{F} = \mathscr{B}^n_\Omega$. Then every continuous function $X : \Omega \to \mathbb{R}$ is a random variable. Indeed, for every $c \in \mathbb{R}$, $\{X \le c\}$ is closed in Ω, so by Example (1.8be) it belongs to $\mathscr{B}^n_\Omega$. Thus the claim follows from Example (1.26).

The next theorem describes an important principle for constructing new probability measures, which will be used repeatedly.

(1.28) Theorem. Distribution of a random variable. *If X is a random variable from a probability space $(\Omega, \mathscr{F}, P)$ to an event space $(\Omega', \mathscr{F}')$, then the prescription*

$$P'(A') := P(X^{-1}A') = P(\{X \in A'\}) \quad \textit{for } A' \in \mathscr{F}'$$

defines a probability measure P' on $(\Omega', \mathscr{F}')$.

To simplify the notation, we will omit the braces in expressions like $P(\{X \in A'\})$, and simply write $P(X \in A')$ in the future.

Proof. By (1.22) the definition of P' makes sense. Furthermore, P' satisfies the conditions (N) and (A). Indeed, on the one hand, $P'(\Omega') = P(X \in \Omega') = P(\Omega) = 1$. On the other hand, if $A'_1, A'_2, \ldots \in \mathscr{F}'$ are pairwise disjoint, so are their preimages $X^{-1}A'_1, X^{-1}A'_2, \ldots$, whence

$$\begin{aligned} P'\Big(\bigcup_{i\geq 1} A'_i\Big) &= P\Big(X^{-1}\bigcup_{i\geq 1} A'_i\Big) = P\Big(\bigcup_{i\geq 1} X^{-1}A'_i\Big) \\ &= \sum_{i\geq 1} P(X^{-1}A'_i) = \sum_{i\geq 1} P'(A'_i)\,. \end{aligned}$$

Hence P' is a probability measure. $\diamond$

Definition. (a) The probability measure P' in Theorem (1.28) is called the *distribution of X under P*, or the *image of P under X*, and is denoted by $P \circ X^{-1}$. (In the literature, one also finds the notations P_X or $\mathscr{L}(X; P)$. The letter $\mathscr{L}$ stands for the more traditional term *law*, or *loi* in French.)

(b) Two random variables are said to be *identically distributed* if they have the same distribution.

At this point, we need to emphasise that the term 'distribution' is used in an inflationary way in stochastics. Apart from the meaning that we have just introduced, it is also generally used as a synonym for probability measure. (In fact, every probability measure is the distribution of a random variable, namely the identity function of the underlying Ω.) This has to be distinguished from two further notions, namely 'distribution function' and 'distribution density', which refer to the real event space $(\mathbb{R}, \mathscr{B})$ and will be introduced now.

Each probability measure P on $(\mathbb{R}, \mathscr{B})$ is already uniquely determined by the function $F_P(c) := P(]-\infty, c])$ for $c \in \mathbb{R}$. Likewise, the distribution of a real-valued random variable X on a probability space $(\Omega, \mathscr{F}, P)$ is uniquely determined by the function $F_X : c \to P(X \leq c)$ on $\mathbb{R}$. This is because any two probability measures on $(\mathbb{R}, \mathscr{B})$ coincide if and only if they agree on all intervals of the form $]-\infty, c]$, by statement (1.8d) and the uniqueness theorem (1.12). This motivates the following concepts.

Definition. For a probability measure P on the real line $(\mathbb{R}, \mathscr{B})$, the function $F_P : c \to P(]-\infty, c])$ from $\mathbb{R}$ to $[0,1]$ is called the (cumulative) *distribution function of* P. Likewise, for a real random variable X on a probability space $(\Omega, \mathscr{F}, P)$, the distribution function $F_X(c) := F_{P \circ X^{-1}}(c) = P(X \le c)$ of its distribution is called the (cumulative) *distribution function of* X.

Every distribution function $F = F_X$ is increasing and right-continuous and has the asymptotic behaviour

$$\lim_{c \to -\infty} F(c) = 0 \quad \text{and} \quad \lim_{c \to +\infty} F(c) = 1\,. \tag{1.29}$$

This follows immediately from Theorem (1.11); see Problem 1.16. Figure 1.4 shows an example. Remarkably, *every* function with these properties is the distribution function of a random variable on the unit interval (equipped with the uniform distribution from Example (1.19)). The term 'quantile' occurring in the name of these random variables will play an important role in statistics, i.e., in Part II; see the definition on p. 231.

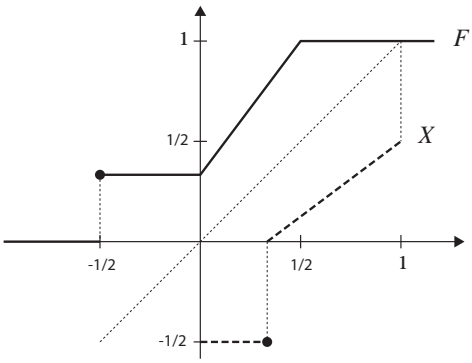


Figure 1.4. Distribution function F (bold) and quantile transformation X (dashed) of the probability measure $\frac{1}{3}\,\delta_{-1/2} + \frac{2}{3}\,\mathcal{U}_{]0,1/2[}$ from (1.20). The dotted lines illustrate that X is obtained from F by reflection at the diagonal. The values at the discontinuities are marked by bullets.

(1.30) Proposition. Quantile transformation. *For every increasing right-continuous function F on $\mathbb{R}$ with limit behaviour* (1.29)*, there exists a real random variable X on the probability space* $(]0,1[, \mathscr{B}_{]0,1[}, \mathcal{U}_{]0,1[})$ *such that $F_X = F$. This X is given explicitly by $X(u) = \inf\{c \in \mathbb{R} : F(c) \ge u\}$, $u \in]0,1[$, and is called the 'quantile transformation'.*

Proof. By (1.29) we have $-\infty < X(u) < \infty$ for all $0 < u < 1$. In fact, X is a left-continuous inverse of F; compare Figure 1.4. Indeed, $X(u) \le c$ holds if and only if $u \le F(c)$; this is because, by the right-continuity of F, the infimum in the definition

of X is in fact a minimum. In particular, $\{X \le c\} =]0, F(c)] \cap]0,1[\in \mathscr{B}_{]0,1[}$. Together with Example (1.26) this shows that X is a random variable. Furthermore, the set $\{X \le c\}$ has Lebesgue measure $F(c)$. Hence F is the distribution function of X. $\diamond$

Since every probability measure P on $(\mathbb{R}, \mathscr{B})$ is uniquely determined by its distribution function, we can rephrase the proposition as follows: Every P on $(\mathbb{R}, \mathscr{B})$ is the distribution of a random variable on the probability space $(]0,1[, \mathscr{B}_{]0,1[}, \mathcal{U}_{]0,1[})$. This fact will repeatedly be useful.

The connection between distribution functions and probability densities is made by the notion of a distribution density.

(1.31) Remark and Definition. *Existence of a distribution density.* Let X be a real random variable on a probability space $(\Omega, \mathscr{F}, P)$. Its distribution $P \circ X^{-1}$ admits a Lebesgue density ρ if and only if

$$F_X(c) = \int_{-\infty}^{c} \rho(x)\, dx \quad \text{for all } c \in \mathbb{R}.$$

Such a ρ is called the *distribution density* of X. In particular, $P \circ X^{-1}$ admits a continuous density ρ if and only if F_X is continuously differentiable, and then $\rho = F_X'$. This follows directly from (1.8d) and the uniqueness theorem (1.12).

Problems

1.1 Let $(\Omega, \mathscr{F})$ be an event space, $A_1, A_2, \ldots \in \mathscr{F}$ and

$$A = \{\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n\}.$$

Show that (a) $A = \bigcap_{N\ge 1} \bigcup_{n \ge N} A_n$, (b) $1_A = \limsup_{n\to\infty} 1_{A_n}$.

1.2 Let Ω be uncountable and $\mathscr{G} = \{\{\omega\} : \omega \in \Omega\}$ the system of the singleton subsets of Ω. Show that $\sigma(\mathscr{G}) = \{A \subset \Omega : A \text{ or } A^c \text{ is countable}\}$.

1.3 [S] Show that the Borel σ-algebra $\mathscr{B}^n$ on $\mathbb{R}^n$ coincides with $\mathscr{B}^{\otimes n}$, the n-fold product of the Borel σ-algebra $\mathscr{B}$ on $\mathbb{R}$.

1.4 Let $\Omega \subset \mathbb{R}^n$ be at most countable. Show that $\mathscr{B}^n_\Omega = \mathscr{P}(\Omega)$.

1.5 Let E_i, $i \in \mathbb{N}$, be countable sets and $\Omega = \prod_{i\ge 1} E_i$ their Cartesian product. Denote by $X_i : \Omega \to E_i$ the projection onto the ith coordinate. Show that the system

$$\mathscr{G} = \big\{\{X_1 = x_1, \ldots, X_k = x_k\} : k \ge 1,\, x_i \in E_i\big\} \cup \big\{\varnothing\big\}$$

is an intersection-stable generator of the product σ-algebra $\bigotimes_{i\ge 1} \mathscr{P}(E_i)$.

1.6 [S] Let $(\Omega_i, \mathscr{F}_i)$, $i = 1, 2$, be two event spaces and $\omega_1 \in \Omega_1$. Show the following. For every $A \in \mathscr{F}_1 \otimes \mathscr{F}_2$, the '$\omega_1$-section' $A_{\omega_1} := \{\omega_2 \in \Omega_2 : (\omega_1, \omega_2) \in A\}$ of A belongs to $\mathscr{F}_2$, and if f is a random variable on $(\Omega_1 \times \Omega_2, \mathscr{F}_1 \otimes \mathscr{F}_2)$ then the function $f(\omega_1, \cdot)$ is a random variable on $(\Omega_2, \mathscr{F}_2)$.

1.7 S *Inclusion–exclusion principle.* Let $(\Omega, \mathscr{F}, P)$ be a probability space and $A_i \in \mathscr{F}$, $i \in I = \{1, \dots, n\}$. For $J \subset I$ let

$$B_J = \bigcap_{j \in J} A_j \cap \bigcap_{j \in I \setminus J} A_j^c ;$$

by convention, an intersection over an empty index set is equal to Ω. Show the following:

(a) For all $K \subset I$,

$$P\Big(\bigcap_{k \in K} A_k\Big) = \sum_{K \subset J \subset I} P(B_J).$$

(b) For all $J \subset I$,

$$P(B_J) = \sum_{J \subset K \subset I} (-1)^{|K \setminus J|} P\Big(\bigcap_{k \in K} A_k\Big).$$

What does this imply for $J = \varnothing$?

1.8 *Bonferroni inequality.* Let $A_1, \dots, A_n$ be any events in a probability space $(\Omega, \mathscr{F}, P)$. Show that

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) \geq \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) .$$

1.9 A certain Chevalier de Méré, who has become famous in the history of probability theory for his gambling problems and their solutions by Pascal, once mentioned to Pascal how surprised he was that when throwing three dice he observed the total sum of 11 more often than the sum of 12, although 11 could be obtained by the combinations 6-4-1, 6-3-2, 5-5-1, 5-4-2, 5-3-3, 4-4-3, and the sum of 12 by as many combinations (which ones?). Can we consider his observation as caused by 'chance' or is there an error in his argument? To solve the problem, introduce a suitable probability space.

1.10 In a pack of six chocolate drinks every carton is supposed to have a straw, but it is missing with probability $1/3$, with probability $1/3$ it is broken and only with probability $1/3$ it is in perfect condition. Let A be the event 'at least one straw is missing and at least one is in perfect condition'. Exhibit a suitable probability space, formulate the event A set-theoretically, and determine its probability.

1.11 S Alice and Bob agree to play a fair game over 7 rounds. Each of them pays €5 as an initial stake, and the winner gets the total of €10. At the score of 2 : 3 they have to stop the game. Alice suggests to split the winnings in this ratio. Should Bob accept the offer? Set up an appropriate model and calculate the probability of winning for Bob.

1.12 *The birthday paradox.* Let p_n be the probability that in a class of n children at least two have their birthday on the same day. For simplicity, we assume here that no birthday is on February 29th, and all other birthdays are equally likely. Show (using the inequality $1 - x \leq e^{-x}$) that

$$p_n \geq 1 - \exp\big(-n(n-1)/730\big),$$

and determine the smallest n such that $p_n \geq 1/2$.

1.13 S *The rencontre problem.* Alice and Bob agree to play the following game: From two completely new, identical sets of playing cards, one is well shuffled. Both piles are put next to each other face down, and then revealed card by card simultaneously. Bob bets (for a stake of €10) that in this procedure at least two identical cards will be revealed at the same time. Alice, however, is convinced that this is 'completely unlikely' and so bets the opposite way. Who do you think is more likely to win? Set up an appropriate model and calculate the probability of winning for Alice. *Hint:* Use Problem 1.7b; the sum that appears can be approximated by the corresponding infinite series.

1.14 Let $X, Y, X_1, X_2, \ldots$ be real random variables on an event space $(\Omega, \mathscr{F})$. Prove the following statements.

(a) $(X, Y) : \Omega \to \mathbb{R}^2$ is a random variable.

(b) $X + Y$ and XY are random variables.

(c) $\sup_{n\in\mathbb{N}} X_n$ and $\limsup_{n\to\infty} X_n$ are random variables (taking values in $\bar{\mathbb{R}}$).

(d) $\{X = Y\} \in \mathscr{F}$, $\{\lim_{n\to\infty} X_n \text{ exists}\} \in \mathscr{F}$, $\{X = \lim_{n\to\infty} X_n\} \in \mathscr{F}$.

1.15 S Let $(\Omega, \mathscr{F}) = (\mathbb{R}, \mathscr{B})$ and $X : \Omega \to \mathbb{R}$ be an arbitrary real function. Verify the following:

(a) If X is piecewise monotone (i.e., $\mathbb{R}$ may be decomposed into at most countably many intervals, on each of which X is either increasing or decreasing), then X is a random variable.

(b) If X is differentiable with (not necessarily continuous) derivative X', then X' is a random variable.

1.16 *Properties of distribution functions.* Let P be a probability measure on $(\mathbb{R}, \mathscr{B})$ and $F(c) = P(]-\infty, c])$, for $c \in \mathbb{R}$, its distribution function. Show that F is increasing and right-continuous, and (1.29) holds.

1.17 Consider the two cases

(a) $\Omega = [0, \infty[$, $\rho(\omega) = e^{-\omega}$, $X(\omega) = (\omega/\alpha)^{1/\beta}$ for $\omega \in \Omega$ and $\alpha, \beta > 0$,

(b) $\Omega =]-\pi/2, \pi/2[$, $\rho(\omega) = 1/\pi$, $X(\omega) = \sin^2 \omega$ for $\omega \in \Omega$.

In each case, show that ρ is a probability density and X a random variable on $(\Omega, \mathscr{B}_\Omega)$, and calculate the distribution density of X with respect to the probability measure P with density ρ. (The distribution of X in case (a) is called the *Weibull distribution* with parameters α, β, in case (b) the *arcsine distribution*.)

1.18 S *Transformation to uniformity.* Prove the following converse to Proposition (1.30): If X is a real random variable with a *continuous* distribution function $F_X = F$, then the random variable $F(X)$ is uniformly distributed on $[0, 1]$. Show further that the continuity of F is necessary for this to hold.

Chapter 2

Stochastic Standard Models

Having described the general structure of stochastic models, we will now discuss how to find suitable models for concrete random phenomena. In general, this can be quite delicate, and requires the right balance between being close to reality yet mathematically tractable. At this stage, however, we confine ourselves to several classical examples, for which the appropriate model is quite obvious. This gives us the opportunity to introduce some fundamental probability distributions along with typical applications. These distributions can be used as building blocks of more complex models, as we will see later on.

2.1 The Uniform Distributions

There are two different types of uniform distributions: the discrete ones on finite sets, and the continuous uniform distributions on Borel subsets of $\mathbb{R}^n$.

2.1.1 Discrete Uniform Distributions

Let us start with the simplest case of a random experiment with only finitely many possible outcomes, i.e., an experiment with a finite sample space Ω. For example, we can think of tossing a coin or rolling a dice several times. In these and many other examples, symmetry suggests the assumption that all single outcomes $\omega \in \Omega$ are equally likely. By Theorem (1.18a) this means that the probability measure P should have the *constant* density $\rho(\omega) = 1/|\Omega|$ (for $\omega \in \Omega$). This leads to the approach $P = \mathcal{U}_\Omega$, where

$$\mathcal{U}_\Omega(A) = \frac{|A|}{|\Omega|} = \frac{\text{number of 'favourable' outcomes}}{\text{number of possible outcomes}} \quad \text{for all } A \subset \Omega\,. \tag{2.1}$$

Definition. For a finite set Ω, the probability measure $\mathcal{U}_\Omega$ on $(\Omega, \mathcal{P}(\Omega))$ defined by (2.1) is called the *(discrete) uniform distribution* on Ω. Sometimes $(\Omega, \mathcal{P}(\Omega), \mathcal{U}_\Omega)$ is also called a *Laplace space* (in honour of Pierre-Simon Laplace, 1749–1827).

Classical examples in which the uniform distribution shows up are tossing a coin or rolling a dice (once or several times), the lottery, playing cards, and many more. Several of these examples will be discussed soon, in particular in Sections 2.2 and 2.3. A less obvious example is the following.

(2.2) Example. *The Bose–Einstein distribution (1924).* Consider a system of n indistinguishable particles that are distributed over N different 'cells'; the cells are of the same type, but distinguishable. For example, one can imagine the seeds in the pits of the Syrian game Kalah, or – and this was Bose's and Einstein's motivation – physical particles, whose phase space is partitioned into finitely many cells. A (macro) state of the system is determined by specifying how many particles populate each cell. Hence,

$$\Omega = \Big\{(k_1, \dots, k_N) \in \mathbb{Z}_+^N : \sum_{j=1}^{N} k_j = n\Big\}.$$

This sample space has cardinality $|\Omega| = \binom{n+N-1}{n}$, since each $(k_1, \dots, k_N) \in \Omega$ is uniquely characterised by a sequence of the form

$$\underbrace{\bullet\cdots\bullet}_{k_1}|\underbrace{\bullet\cdots\bullet}_{k_2}|\cdots|\underbrace{\bullet\cdots\bullet}_{k_N},$$

where the blocks of $k_1, \dots, k_N$ balls are separated from each other by a total of $N-1$ vertical bars. To determine a state, we only have to place the n balls (resp. the $N-1$ vertical bars) into the $n+N-1$ available positions. Hence, the uniform distribution $\mathcal{U}_\Omega$ on Ω is given by $\mathcal{U}_\Omega(\{\omega\}) = 1/\binom{n+N-1}{n}$, $\omega \in \Omega$. The assumption of a uniform distribution agrees with the experimental results in the case of so-called bosons (i.e., particles with integer spin, such as photons and mesons).

Physicists often speak of Bose–Einstein 'statistics'. In this traditional terminology, statistics means the same as 'probability distribution' and has nothing in common with statistics in today's mathematical sense.

2.1.2 Continuous Uniform Distributions

Let us begin with a motivating example.

(2.3) Example. *Random choice of a direction.* Imagine we spin a roulette wheel. After the wheel has stopped, into which direction is the zero facing? The angle it forms relative to a fixed direction is a number in the interval $\Omega = [0, 2\pi[$, which is equipped with the Borel σ-algebra $\mathscr{F} := \mathscr{B}_\Omega$. Which probability measure P describes the situation? For every $n \geq 1$, Ω can be partitioned in the n disjoint intervals $[\frac{k}{n}2\pi, \frac{k+1}{n}2\pi[$ with $0 \leq k < n$. By symmetry, each of these should receive the same probability, which is then necessarily equal to $1/n$. That is, we should have

$$P\Big(\Big[\frac{k}{n}2\pi, \frac{k+1}{n}2\pi\Big[\Big) = \frac{1}{n} = \int_{\frac{k}{n}2\pi}^{\frac{k+1}{n}2\pi} \frac{1}{2\pi}\,dx$$

for $0 \leq k < n$ and, by additivity,

$$P\Big(\Big[\frac{k}{n}2\pi, \frac{l}{n}2\pi\Big[\Big) = \int_{\frac{k}{n}2\pi}^{\frac{l}{n}2\pi} \frac{1}{2\pi}\,dx$$

for $0 \le k < l \le n$. By Theorems (1.12) and (1.18), there exists only one probability measure P with this property, namely the one with the constant Lebesgue density $1/2\pi$ on $[0, 2\pi[$. This P reflects the idea that all possible directions are equally likely.

Definition. Let $\Omega \subset \mathbb{R}^n$ be a Borel set with n-dimensional volume $0 < \lambda^n(\Omega) < \infty$; cf. (1.17). The probability measure $\mathcal{U}_\Omega$ on $(\Omega, \mathscr{B}^n_\Omega)$ with constant Lebesgue density $\rho(x) = 1/\lambda^n(\Omega)$, which is given by

$$\mathcal{U}_\Omega(A) = \int_A \frac{1}{\lambda^n(\Omega)}\, dx = \frac{\lambda^n(A)}{\lambda^n(\Omega)}\,, \quad A \in \mathscr{B}^n_\Omega\,,$$

is called the *(continuous) uniform distribution* on Ω.

Note that, depending on the context, $\mathcal{U}_\Omega$ can either stand for a discrete or a continuous distribution. But both cases are completely analogous. The favourable resp. possible outcomes are simply counted in the discrete case (2.1), whereas in the continuous case they are measured with Lebesgue measure. The following application of continuous uniform distributions is an example of historical interest, and also a little taster from so-called stochastic geometry.

(2.4) Example. *Bertrand's paradox.* Given a circle with radius $r > 0$, a chord is chosen 'at random'. What is the probability that it is longer than the sides of an inscribed equilateral triangle? (This problem appeared in 1889 in a textbook of the French mathematician J. L. F. Bertrand, 1822–1900.)

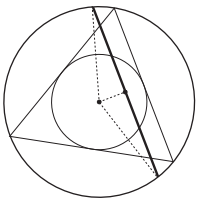


Figure 2.1. The geometry of Bertrand's paradox. The incircle of the inscribed equilateral triangle has half the radius.

The answer depends on what one considers a 'random choice' or, in other words, it depends on the method the chord is chosen.

First approach. The chord is uniquely determined by its midpoint (as long as it is not the centre of the circle; this case can be neglected). Hence, a possible sample space is $\Omega_1 = \{x \in \mathbb{R}^2 : |x| < r\}$, and it seems reasonable to interpret 'choosing at random' by taking the uniform distribution $\mathcal{U}_{\Omega_1}$ as the underlying probability measure. The event 'the chord is longer than the sides of the inscribed equilateral triangle'

is then described by the set $A_1 = \{x \in \Omega_1 : |x| < r/2\}$, cf. Figure 2.1. Consequently,

$$\mathcal{U}_{\Omega_1}(A_1) = \frac{\pi(r/2)^2}{\pi r^2} = \frac{1}{4}\,.$$

Second approach. The chord is uniquely determined by the angle under which it is seen from the centre of the circle, and the direction of its perpendicular bisector; the latter is irrelevant by the rotational symmetry of the problem. The angle falls into $\Omega_2 =]0, \pi]$. The relevant event is $A_2 =]2\pi/3, \pi]$. If we again use the uniform distribution, it follows that

$$\mathcal{U}_{\Omega_2}(A_2) = \frac{\pi/3}{\pi} = \frac{1}{3}\,.$$

Third approach. The chord is also uniquely determined by its distance and direction from the centre; the latter can again be ignored. Hence, we can also take the sample space $\Omega_3 = [0, r[$. Then $A_3 = [0, r/2[$ is the event we are interested in, and we obtain $\mathcal{U}_{\Omega_3}(A_3) = 1/2$.

In Bertrand's times, this apparent paradox cast doubt on the legitimacy of non-discrete probability spaces. Today it is clear that the three versions describe different methods of choosing the chord 'at random', and it is all but surprising that the probability we are looking for depends on the choice of the method.

> Some readers may consider this way of resolving the paradox as a cheap way out, because they think that there must be a unique 'natural' interpretation of 'choosing at random'. In fact the latter is true, but only if we reformulate the problem by requiring that the object chosen 'at random' is not a chord, but a straight line that intersects the circle. Such a random straight line is best described by the third approach, since one can show that this is the only case in which the probability that a random straight line hits a set A is invariant under rotations and translations of A.

This example demonstrates that the choice of a suitable model can be non-trivial, even in a simple case like this, which only involves uniform distributions. This is a main problem in all applications.

2.2 Urn Models with Replacement

The so-called urn models form a simple class of stochastic models with finite sample space. Their significance comes from the observation that many random experiments can be thought of as picking balls of different colours at random from a container, which is traditionally called an 'urn'. In this section we consider the case that the balls are replaced after being drawn. The case without replacement follows in the next section.

2.2.1 Ordered Samples

We begin with two examples.

(2.5) Example. *Investigation of a fish pond.* Consider a fish pond inhabited by several species of fish, and let E be the set of species. Suppose the pond contains N_a fish of species $a \in E$, so the total population size is $\sum_{a\in E} N_a = N$. We repeat the following procedure n times. A fish is caught, examined for parasites, say, and then thrown back into the pond. What is the probability that a random sample exhibits a specific sequence of species?

(2.6) Example. *A survey.* A local radio station interviews passers-by on the high street concerning a question of local interest, for instance the construction of a new football stadium. Let E be the set of viewpoints expressed (possible location, rejection on principle, etc.). The station interviews n people. What is the probability that a given sequence of views is observed?

Such problems, where samples are picked randomly from a given pool, are often formulated in an abstract way as *urn models*. An urn contains a certain number of coloured but otherwise identical balls. Let E be the set of colours, where $2 \le |E| < \infty$. A sample of size n is taken from the urn with replacement, meaning that n times a ball is drawn at random and returned immediately. We are interested in the colour of each ball we take out. Thus, the sample space is $\Omega = E^n$, equipped with the σ-algebra $\mathscr{F} = \mathscr{P}(\Omega)$. Which probability measure P describes the situation adequately?

To find out, we proceed as follows. We imagine that the balls are labelled with the numbers $1, \dots, N$; the labels of the balls with colour $a \in E$ are collected in a class $C_a \subset \{1, \dots, N\}$. Thus $|C_a| = N_a$. If we could observe the labels, we would describe our experiment by the sample space $\overline{\Omega} = \{1, \dots, N\}^n$ (with the σ-algebra $\overline{\mathscr{F}} = \mathscr{P}(\overline{\Omega})$), and since all balls are identical (up to their colour), we would choose our probability measure to be the uniform distribution $\bar{P} = \mathcal{U}_{\overline{\Omega}}$. So, using the trick of increasing the observation depth artificially by labelling the balls, we arrive at a plausible stochastic model.

We now return to the true sample space $\Omega = E^n$. As we have seen in Section 1.3, this transition can be described by constructing a suitable random variable $X : \overline{\Omega} \to \Omega$. The colour of the ith draw is specified by the random variable

$$X_i : \overline{\Omega} \to E, \quad \bar{\omega} = (\bar{\omega}_1, \dots, \bar{\omega}_n) \to a \text{ if } \bar{\omega}_i \in C_a \,.$$

The sequence of colours in our sample is then given by the n-step random variable $X = (X_1, \dots, X_n) : \overline{\Omega} \to \Omega$.

What is the distribution of X? For every $\omega = (\omega_1, \dots, \omega_n) \in E^n$ we have

$$\{X = \omega\} = C_{\omega_1} \times \cdots \times C_{\omega_n}$$

and thus

$$\bar{P} \circ X^{-1}(\{\omega\}) = \bar{P}(X = \omega) = \frac{|C_{\omega_1}| \dots |C_{\omega_n}|}{|\overline{\Omega}|} = \prod_{i=1}^{n} \rho(\omega_i)\,,$$

where $\rho(a) = |C_a|/N = N_a/N$ is the proportion of balls of colour a.

Definition. For every density ρ on E, the (discrete) density

$$\rho^{\otimes n}(\omega) = \prod_{i=1}^{n} \rho(\omega_i)$$

on E^n is called the *n-fold product density of* ρ, and the corresponding probability measure P on E^n is the *n-fold product measure of* ρ. (We will not introduce an extra notation for P, but instead we will use the notation $\rho^{\otimes n}$ for P as well.)

In the special case when $E = \{0, 1\}$ and $\rho(1) = p \in [0, 1]$, one obtains the product density

$$\rho^{\otimes n}(\omega) = p^{\sum_{i=1}^{n} \omega_i} (1-p)^{\sum_{i=1}^{n} (1-\omega_i)}$$

on $\{0, 1\}^n$, and P is called the *Bernoulli measure* or the *Bernoulli distribution* for n trials with 'success probability' p. (The name refers to Jakob Bernoulli, 1654–1705.)

2.2.2 Unordered Samples

In the urn model, one is usually not interested in the (temporal) order in which the colours were selected, but only in how many balls of each colour were drawn. (In particular, this applies to the situations of Examples (2.5) and (2.6).) This corresponds to a lower observation depth and leads to the sample space

$$\widehat{\Omega} = \left\{ \vec{k} = (k_a)_{a \in E} \in \mathbb{Z}_+^E,\ \sum_{a \in E} k_a = n \right\},$$

which consists of the integral grid points in the simplex spanned by the $|E|$ vertices $(n\delta_{a,b})_{b \in E}$, $a \in E$; cf. Figure 2.3 on p. 37. The transition to $\widehat{\Omega}$ is described by the random variable

$$S : \Omega \to \widehat{\Omega}, \quad \omega = (\omega_1, \dots, \omega_n) \to \big(S_a(\omega)\big)_{a \in E}\,; \tag{2.7}$$

where $S_a(\omega) = \sum_{i=1}^{n} 1_{\{a\}}(\omega_i)$ is the number of occurrences of colour a in sample point $\omega \in E^n$. $S(\omega)$ is called the *histogram* of ω. It can be visualised by plotting a rectangle of width 1 and height $S_a(\omega)$ for every $a \in E$. Notice that the total area of all rectangles is n.

Now, for $P = \rho^{\otimes n}$ and $\vec{k} = (k_a)_{a \in E} \in \widehat{\Omega}$, we find

$$P(S = \vec{k}) = \sum_{\omega \in \Omega:\, S(\omega) = \vec{k}} \prod_{i=1}^{n} \rho(\omega_i) = \binom{n}{\vec{k}} \prod_{a \in E} \rho(a)^{k_a}\,.$$