Dirk P. Kroese and Joshua C.C. Chan

Statistical Modeling and Computation

An Inclusive Approach to Statistics

July 19, 2013

Springer

In memory of Reuven Rubinstein, my Friend and Mentor

Dirk Kroese

To Raquel

Joshua Chan

Preface

Statistics provides one of the few principled means to extract information from random data, and has perhaps more interdisciplinary connections than any other field of science. However, for a beginning student of statistics the abundance of mathematical concepts, statistical philosophies, and numerical techniques can seem overwhelming. The purpose of this book is to provide a comprehensive and accessible introduction to modern statistics, illuminating its many facets, both from a classical (frequentist) and Bayesian point of view. The book offers an integrated treatment of mathematical statistics and modern statistical computation.

The book is aimed at beginning students of statistics and practitioners who would like to fully understand the theory and key numerical techniques of statistics. It is based on a progression of undergraduate statistics courses at The University of Queensland and The Australian National University. Parts of the book have also been successfully tested at The University of New South Wales. Emphasis is laid on the mathematical and computational aspects of statistics. No prior knowledge of statistics is required, but we assume that the reader has a basic knowledge of mathematics, which forms an essential basis for the development of the statistical theory. Starting from scratch, the book gradually builds up to an advanced undergraduate level, providing a solid basis for possible postgraduate research. Throughout the text we illustrate the theory by providing working code in Matlab, rather than relying on black-box statistical packages. We make frequent use of the symbol ☞ in the margin to facilitate cross-referencing between related pages. The book is accompanied by the website `www.statmodcomp.org` from which the Matlab code and data files can be downloaded. In addition, we provide an R equivalent for each Matlab program.

The book is structured into three parts. In Part I we introduce the fundamentals of probability theory. We discuss models for random experiments, conditional probability and independence, random variables, and probability distributions. Moreover, we explain how to carry out random experiments on a computer.

In Part II we introduce the general framework for statistical modeling and inference, both from a classical and Bayesian perspective. We discuss a variety of common models for data, such as independent random samples, linear regression, and ANOVA models. Once a model for the data is determined one can carry out a mathematical analysis of the model on the basis of the available data. We discuss a wide range of concepts and techniques for statistical inference, including likelihood-based estimation and hypothesis testing, sufficiency, confidence intervals, and kernel density estimation. We encompass both classical and Bayesian approaches, and also highlight popular Monte Carlo sampling techniques.

In Part III we address the statistical analysis and computation of a variety of advanced models, such as generalized linear models, autoregressive and moving average models, Gaussian models, and state space models. Particular attention is paid to fast numerical techniques for classical and Bayesian inference on these models. Throughout the book our leading principle is that the mathematical formulation of a statistical model goes hand in hand with the specification of its simulation counterpart.

The book contains a large number of illustrative examples and problem sets (with solutions). To keep the book fully self-contained, we include the more technical proofs and mathematical theory in an appendix. A separate appendix features a concise introduction to MATLAB.

Brisbane and Canberra, *Dirk Kroese*
July 19, 2013 *Joshua Chan*

Acknowledgements

This book has benefited from the input of many people. We thank Zdravko Botev, Tim Brereton, Hyun Choi, Eric Eisenstat, Eunice Foo, Catherine Forbes, Patricia Galvan, Ivan Jeliazkov, Ross McVinish, Gary Koop, Rongrong Qu, Ad Ridder, Leonardo Rojas–Nandayapa, John Stachurski, Rodney Strachan, Mingzhu Sun, Thomas Taimre, Justin Tobias, Elisse Yulian, and Bo Zhang for their valuable comments and suggestions on previous drafts of the book.

Contents

Part I Fundamentals of Probability

Abbreviations and Acronyms

ANOVA	analysis of variance
AR	autoregressive
ARMA	autoregressive moving average
cdf	cumulative distribution function
EM	expectation–maximization
iid	independent and identically distributed
pdf	probability density function (discrete or continuous)
PGF	probability generating function
KDE	kernel density estimate/estimator
MA	moving average
MCMC	Markov chain Monte Carlo
MGF	moment generating function
ML(E)	maximum likelihood (estimate/estimator)
PRESS	predicted residual sum of squares

Mathematical Notation

Throughout this book we use notation in which different fonts and letter cases signify different types of mathematical objects. For example, vectors $\mathbf{a}, \mathbf{b}, \mathbf{x}, \ldots$ are written in lowercase boldface font, and matrices A, B, X in uppercase normal font. Sans serif fonts indicate probability distributions, such as N, Exp, and Bin. Probability and expectation symbols are written in black board bold font: $\mathbb{P}$ and $\mathbb{E}$. Matlab code and functions will always be written in `typewriter` font.

Traditionally, classical and Bayesian statistics use a *different* notation system for random variables and their probability density functions. In classical statistics and probability theory random variables usually are denoted by uppercase letters $X, Y, Z, \ldots$, and their outcomes by lower case letters $x, y, z, \ldots$. Bayesian statisticians typically use lower case letters for both. More importantly, in the Bayesian notation system it is common to use the *same* letter f (or p) for different probability densities, as in $f(x,y) = f(x)f(y)$. Classical statisticians and probabilists would prefer a different symbol for each function, as in $f(x,y) = f_X(x)f_Y(y)$. We will predominantly use the classical notation, especially in the first part of the book. However, when dealing with Bayesian models and inference, such as in Chapters 8 and 11, it will be convenient to switch to the Bayesian notation system. Here is a list of frequently used symbols:

$\approx$	is approximately
$\propto$	is proportional to
∞	infinity
$\otimes$	Kronecker product
$\overset{\text{def}}{=}$	is defined as
$\sim$	is distributed as
$\overset{\text{iid}}{\sim}$, $\sim_{\text{iid}}$	are independent and identically distributed as
$\overset{\text{approx.}}{\sim}$	is approximately distributed as
$\mapsto$	maps to
$A \cup B$	union of sets A and B
$A \cap B$	intersection of sets A and B
A^c	complement of set A
$A \subset B$	A is a subset of B
$\emptyset$	empty set
$\lVert\mathbf{x}\rVert$	Euclidean norm of vector $\mathbf{x}$
∇f	gradient of f
$\nabla^2 f$	Hessian of f
$A^\top$, $\mathbf{x}^\top$	transpose of matrix A or vector $\mathbf{x}$
$\mathrm{diag}(\mathbf{a})$	diagonal matrix with diagonal entries defined by $\mathbf{a}$
$\mathrm{tr}(A)$	trace of matrix A

$\det(A)$	determinant of matrix A
$\|A\|$	absolute value of the determinant of matrix A. Also, number of elements in set A, or absolute value of real number A
argmax	$\operatorname{argmax} f(x)$ is a value x^* for which $f(x^*) \geqslant f(x)$ for all x
d	differential symbol
$\mathbb{E}$	expectation
e	Euler's constant $\lim_{n\to\infty}(1+1/n)^n = 2.71828\ldots$
i	the square root of -1
$\mathrm{I}_A, \mathrm{I}\{A\}$	indicator function: equal to 1 if the condition/event A holds, and 0 otherwise.
ln	(natural) logarithm
$\mathbb{N}$	set of natural numbers $\{0, 1, \ldots\}$
φ	pdf of the standard normal distribution
$\varPhi$	cdf of the standard normal distribution
$\mathbb{P}$	probability measure
$\mathcal{O}$	big-O order symbol: $f(x) = \mathcal{O}(g(x))$ if $\|f(x)\| \leqslant \alpha g(x)$ for some constant α as $x \to a$
o	little-o order symbol: $f(x) = o(g(x))$ if $f(x)/g(x) \to 0$ as $x \to a$
$\mathbb{R}$	the real line = one-dimensional Euclidean space
$\mathbb{R}_+$	positive real line: $[0, \infty)$
$\mathbb{R}^n$	n-dimensional Euclidean space
$\widehat{\boldsymbol{\theta}}$	estimate/estimator
$\mathbf{x}, \mathbf{y}$	vectors
$\mathbf{X}, \mathbf{Y}$	random vectors
$\mathbb{Z}$	set of integers $\{\ldots, -1, 0, 1, \ldots\}$

Probability Distributions

Ber	Bernoulli distribution
Beta	beta distribution
Bin	binomial distribution
Cauchy	Cauchy distribution
χ^2	chi-squared distribution
Dirichlet	Dirichlet distribution
DU	discrete uniform distribution
Exp	exponential distribution
F	F distribution
Gamma	gamma distribution
Geom	geometric distribution
InvGamma	inverse-gamma distribution
Mnom	multinomial distribution
N	normal or Gaussian distribution
Poi	Poisson distribution
t	Student's t distribution
TN	truncated normal distribution
U	uniform distribution
Weib	Weibull distribution

Part I
Fundamentals of Probability

In Part I of the book we consider the *probability* side of statistics. In particular, we will consider how random experiments can be modelled mathematically, and how such modeling enables us to compute various properties of interest for those experiments.

Chapter 1
Probability Models

1.1 Random Experiments

The basic notion in probability is that of a **random experiment**: an experiment whose outcome cannot be determined in advance, but which is nevertheless subject to analysis. Examples of random experiments are:

1. tossing a die and observing its face value,
2. measuring the amount of monthly rainfall in a certain location,
3. counting the number of calls arriving at a telephone exchange during a fixed time period,
4. selecting at random fifty people and observing the number of left-handers,
5. choosing at random ten people and measuring their heights.

The goal of *probability* is to understand the behavior of random experiments by analyzing the corresponding *mathematical models*. Given a mathematical model for a random experiment one can calculate quantities of interest such as probabilities and expectations. Moreover, such mathematical models can typically be implemented on a computer, so that it becomes possible to *simulate* the experiment. Conversely, any computer implementation of a random experiment implicitly defines a mathematical model. Mathematical models for random experiments are also the basis of *statistics*, where the objective is to infer which of several competing models best fits the observed data. This often involves the estimation of model parameters from the data.

Example 1.1 (Coin Tossing). One of the most fundamental random experiments is the one where a coin is tossed a number of times. Indeed, much of probability theory can be based on this simple experiment. To better understand how this coin toss experiment behaves, we can carry it out on a computer, using programs such as Matlab. The following simple Matlab program simulates a sequence of 100 tosses with a fair coin (that is, Heads and Tails are equally likely), and plots the results in a bar chart.

```
x = (rand(1,100) < 0.5)   % generate the coin tosses
bar(x)                    % plot the results in a bar chart
```

The function `rand` draws uniform random numbers from the interval $[0, 1]$ — in this case a 1×100 vector of such numbers. By testing whether the uniform numbers are less than 0.5, we obtain a vector `x` of 1s and 0s, indicating Heads and Tails, say. Typical outcomes for three such experiments are given in Figure 1.1.

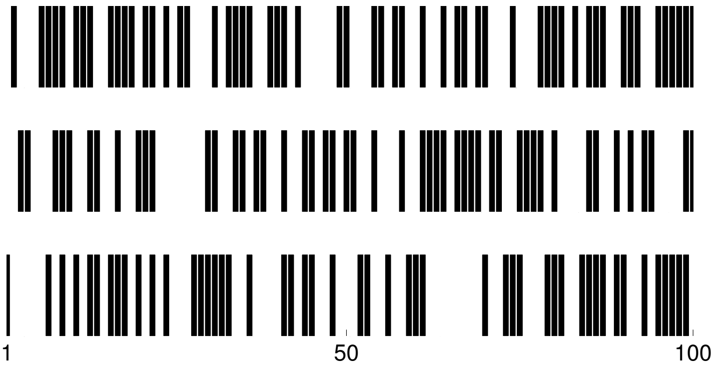

Fig. 1.1 Three experiments where a fair coin is tossed 100 times. The dark bars indicate when "Heads" (=1) appears.

We can also plot the average number of Heads against the number of tosses. In the same MATLAB program, this is accomplished by adding two lines of code:

```
y = cumsum(x)./[1:100] % calculate the cumulative sum and
             % divide this elementwise by the vector [1:100]
plot(y)      % plot the result in a line graph
```

The result of three such experiments is depicted in Figure 1.2. Notice that the average number of Heads seems to converge to 0.5, but there is a lot of random fluctuation.

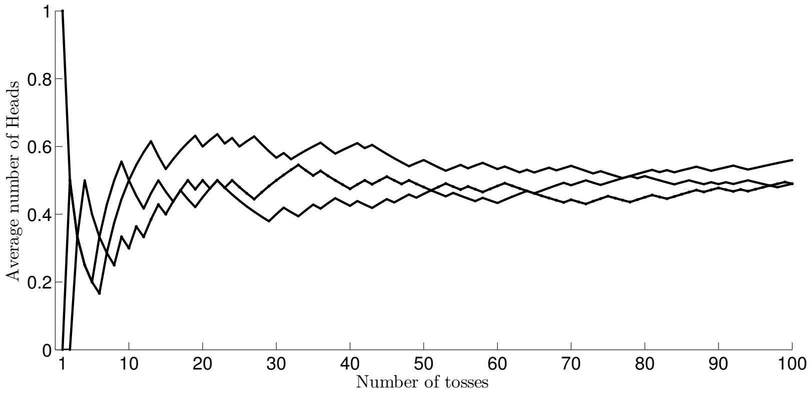

Fig. 1.2 The average number of Heads in n tosses, where $n = 1, \ldots, 100$.

Similar results can be obtained for the case where the coin is *biased*, with a probability of Heads of p, say. Here are some typical *probability* questions.

- What is the probability of x Heads in 100 tosses?
- What is the expected number of Heads?
- How long does one have to wait until the first Head is tossed?
- How fast does the average number of Heads converge to p?

A statistical analysis would start from observed data of the experiment — for example, all the outcomes of 100 tosses are known. Suppose the probability of Heads p is not known. Typical *statistics* questions are:

- Is the coin fair?
- How can p be best estimated from the data?
- How accurate/reliable would such an estimate be?

The mathematical models that are used to describe random experiments consist of three building blocks: a *sample space*, a set of *events*, and a *probability*. We will now describe each of these objects.

1.2 Sample Space

Although we cannot predict the outcome of a random experiment with certainty, we usually can specify a set of possible outcomes. This gives the first ingredient in our model for a random experiment.

Definition 1.1. (Sample Space). The **sample space** Ω of a random experiment is the set of all possible outcomes of the experiment.

Examples of random experiments with their sample spaces are:

1. Cast two dice consecutively and observe their face values.
$$\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}\ .$$
2. Measure the lifetime of a machine in days.
$$\Omega = \mathbb{R}_+ = \{\text{ positive real numbers }\}\,.$$
3. Count the number of arriving calls at an exchange during a specified time interval.
$$\Omega = \{0,1,\ldots\}\,.$$
4. Measure the heights of 10 people.
$$\Omega = \{(x_1,\ldots,x_{10}) : x_i \geqslant 0, i=1,\ldots,10\} = \mathbb{R}_+^{10}\ .$$
Here $(x_1,\ldots,x_{10})$ represents the outcome that the height of the first selected person is x_1, the height of the second person is x_2, and so on.

Notice that for modeling purposes it is often easier to take the sample space larger than is strictly necessary. For example, the actual lifetime of a machine would in reality not span the entire positive real axis, and the heights of the 10 selected people would not exceed 9 feet.

1.3 Events

Often we are not interested in a single outcome but in whether or not one of a *group* of outcomes occurs.

Definition 1.2. (Event). An **event** is a subset of the sample space Ω to which a probability can be assigned.

Events will be denoted by capital letters $A, B, C, \ldots$. We say that event A **occurs** if the outcome of the experiment is one of the elements in A.

Examples of events are:

1. The event that the sum of two dice is 10 or more:
$$A = \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}\ .$$
2. The event that a machine is functioning for less than 1000 days:
$$A = [0,1000)\ .$$

3. The event that out of a group of 50 people 5 are left-handed:

$$A = \{5\} .$$

Example 1.2 (Coin Tossing). Suppose that a coin is tossed 3 times, and that we record either Heads or Tails at every toss. The sample space can then be written as

$$\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\} ,$$

where, for instance, HTH means that the first toss is Heads, the second Tails, and the third Heads. An alternative (but equivalent) sample space is the set $\{0, 1\}^3$ of binary vectors of length 3; for example, HTH corresponds to (1,0,1) and THH to (0,1,1).

The event A that the third toss is Heads is

$$A = \{\text{HHH, HTH, THH, TTH}\} .$$

Since events are sets, we can apply the usual set operations to them, as illustrated in the *Venn diagrams* in Figure 1.3.

1. The set $A \cap B$ (A **intersection** B) is the event that A *and* B both occur.
2. The set $A \cup B$ (A **union** B) is the event that A *or* B *or* both occur.
3. The event A^c (A **complement**) is the event that A does *not* occur.
4. If $B \subset A$ (B is a **subset** of A) then event B is said to *imply* event A.

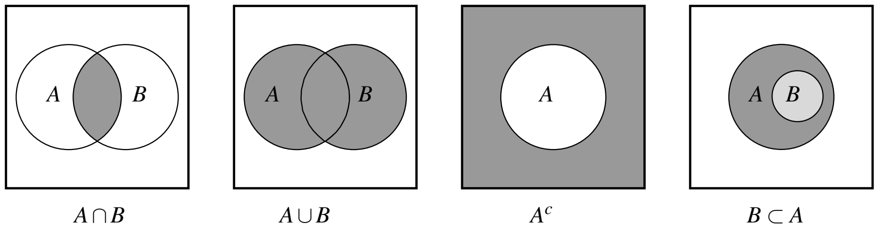

Fig. 1.3 Venn diagrams of set operations. Each square represents the sample space Ω.

Two events A and B which have no outcomes in common, that is, $A \cap B = \emptyset$ (empty set), are called **disjoint** events.

Example 1.3 (Casting Two Dice). Suppose we cast two dice consecutively. The sample space is $\Omega = \{(1, 1), (1, 2), \ldots, (1, 6), (2, 1), \ldots, (6, 6)\}$. Let $A = \{(6, 1), \ldots, (6, 6)\}$ be the event that the first die is 6, and let $B = \{(1, 6), \ldots, (6, 6)\}$ be the event that the second die is 6. Then $A \cap B = \{(6, 1), \ldots, (6, 6)\} \cap \{(1, 6), \ldots, (6, 6)\} = \{(6, 6)\}$ is the event that both dice are 6.

Example 1.4 (System Reliability). In Figure 1.4 three systems are depicted, each consisting of 3 unreliable components. The *series* system works if all components work; the *parallel* system works if at least one of the components works; and the *2-out-of-3 system* works if at least 2 out of 3 components work.

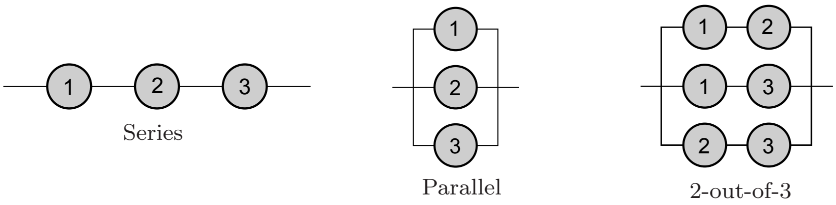

Fig. 1.4 Three unreliable systems.

Let A_i be the event that the i-th component is functioning, $i = 1, 2, 3$; and let D_a, D_b, D_c be the events that respectively the series, parallel, and 2-out-of-3 system is functioning. Then, $D_a = A_1 \cap A_2 \cap A_3$ and $D_b = A_1 \cup A_2 \cup A_3$. Also,

$$\begin{aligned} D_c &= (A_1 \cap A_2 \cap A_3) \cup (A_1^c \cap A_2 \cap A_3) \cup (A_1 \cap A_2^c \cap A_3) \cup (A_1 \cap A_2 \cap A_3^c) \\ &= (A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_2 \cap A_3) \ . \end{aligned}$$

Two useful results in the theory of sets are the following, due to De Morgan:

Theorem 1.1. (De Morgan's Laws). If $\{A_i\}$ is a collection of sets, then

$$\left(\bigcup_i A_i\right)^c = \bigcap_i A_i^c \tag{1.1}$$

and

$$\left(\bigcap_i A_i\right)^c = \bigcup_i A_i^c \ . \tag{1.2}$$

Proof. If we interpret A_i as the event that component i works in Example 1.4, then the left-hand side of (1.1) is the event that the parallel system is not working. The right-hand side of (1.1) is the event that all components are not working. Clearly these two events are identical. The proof for (1.2) follows from a similar reasoning; see also Problem 1.2. □

☞ 18

1.4 Probability

The third ingredient in the model for a random experiment is the specification of the probability of the events. It tells us how *likely* it is that a particular event will occur.

Definition 1.3. (Probability). A **probability** $\mathbb{P}$ is a function which assigns a number between 0 and 1 to each event, and which satisfies the following rules:

1. $0 \leqslant \mathbb{P}(A) \leqslant 1$.
2. $\mathbb{P}(\Omega) = 1$.
3. For any sequence $A_1, A_2, \ldots$ of *disjoint* events we have

Sum Rule:
$$\mathbb{P}\big(\bigcup_i A_i\big) = \sum_i \mathbb{P}(A_i)\ . \tag{1.3}$$

The crucial property (1.3) is called the **sum rule** of probability. It simply states that if an event can happen in several distinct ways (expressed as a union of events, none of which are overlapping), then the probability that at least one of these events happens (that is, the probability of the union) is simply the sum of the probabilities of the individual events. Figure 1.5 illustrates that the probability $\mathbb{P}$ has the properties of a *measure*. However, instead of measuring lengths, areas, or volumes, $\mathbb{P}(A)$ measures the likelihood or probability of an event A as a number between 0 and 1.

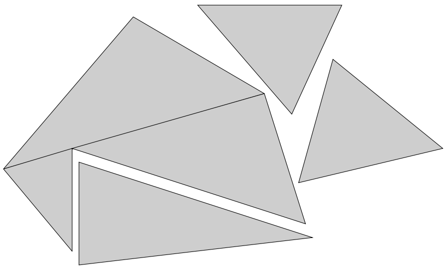

Fig. 1.5 A probability rule $\mathbb{P}$ has exactly the same properties as an area measure. For example, the total area of the union of the non-overlapping triangles is equal to the sum of the areas of the individual triangles.

The following theorem lists some important properties of a probability measure. These properties are direct consequences of the three rules defining a probability measure.

Theorem 1.2. (Properties of a Probability). Let A and B be events and $\mathbb{P}$ a probability. Then,

1. $\mathbb{P}(\emptyset) = 0$,
2. if $A \subset B$, then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$,
3. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Proof.

1. Since $\Omega = \Omega \cup \emptyset$ and $\Omega \cap \emptyset = \emptyset$, it follows from the sum rule that $\mathbb{P}(\Omega) = \mathbb{P}(\Omega) + \mathbb{P}(\emptyset)$. Therefore, by Rule 2 of Definition 1.3, we have $1 = 1 + \mathbb{P}(\emptyset)$; from which it follows that $\mathbb{P}(\emptyset) = 0$.
2. If $A \subset B$, then $B = A \cup (B \cap A^c)$, where A and $B \cap A^c$ are disjoint. Hence, by the sum rule, $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$, which (by Rule 1) is greater than or equal to $\mathbb{P}(A)$.
3. $\Omega = A \cup A^c$, where A and A^c are disjoint. Hence, by the sum rule and Rule 2: $1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$, and thus $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
4. Write $A \cup B$ as the disjoint union of A and $B \cap A^c$. Then, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$. Also, $B = (A \cap B) \cup (B \cap A^c)$, so that $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \cap A^c)$. Combining these two equations gives $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. □

We have now completed our general model for a random experiment. Of course for any *specific* model we must carefully specify the sample space Ω and probability $\mathbb{P}$ that best describe the random experiment.

Example 1.5 (Casting a Die). Consider the experiment where a fair die is cast. How should we specify Ω and $\mathbb{P}$? Obviously, $\Omega = \{1, 2, \ldots, 6\}$; and common sense dictates that we should define $\mathbb{P}$ by

$$\mathbb{P}(A) = \frac{|A|}{6}, \quad A \subset \Omega ,$$

where $|A|$ denotes the number of elements in set A. For example, the probability of getting an even number is $\mathbb{P}(\{2, 4, 6\}) = 3/6 = 1/2$.

In many applications the sample space is *countable*: $\Omega = \{a_1, a_2, \ldots, a_n\}$ or $\Omega = \{a_1, a_2, \ldots\}$. Such a sample space is said to be **discrete**. The easiest way to specify a probability $\mathbb{P}$ on a discrete sample space is to first assign a probability p_i to each **elementary event** $\{a_i\}$ and then to define

$$\mathbb{P}(A) = \sum_{i : a_i \in A} p_i \text{ for all } A \subset \Omega .$$

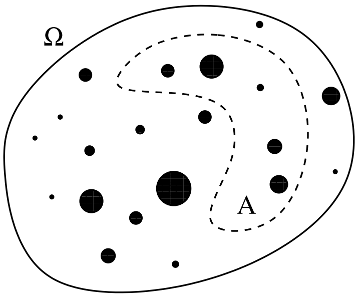

Fig. 1.6 A discrete sample space.

This idea is graphically represented in Figure 1.6. Each element a_i in the sample space is assigned a probability weight p_i represented by a black dot. To find the probability of an event A we have to sum up the weights of all the elements in the set A.

Again, it is up to the modeler to properly specify these probabilities. Fortunately, in many applications all elementary events are *equally likely*, and thus the probability of each elementary event is equal to 1 divided by the total number of elements in Ω. In such case the probability of an event $A \subset \Omega$ is simply

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\text{Number of elements in } A}{\text{Number of elements in } \Omega} \ ,$$

provided that the total number of elements in Ω is finite. The calculation of such probabilities thus reduces to *counting*; see Problem 1.6. ☞ 19

When the sample space is not countable, for example $\Omega = \mathbb{R}_+$, it is said to be **continuous**.

Example 1.6 (Drawing a Random Point in the Unit Interval). We draw at random a point in the interval $[0, 1]$ such that each point is equally likely to be drawn. How do we specify the model for this experiment?

The sample space is obviously $\Omega = [0, 1]$, which is a continuous sample space. We cannot define $\mathbb{P}$ via the elementary events $\{x\}$, $x \in [0, 1]$ because each of these events has probability 0. However, we can define $\mathbb{P}$ as follows. For each $0 \leqslant a \leqslant b \leqslant 1$, let

$$\mathbb{P}([a, b]) = b - a \ .$$

This completely defines $\mathbb{P}$. In particular, the probability that a point will fall into any (sufficiently nice) set A is equal to the *length* of that set.

Describing a random experiment by specifying explicitly the sample space and the probability measure is not always straightforward or necessary. Sometimes it is useful to model only certain *observations* on the experiment. This is where *random variables* come into play, and we will discuss these in Chapter 2. ☞ 23

1.5 Conditional Probability and Independence

How do probabilities change when we know that some event $B \subset \Omega$ has occurred? Thus, we know that the outcome lies in B. Then A will occur if and only if $A \cap B$ occurs, and the relative chance of A occurring is therefore $\mathbb{P}(A \cap B)/\mathbb{P}(B)$, which is called the *conditional probability* of A given B. The situation is illustrated in Figure 1.7.

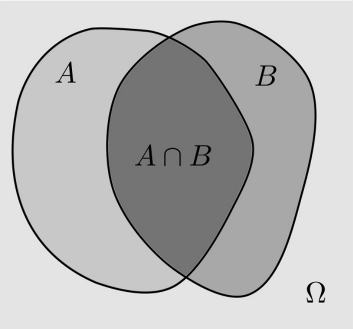

Fig. 1.7 What is the probability that A occurs given that the outcome is known to lie in B?

Definition 1.4. (Conditional Probability). The **conditional probability** of A given B (with $\mathbb{P}(B) \neq 0$) is defined as:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} . \tag{1.4}$$

Example 1.7 (Casting Two Dice). We cast two fair dice consecutively. Given that the sum of the dice is 10, what is the probability that one 6 is cast? Let B be the event that the sum is 10:

$$B = \{(4,6),(5,5),(6,4)\} .$$

Let A be the event that one 6 is cast:

$$A = \{(1,6),\ldots,(5,6),(6,1),\ldots,(6,5)\} .$$

Then, $A \cap B = \{(4,6),(6,4)\}$. And, since for this experiment all elementary events are equally likely, we have

$$\mathbb{P}(A \mid B) = \frac{2/36}{3/36} = \frac{2}{3} .$$

Example 1.8 (Monty Hall Problem). Consider a quiz in which the final contestant is to choose a prize which is hidden behind one of three curtains (A, B, or C). Suppose without loss of generality that the contestant chooses curtain A. Now the quiz master (Monty) always opens one of the other curtains: if the prize is behind B, Monty opens C; if the prize is behind C, Monty opens B; and if the prize is behind A, Monty opens B or C with equal probability, e.g., by tossing a coin (of course the contestant does not see Monty tossing the coin!).

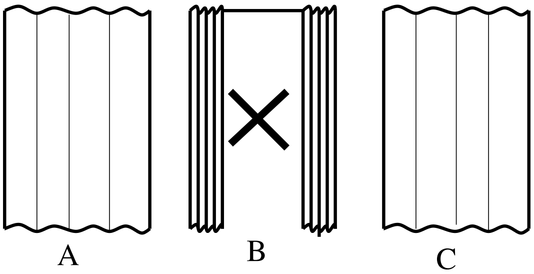

Fig. 1.8 Given that Monty opens curtain B, should the contestant stay with his/her original choice (A) or switch to the other unopened curtain (C)?

Suppose, again without loss of generality, that Monty opens curtain B. The contestant is now offered the opportunity to switch to curtain C. Should the contestant stay with his/her original choice (A) or switch to the other unopened curtain (C)?

Notice that the sample space here consists of 4 possible outcomes: Ac: The prize is behind A, and Monty opens C; Ab: The prize is behind A, and Monty opens B; Bc: The prize is behind B, and Monty opens C; and Cb: The prize is behind C, and Monty opens B. Let A, B, C be the events that the prize is behind A, B, and C, respectively. Note that $A = \{Ac, Ab\}$, $B = \{Bc\}$, and $C = \{Cb\}$; see Figure 1.9.

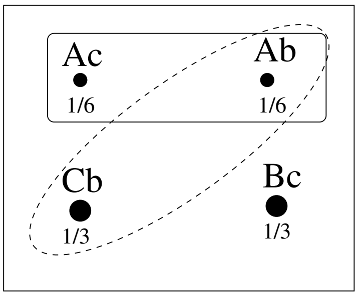

Fig. 1.9 The sample space for the Monty Hall problem.

Now, obviously $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C)$, and since Ac and Ab are equally likely, we have $\mathbb{P}(\{Ab\}) = \mathbb{P}(\{Ac\}) = 1/6$. Monty opening curtain B means that we have information that event $\{Ab, Cb\}$ has occurred. The probability that the prize is behind A given this event is therefore

$$\mathbb{P}(A \,|\, \text{B is opened}) = \frac{\mathbb{P}(\{Ac, Ab\} \cap \{Ab, Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\mathbb{P}(\{Ab\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{3}} = \frac{1}{3} \,.$$

This is what is to be expected: the fact that Monty opens a curtain does not give any extra information that the prize is behind A. Obviously, $\mathbb{P}(B \,|\, \text{B is opened}) = 0$. It follows then that $\mathbb{P}(C \,|\, \text{B is opened})$ must be $2/3$, since the conditional probabilities must sum up to 1. Indeed,

$$\mathbb{P}(C \,|\, \text{B is opened}) = \frac{\mathbb{P}(\{Cb\} \cap \{Ab, Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\mathbb{P}(\{Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\frac{1}{3}}{\frac{1}{6} + \frac{1}{3}} = \frac{2}{3} \,.$$

Hence, given the information that B is opened, it is twice as likely that the prize is behind C than behind A. Thus, the contestant should switch!

1.5.1 Product Rule

By the definition of conditional probability (1.4) we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \,|\, A) \,.$$

It is not difficult to generalize this to n intersections $A_1 \cap A_2 \cap \cdots \cap A_n$, which we abbreviate as $A_1 A_2 \cdots A_n$. This gives the **product rule** of probability. We leave the proof as an exercise; see Problem 1.11.

☞ 20

Theorem 1.3. (Product Rule). Let $A_1, \ldots, A_n$ be a sequence of events with $\mathbb{P}(A_1 \cdots A_{n-1}) > 0$. Then,

$$\mathbb{P}(A_1 \cdots A_n) = \\ \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_1 A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1}) \,. \tag{1.5}$$

Example 1.9 (Urn Problem). We draw consecutively 3 balls from an urn with 5 white and 5 black balls, without putting them back. What is the probability that all drawn balls will be black?

Let A_i be the event that the i-th ball is black. We wish to find the probability of $A_1 A_2 A_3$, which by the product rule (1.5) is

$$\mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_1 A_2) = \frac{5}{10}\,\frac{4}{9}\,\frac{3}{8} \approx 0.083 \,.$$

Example 1.10 (Birthday Problem). What is the probability that in a group of n people all have different birthdays? We can use the product rule. Let A_i be the event that the first i people have different birthdays, $i = 1, 2, \ldots$. Note that $\cdots \subset A_3 \subset A_2 \subset A_1$. Therefore, $A_n = A_1 \cap A_2 \cap \cdots \cap A_n$, and thus by the product rule

$$\mathbb{P}(A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_2) \cdots \mathbb{P}(A_n \,|\, A_{n-1}) \ .$$

Now $\mathbb{P}(A_k \,|\, A_{k-1}) = (365 - k + 1)/365$, because given that the first $k-1$ people have different birthdays, there are no duplicate birthdays among the first k people if and only if the birthday of the k-th person is chosen from the $365 - (k-1)$ remaining birthdays. Thus, we obtain

$$\mathbb{P}(A_n) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365 - n + 1}{365}, \quad n \geqslant 1 \ . \tag{1.6}$$

A graph of $\mathbb{P}(A_n)$ against n is given in Figure 1.10. Note that the probability $\mathbb{P}(A_n)$ rapidly decreases to zero. For $n = 23$ the probability of having no duplicate birthdays is already less than 1/2.

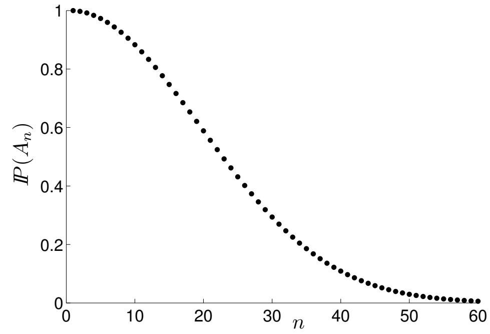

Fig. 1.10 The probability of having no duplicate birthday in a group of n people against n.

1.5.2 Law of Total Probability and Bayes' Rule

Suppose that $B_1, B_2, \ldots, B_n$ is a **partition** of Ω. That is, $B_1, B_2, \ldots, B_n$ are disjoint and their union is Ω; see Figure 1.11.

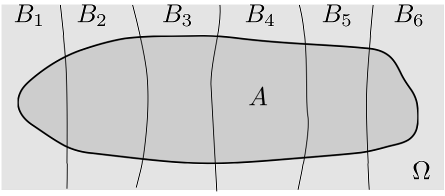

Fig. 1.11 A partition $B_1, \ldots, B_6$ of the sample space Ω. Event A is partitioned into events $A \cap B_1$, $\ldots$, $A \cap B_6$.

A partitioning of the state space can sometimes make it easier to calculate probabilities via the following theorem.

Theorem 1.4. (Law of Total Probability). Let A be an event and let $B_1, B_2, \ldots, B_n$ be a partition of Ω. Then,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \,|\, B_i)\, \mathbb{P}(B_i) \; . \tag{1.7}$$

Proof. The sum rule gives $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i)$, and by the product rule we have $\mathbb{P}(A \cap B_i) = \mathbb{P}(A \,|\, B_i)\, \mathbb{P}(B_i)$. □

Combining the law of total probability with the definition of conditional probability gives **Bayes' Rule**:

Theorem 1.5. (Bayes Rule). Let A be an event with $\mathbb{P}(A) > 0$ and let $B_1, B_2, \ldots, B_n$ be a partition of Ω. Then,

$$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(A \,|\, B_j)\, \mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A \,|\, B_i)\, \mathbb{P}(B_i)} \; . \tag{1.8}$$

Proof. By definition, $\mathbb{P}(B_j \,|\, A) = \mathbb{P}(A \cap B_j)/\mathbb{P}(A) = \mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j)/\mathbb{P}(A)$. Now apply the law of total probability to $\mathbb{P}(A)$. □

Example 1.11 (Quality Control Problem). A company has three factories (1, 2, and 3) that produce the same chip, each producing 15%, 35%, and 50% of the total production. The probability of a faulty chip at factory 1, 2, 3 is 0.01, 0.05, 0.02, respectively. Suppose we select randomly a chip from the total production and this chip turns out to be faulty. What is the conditional probability that this chip has been produced in factory 1?

Let B_i denote the event that the chip has been produced in factory i. The $\{B_i\}$ form a partition of Ω. Let A denote the event that the chip is faulty. We are given the information that $\mathbb{P}(B_1) = 0.15, \mathbb{P}(B_2) = 0.35, \mathbb{P}(B_3) = 0.5$ as well as $\mathbb{P}(A \,|\, B_1) = 0.01$, $\mathbb{P}(A \,|\, B_2) = 0.05$, $\mathbb{P}(A \,|\, B_3) = 0.02$.

We wish to find $\mathbb{P}(B_1 \,|\, A)$, which by Bayes' rule is given by

$$\mathbb{P}(B_1 \,|\, A) = \frac{0.15 \times 0.01}{0.15 \times 0.01 + 0.35 \times 0.05 + 0.5 \times 0.02} = 0.052 \; .$$

1.5.3 Independence

Independence is a very important concept in probability and statistics. Loosely speaking it models the *lack of information* between events. We say events A and B are *independent* if the knowledge that B has occurred does not change the probability that A occurs. More precisely, A and B are said to be independent if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Since $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$, an alternative definition of independence is: A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. This definition covers the case where $B = \emptyset$.

We can extend the definition to arbitrarily many events (compare with the product rule (1.5)):

Definition 1.5. (Independence). The events $A_1, A_2, \ldots,$ are said to be **independent** if for any k and any choice of distinct indices $i_1, \ldots, i_k$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\,\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}) \ . \tag{1.9}$$

Remark 1.1. In most cases independence of events is a *model assumption.* That is, $\mathbb{P}$ is chosen such that certain events are independent.

Example 1.12 (Coin Tossing and the Binomial Law). We toss a coin n times. The sample space can be written as the set of binary n-tuples:

$$\Omega = \{(\underbrace{0, \ldots, 0}_{n \text{ times}}), \ldots, (1, \ldots, 1)\} \ .$$

Here, 0 represents Tails and 1 represents Heads. For example, the outcome $(0, 1, 0, 1, \ldots)$ means that the first time Tails is thrown, the second time Heads, the third times Tails, the fourth time Heads, etc.

How should we define $\mathbb{P}$? Let A_i denote the event of Heads at the i-th throw, $i = 1, \ldots, n$. Then, $\mathbb{P}$ should be such that the following holds.

- The events $A_1, \ldots, A_n$ should be *independent* under $\mathbb{P}$.
- $\mathbb{P}(A_i)$ should be the same for all i. Call this known or unknown probability p $(0 \leqslant p \leqslant 1)$.

These two rules completely specify $\mathbb{P}$. For example, the probability that the first k throws are Heads and the last $n - k$ are Tails is

$$\begin{aligned}\mathbb{P}(\{(\underbrace{1, 1, \ldots, 1}_{k \text{ times}}, \underbrace{0, 0, \ldots, 0}_{n-k \text{ times}})\}) &= \mathbb{P}(A_1 \cap \cdots \cap A_k \cap A_{k+1}^c \cap \cdots \cap A_n^c) \\ &= \mathbb{P}(A_1) \cdots \mathbb{P}(A_k)\,\mathbb{P}(A_{k+1}^c) \cdots \mathbb{P}(A_n^c) = p^k (1-p)^{n-k}.\end{aligned}$$

Note that if A_i and A_j are independent, then so are A_i and A_j^c; see Problem 1.12.

Let B_k be the event that k Heads are thrown in total. The probability of this event is the sum of the probabilities of elementary events $\{(x_1, \ldots, x_n)\}$ for which $x_1 + \cdots + x_n = k$. Each of these events has probability $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ of these. We thus obtain the **binomial law**:

$$\mathbb{P}(B_k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n\ . \tag{1.10}$$

Example 1.13 (Geometric Law). There is another important law associated with the coin toss experiment. Let C_k be the event that Heads appears for the first time at the k-th toss, $k = 1, 2, \ldots$. Then, using the same events $\{A_i\}$ as in the previous example, we can write

$$C_k = A_1^c \cap A_2^c \cap \cdots \cap A_{k-1}^c \cap A_k\ .$$

Using the independence of $A_1^c, \ldots, A_{k-1}^c, A_k$, we obtain the **geometric law**:

$$\begin{aligned}\mathbb{P}(C_k) &= \mathbb{P}(A_1^c) \cdots \mathbb{P}(A_{k-1}^c)\, \mathbb{P}(A_k) \\ &= \underbrace{(1-p) \cdots (1-p)}_{k-1 \text{ times}}\, p = (1-p)^{k-1}\, p\ .\end{aligned}$$

1.6 Problems

1.1. For each of the five random experiments at the beginning of Section 1.1 define a convenient sample space.

1.2. Interpret De Morgan's rule (1.2) in terms of an unreliable series system.

1.3. Let $\mathbb{P}(A) = 0.9$ and $\mathbb{P}(B) = 0.8$. Show that $\mathbb{P}(A \cap B) \geqslant 0.7$.

1.4. Throw two fair dice one after the other.

a. What is the probability that the second die is 3, given that the sum of the dice is 6?
b. What is the probability that the first die is 3 and the second is not 3?

1.5. An "expert" wine taster has to try to match 6 glasses of wine to 6 wine labels. Each label can only be chosen once.

a. Formulate a sample space Ω for this experiment.
b. Assuming the wine taster is a complete fraud, define an appropriate probability $\mathbb{P}$ on the sample space.

c. What is the probability that the wine taster guesses 4 labels correctly, assuming he/she guesses them randomly?

1.6. Many counting problems can be cast into the framework of drawing k balls from an urn with n balls, numbered $1, \ldots, n$; see Figure 1.12.

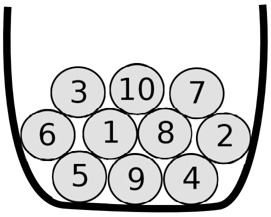

Fig. 1.12 Draw k balls from an urn with $n = 10$ numbered balls.

The drawing can be done in several ways. Firstly, the k balls could be drawn one-by-one or all at the same time. In the first case the **order** in which the balls are drawn can be noted. In the second case we can still assume that the balls are drawn one-by-one, but we do not note the order. Secondly, once a ball is drawn, it can either be put back into the urn or be left out. This is called drawing with and without **replacement**, respectively. There are thus four possible random experiments. Prove that for each of these experiments the total number of possible outcomes is the following.

1. Ordered, with replacement: n^k.
2. Ordered, without replacement: ${}^nP_k = n(n-1)\cdots(n-k+1)$.
3. Unordered, without replacement: ${}^nC_k = \binom{n}{k} = \frac{{}^nP_k}{k!} = \frac{n!}{(n-k)!\,k!}$.
4. Unordered, with replacement: $\binom{n+k-1}{k}$.

Provide a sample space for each of these experiments. Hint: it is important to use a notation that clearly shows whether the arrangements of numbers are ordered or not. Denote ordered arrangements by *vectors*, e.g., $(1, 1, 2)$, and unordered arrangements by *sets*, e.g., $\{1, 2, 3\}$ or *multisets*, e.g., $\{1, 1, 2\}$.

1.7. Formulate the birthday problem in terms of an urn experiment, as in Problem 1.6, and derive the probability (1.6) by counting.

1.8. Three cards are drawn from a full deck of cards, noting the order. The cards may be numbered from 1 to 52.

a. Give the sample space. Is each elementary event equally likely?
b. What is the probability that we draw three Aces?
c. What is the probability that we draw one Ace, one King, and one Queen (not necessarily in that order)?
d. What is the probability that we draw no pictures (no A, K, Q, or J)?

1.9. In a group of 20 people there are three brothers. The group is separated at random into two groups of 10. What is the probability that the brothers are in the same group?

1.10. Two fair dice are thrown.

a. Find the probability that both dice show the same face.
b. Find the same probability, using the extra information that the sum of the dice is not greater than 4.

1.11. Prove the product rule (1.5). Hint: first show it for the case of 3 events:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)\,\mathbb{P}(C \mid A \cap B) \ .$$

1.12. If A and B are independent events, then A and B^c are also independent. Prove this.

1.13. Select at random 3 people from a large population. What is the probability that they all have the same birthday?

1.14. In a large population 40% votes for A and 60% for B. Suppose we select at random 10 people. What is the probability that in this group exactly 4 people will vote for A?

1.15. A certain AIDS test has a 0.98 probability of giving a Positive result when the blood is infected, and a 0.07 probability of giving a Positive result when the blood is not infected (a so-called false positive). Suppose 1% of the population carries the HIV virus.

a. Using the law of total probability, what is the probability that the test is Positive for a randomly selected person?
b. What is the probability that a person is indeed infected, *given* that the test yields a Positive result?

1.16. A box has three identical-looking coins. However the probability of success (Heads) is different for each coin: coin 1 is fair, coin 2 has a success probability of 0.4 and coin 3 has a success probability of 0.6. We pick one coin at random and throw it 100 times. Suppose 43 Heads come up. Using this information assess the probability that coin 1, 2, or 3 was chosen.

1.17. In a binary communication channel, 0s and 1s are transmitted with equal probability. The probability that a 0 is correctly received (as a 0) is 0.95. The probability that a 1 is correctly received (as a 1) is 0.99. Suppose we receive a 0, what is the probability that, in fact, a 1 was sent?

1.18. A fair coin is tossed 20 times.

a. What is the probability of exactly 10 Heads?
b. What is the probability of 15 or more Heads?

1.19. Two fair dice are cast (at the same time) until their sum is 12.

a. What is the probability that we have to wait exactly 10 tosses?
b. What is the probability that we do not have to wait more than 100 tosses?

1.20. Independently throw 10 balls into one of three boxes, numbered 1, 2, and 3, with probabilities 1/4, 1/2, and 1/4, respectively.

a. What is the probability that box 1 has 2 balls, box 2 has 5 balls, and box 3 has 3 balls?
b. What is the probability that box 1 remains empty?

1.21. Implement a Matlab program that performs 100 tosses with a fair die. Hint: use the `rand` and `ceil` functions, where `ceil(x)` returns the smallest integer larger than or equal to `x`.

1.22. For each of the four urn experiments in Problem 1.6 implement a Matlab program that simulates the experiment. Hint: in addition to the functions `rand` and `ceil`, you may wish to use the `sort` function.

1.23. Verify your answers for Problem 1.20 with a computer simulation, where the experiment is repeated many times.

Chapter 2

Random Variables and Probability Distributions

Specifying a model for a random experiment via a complete description of the sample space Ω and probability measure $\mathbb{P}$ may not always be necessary or convenient. In practice we are only interested in certain *numerical measurements* pertaining to the experiment. Such random measurements can be included into the model via the notion of a *random variable*.

2.1 Random Variables

Definition 2.1. (Random Variable). A **random variable** is a *function* from the sample space Ω to $\mathbb{R}$.

Example 2.1 (Sum of Two Dice). We throw two fair dice and note the sum of their face values. If we throw the dice consecutively and observe both throws, the sample space is $\Omega = \{(1,1),\ldots,(6,6)\}$. The function X defined by $X(i,j) = i + j$ is a random variable which maps the outcome (i,j) to the sum $i + j$, as depicted in Figure 2.1.

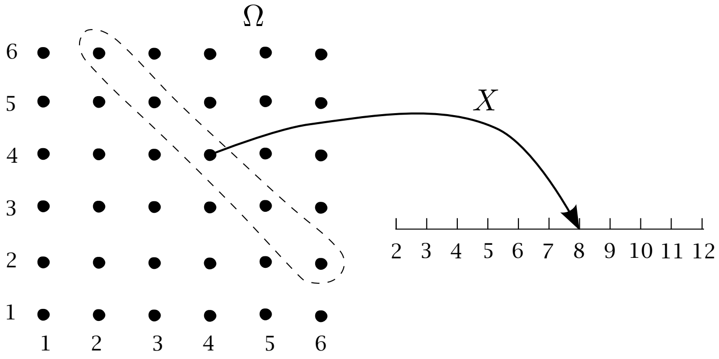

Fig. 2.1 Random variable X represents the sum of two dice.

Note that five outcomes in the sample space are mapped to 8. A natural notation for the corresponding set of outcomes is $\{X = 8\}$. Since all outcomes in Ω are equally likely, we have

$$\mathbb{P}(\{X = 8\}) = \frac{5}{36} \; .$$

This notation is very suggestive and convenient. From a non-mathematical viewpoint we can interpret X as a "random" variable. That is, a variable that can take several values with certain probabilities. In particular, it is not difficult to check that

$$\mathbb{P}(\{X = x\}) = \frac{6 - |7 - x|}{36}, \quad x = 2, \ldots, 12 \; .$$

Although random variables are, mathematically speaking, *functions*, it is often convenient to view them as observations of a random experiment that has not yet taken place. In other words, a random variable is considered as a measurement that becomes available *tomorrow*, while all the thinking about the measurement can be carried out *today*. For example, we can specify today exactly the probabilities pertaining to the random variables.

We often denote random variables with *capital* letters from the last part of the alphabet, e.g., X, $X_1, X_2, \ldots, Y, Z$. Random variables allow us to use natural and intuitive notations for certain events, such as $\{X = 10\}$, $\{X > 1000\}$, $\{\max(X, Y) \leqslant Z\}$, etc.

☞ 17 **Example 2.2 (Coin Tossing).** In Example 1.12 we constructed a probability model for the random experiment where a biased coin is tossed n times. Suppose we are not interested in a specific outcome but only in the total number of Heads, X, say. In particular, we would like to know the probability that X takes certain values between 0 and n. Example 1.12 suggests that

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n \; , \tag{2.1}$$

providing all the information about X that we could possibly wish to know. To justify (2.1) mathematically, we can reason as in Example 2.1. First, define X as the function that assigns to each outcome $\omega = (x_1, \ldots, x_n)$ the number $x_1 + \cdots + x_n$. Thus, X is a random variable in mathematical terms; that is, a function. Second, the event B_k that there are exactly k Heads in n throws can be written as

$$B_k = \{\omega \in \Omega : X(\omega) = k\} \; .$$

If we write this as $\{X = k\}$, and further abbreviate $\mathbb{P}(\{X = k\})$ to $\mathbb{P}(X = k)$, then we obtain (2.1) directly from (1.10).

We give some more examples of random variables without specifying the sample space.

1. The number of defective transistors out of 100 inspected ones.
2. The number of bugs in a computer program.
3. The amount of rain in a certain location in June.
4. The amount of time needed for an operation.

The set of all possible values that a random variable X can take is called the **range** of X. We further distinguish between discrete and continuous random variables:

- **Discrete** random variables can only take *countably many* values.
- **Continuous** random variables can take a continuous range of values; for example, any value on the positive real line $\mathbb{R}_+$.

2.2 Probability Distribution

Let X be a random variable. We would like to designate the probabilities of events such as $\{X = x\}$ and $\{a \leqslant X \leqslant b\}$. If we can specify all probabilities involving X, we say that we have determined the **probability distribution** of X. One way to specify the probability distribution is to give the probabilities of all events of the form $\{X \leqslant x\}$, $x \in \mathbb{R}$. This leads to the following definition.

Definition 2.2. (Cumulative Distribution Function). The **cumulative distribution function** (cdf) of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = \mathbb{P}(X \leqslant x), \quad x \in \mathbb{R} \; .$$

Note that we have used $\mathbb{P}(X \leqslant x)$ as a shorthand notation for $\mathbb{P}(\{X \leqslant x\})$. From now on we will use this type of abbreviation throughout the book. In Figure 2.2 the graph of a general cdf is depicted.

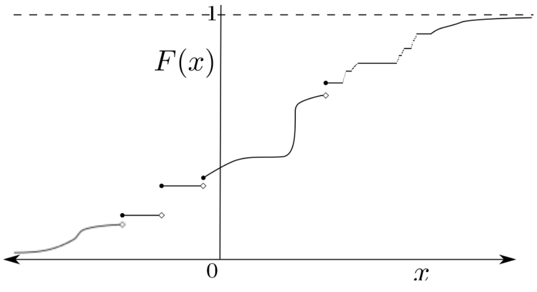

Fig. 2.2 A cumulative distribution function (cdf).

Theorem 2.1. (Properties of Cdf). Let F be the cdf of a random variable X. Then,

1. F is bounded between 0 and 1: $0 \leqslant F(x) \leqslant 1$,
2. F is increasing: if $x \leqslant y$, then $F(x) \leqslant F(y)$,
3. F is right-continuous: $\lim_{h \downarrow 0} F(x+h) = F(x)$.

Proof.

1. ☞ 9 Let $A = \{X \leqslant x\}$. By Rule 1 in Definition 1.3, $0 \leqslant \mathbb{P}(A) \leqslant 1$.
2. Suppose $x \leqslant y$. Define $A = \{X \leqslant x\}$ and $B = \{X \leqslant y\}$. Then, $A \subset B$ ☞ 10 and, by Theorem 1.2, $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.
3. Take any sequence $x_1, x_2, \ldots$ decreasing to x. We have to show that $\lim_{n\to\infty} \mathbb{P}(X \leqslant x_n) = \mathbb{P}(X \leqslant x)$ or, equivalently, $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$, where $A_n = \{X > x_n\}$ and $A = \{X > x\}$. Let $B_n = \{x_{n-1} \geqslant X > x_n\}$, $n = 1, 2, \ldots$, with x_0 defined as ∞. Then, $A_n = \cup_{i=1}^n B_i$ and $A = \cup_{i=1}^\infty B_i$. Since the $\{B_i\}$ are disjoint, we have by the sum rule:

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \overset{\text{def}}{=} \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \lim_{n\to\infty} \mathbb{P}(A_n) \ ,$$

as had to be shown. □

Conversely, any function F with the above properties can be used to specify the distribution of a random variable X.

If X has cdf F, then the probability that X takes a value in the interval $(a, b]$ (excluding a, including b) is given by

$$\mathbb{P}(a < X \leqslant b) = F(b) - F(a) \ .$$

To see this, note that $\mathbb{P}(X \leqslant b) = \mathbb{P}(\{X \leqslant a\} \cup \{a < X \leqslant b\})$, where the events $\{X \leqslant a\}$ and $\{a < X \leqslant b\}$ are disjoint. Thus, by the sum rule: $F(b) = F(a) + \mathbb{P}(a < X \leqslant b)$, which leads to the result above. Note however that

$$\begin{aligned}
\mathbb{P}(a \leqslant X \leqslant b) &= F(b) - F(a) + \mathbb{P}(X = a) \\
&= F(b) - F(a) + F(a) - F(a-) \\
&= F(b) - F(a-) \ ,
\end{aligned}$$

where $F(a-)$ denotes the limit from below: $\lim_{x \uparrow a} F(x)$.

2.2.1 Discrete Distributions

Definition 2.3. (Discrete Distribution). A random variable X is said to have a **discrete distribution** if $\mathbb{P}(X = x_i) > 0$, $i = 1, 2, \ldots$ for some finite or countable set of values $x_1, x_2, \ldots$, such that $\sum_i \mathbb{P}(X = x_i) = 1$. The **discrete probability density function (pdf)** of X is the function f defined by $f(x) = \mathbb{P}(X = x)$.

We sometimes write f_X instead of f to stress that the discrete probability density function refers to the discrete random variable X. The easiest way to specify the distribution of a discrete random variable is to specify its pdf. Indeed, by the sum rule, if we know $f(x)$ for all x, then we can calculate all possible probabilities involving X. Namely, ☞ 9

$$\mathbb{P}(X \in B) = \sum_{x \in B} f(x) \tag{2.2}$$

for any subset B in the range of X, as illustrated in Figure 2.3.

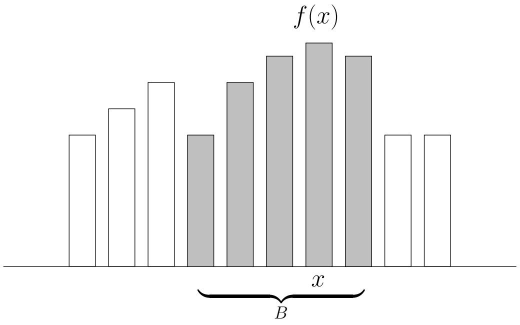

Fig. 2.3 Discrete probability density function.

Example 2.3 (Sum of Two Dice, Continued). Toss two fair dice and let X be the sum of their face values. The discrete pdf is given in Table 2.1, which follows directly from Example 2.1.

Table 2.1 Discrete pdf of the sum of two fair dice.

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

2.2.2 Continuous Distributions

Definition 2.4. (Continuous Distribution). A random variable X with cdf F is said to have a **continuous distribution** if there exists a positive function f with *total integral 1* such that for all $a < b$,

$$\mathbb{P}(a < X \leqslant b) = F(b) - F(a) = \int_a^b f(u)\,\mathrm{d}u\ . \tag{2.3}$$

Function f is called the **probability density function (pdf)** of X.

Remark 2.1. Note that we use the *same* notation f for both the discrete and the continuous pdf, to stress the similarities between the discrete and continuous case. We will even drop the qualifier "discrete" or "continuous" when it is clear from the context with which case we are dealing. Henceforth we will use the notation $X \sim f$ and $X \sim F$ to indicate that X is distributed according to pdf f or cdf F.

In analogy to the discrete case (2.2), once we know the pdf, we can calculate any probability of interest by means of integration:

$$\mathbb{P}(X \in B) = \int_B f(x)\,\mathrm{d}x\ , \tag{2.4}$$

as illustrated in Figure 2.4.

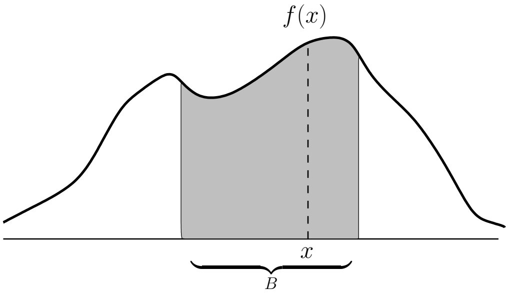

Fig. 2.4 Probability density function (pdf).

Suppose that f and F are the pdf and cdf of a continuous random variable X, as in Definition 2.4. Then F is simply a *primitive* (also called anti-derivative) of f:

$$F(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u\ .$$

Conversely, f is the *derivative* of the cdf F:

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x) = F'(x) \; .$$

It is important to understand that in the continuous case $f(x)$ is not equal to the probability $\mathbb{P}(X = x)$, because the latter is 0 for all x. Instead, we interpret $f(x)$ as the *density* of the probability distribution at x, in the sense that for any small h,

$$\mathbb{P}(x \leqslant X \leqslant x + h) = \int_x^{x+h} f(u)\,\mathrm{d}u \approx h\, f(x) \; . \tag{2.5}$$

Note that $\mathbb{P}(x \leqslant X \leqslant x + h)$ is equal to $\mathbb{P}(x < X \leqslant x + h)$ in this case.

Example 2.4 (Random Point in an Interval). Draw a random number X from the interval of real numbers $[0, 2]$, where each number is equally likely to be drawn. What are the pdf f and cdf F of X? Using the same reasoning as in Example 1.6, we see that ☞ 11

$$\mathbb{P}(X \leqslant x) = F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x/2 & \text{if } 0 \leqslant x \leqslant 2, \\ 1 & \text{if } x > 2. \end{cases}$$

By differentiating F we find

$$f(x) = \begin{cases} 1/2 & \text{if } 0 \leqslant x \leqslant 2, \\ 0 & \text{otherwise.} \end{cases}$$

Note that this density is *constant* on the interval $[0, 2]$ (and zero elsewhere), reflecting the fact that each point in $[0, 2]$ is equally likely to be drawn.

2.3 Expectation

Although all probability information about a random variable is contained in its cdf or pdf, it is often useful to consider various numerical characteristics of a random variable. One such number is the *expectation* of a random variable, which is a "weighted average" of the values that X can take. Here is a more precise definition.

Definition 2.5. (Expectation of a Discrete Random Variable). Let X be a *discrete* random variable with pdf f. The **expectation** (or expected value) of X, denoted as $\mathbb{E}X$, is defined as

$$\mathbb{E}X = \sum_x x\,\mathbb{P}(X = x) = \sum_x x\, f(x) \; . \tag{2.6}$$

The expectation of X is sometimes written as μ_X. It is assumed that the sum in (2.6) is well-defined — possibly ∞ or $-\infty$. One way to interpret the expectation is as a *long-run average payout.* Suppose in a game of dice the payout X (dollars) is the largest of the face values of two dice. To play the game a fee of d dollars must be paid. What would be a fair amount for d? The answer is

$$\begin{aligned} d &= \mathbb{E}X = 1 \times \mathbb{P}(X=1) + 2 \times \mathbb{P}(X=2) + \cdots + 6 \times \mathbb{P}(X=6) \\ &= 1 \times \frac{1}{36} + 2 \times \frac{3}{36} + 3 \times \frac{5}{36} + 4 \times \frac{7}{36} + 5 \times \frac{9}{36} + 6 \times \frac{11}{36} = \frac{161}{36} \approx 4.47\,. \end{aligned}$$

Namely, if the game is played many times, the long-run fraction of tosses where the maximum face value is 1, 2,. . . , 6, is $\frac{1}{36}, \frac{3}{36}, \ldots, \frac{11}{36}$, respectively. Hence, the long-run average payout of the game is the weighted sum of $1, 2, \ldots, 6$, where the weights are the long-run fractions (probabilities). The game is “fair” if the long-run average profit $\mathbb{E}X - d$ is zero.

The expectation can also be interpreted as a *center of mass.* Imagine that point masses with weights $p_1, p_2, \ldots, p_n$ are placed at positions $x_1, x_2, \ldots, x_n$ on the real line; see Figure 2.5.

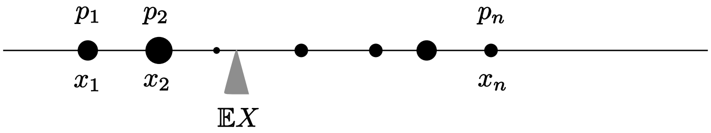

Fig. 2.5 The expectation as a center of mass.

The center of mass — the place where the weights are balanced — is

$$\text{center of mass} = x_1\, p_1 + \cdots + x_n\, p_n\ ,$$

which is exactly the expectation of the discrete variable X that takes values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. An obvious consequence of this interpretation is that for a *symmetric* pdf the expectation is equal to the symmetry point (provided that the expectation exists). In particular, suppose that $f(c+y) = f(c-y)$ for all y. Then,

$$\begin{aligned} \mathbb{E}X &= c\, f(c) + \sum_{x>c} x f(x) + \sum_{x<c} x f(x) \\ &= c\, f(c) + \sum_{y>0} (c+y) f(c+y) + \sum_{y>0} (c-y) f(c-y) \\ &= c\, f(c) + \sum_{y>0} c\, f(c+y) + c \sum_{y>0} f(c-y) = c \sum_{x} f(x) = c\,. \end{aligned}$$

For continuous random variables we can define the expectation in a similar way, replacing the sum with an integral.

Definition 2.6. (Expectation of a Continuous Random Variable). Let X be a *continuous* random variable with pdf f. The **expectation** (or expected value) of X, denoted as $\mathbb{E}X$, is defined as

$$\mathbb{E}X = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x \; . \tag{2.7}$$

If X is a random variable, then a function of X, such as X^2 or $\sin(X)$, is also a random variable. The following theorem simply states that the expected value of a function of X is the weighted average of the values that this function can take.

Theorem 2.2. (Expectation of a Function of a Random Variable). If X is *discrete* with pdf f, then for any real-valued function g

$$\mathbb{E}\, g(X) = \sum_x g(x)\, f(x) \; .$$

Similarly, if X is *continuous* with pdf f, then

$$\mathbb{E}\, g(X) = \int_{-\infty}^{\infty} g(x)\, f(x)\, \mathrm{d}x \; .$$

Proof. The proof is given for the discrete case only, as the continuous case can be proven in a similar way. Let $Y = g(X)$, where X is a discrete random variable with pdf f_X and g is a function. Let f_Y be the (discrete) pdf of the random variable Y. It can be expressed in terms of f_X in the following way:

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x) = \sum_{x:g(x)=y} f_X(x) \; .$$

Thus, the expectation of Y is

$$\begin{aligned}\mathbb{E}Y &= \sum_y y\, f_Y(y) = \sum_y y \sum_{x:g(x)=y} f_X(x) = \sum_y \sum_{x:g(x)=y} y f_X(x) \\ &= \sum_x g(x)\, f_X(x) \; .\end{aligned}$$

□

Example 2.5 (Die Experiment and Expectation). Find $\mathbb{E}X^2$ if X is the outcome of the toss of a fair die. We have

$$\mathbb{E}X^2 = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + \cdots + 6^2 \times \frac{1}{6} = \frac{91}{6} \, .$$

An important consequence of Theorem 2.2 is that the expectation is "linear".

Theorem 2.3. (Properties of the Expectation). For any real numbers a and b, and functions g and h,

1. $\mathbb{E}[a\,X + b] = a\,\mathbb{E}X + b$,
2. $\mathbb{E}[g(X) + h(X)] = \mathbb{E}g(X) + \mathbb{E}h(X)$.

Proof. Suppose X has pdf f. The first statement follows (in the discrete case) from

$$\mathbb{E}(aX + b) = \sum_x (ax + b) f(x) = a \sum_x x\, f(x) + b \sum_x f(x) = a\,\mathbb{E}X + b \, .$$

Similarly, the second statement follows from

$$\begin{aligned}\mathbb{E}(g(X) + h(X)) &= \sum_x (g(x) + h(x)) f(x) = \sum_x g(x) f(x) + \sum_x h(x) f(x) \\ &= \mathbb{E}g(X) + \mathbb{E}h(X) \, .\end{aligned}$$

The continuous case is proven analogously, simply by replacing sums with integrals. □

Another useful numerical characteristic of the distribution of X is the *variance* of X. This number, sometimes written as σ_X^2, measures the *spread* or dispersion of the distribution of X.

Definition 2.7. (Variance and Standard Deviation). The **variance** of a random variable X, denoted as $\mathrm{Var}(X)$, is defined as

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 \, . \tag{2.8}$$

The square root of the variance is called the **standard deviation**. The number $\mathbb{E}X^r$ is called the r-th **moment** of X.

Theorem 2.4. (Properties of the Variance). For any random variable X the following properties hold for the variance.

1. $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
2. $\mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X)$.

Proof. Write $\mathbb{E}X = \mu$, so that $\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2)$. By the linearity of the expectation, the last expectation is equal to the sum $\mathbb{E}X^2 - 2\,\mu\,\mathbb{E}X + \mu^2 = \mathbb{E}X^2 - \mu^2$, which proves the first statement. To prove the second statement, note that the expectation of $a + bX$ is equal to $a + b\mu$. Consequently,

$$\mathrm{Var}(a + bX) = \mathbb{E}(a + bX - (a + b\mu))^2 = \mathbb{E}(b^2(X-\mu)^2) = b^2\mathrm{Var}(X) \; .$$

□

Note that Property 1 in Theorem 2.4 implies that $\mathbb{E}X^2 \geqslant (\mathbb{E}X)^2$, because $\mathrm{Var}(X) \geqslant 0$. This is a special case of a much more general result, regarding the expectation of convex functions. A real-valued function $h(x)$ is said to be **convex** if for each x_0 there exist constants a and b such that (1) $h(x) \geqslant ax+b$ for all x and (2) $h(x_0) = ax_0+b$. Examples are the functions $x \mapsto x^2$, $x \mapsto \mathrm{e}^x$, and $x \mapsto -\ln x$.

Theorem 2.5. (Jensen's Inequality). Let $h(x)$ be a convex function and X a random variable. Then,

$$\mathbb{E}h(X) \geqslant h(\mathbb{E}X) \; . \tag{2.9}$$

Proof. Let $x_0 = \mathbb{E}X$. Because h is convex, there exists constants a and b such that $h(X) \geqslant aX + b$ and $h(x_0) = ax_0 + b$. Hence, $\mathbb{E}h(X) \geqslant \mathbb{E}(aX + b) = ax_0 + b = h(x_0) = h(\mathbb{E}X)$. □

2.4 Transforms

Many probability calculations — such as the evaluation of expectations and variances — are facilitated by the use of *transforms*. We discuss here a number of such transforms.

Definition 2.8. (Probability Generating Function). Let X be a *non-negative* and *integer-valued* random variable with discrete pdf f. The **probability generating function** (PGF) of X is the function G defined by

$$G(z) = \mathbb{E}\, z^X = \sum_{x=0}^{\infty} z^x\, f(x)\;, \quad |z| < R\;,$$

where $R \geqslant 1$ is the **radius of convergence**.

Example 2.6 (Poisson Distribution). Let X have a discrete pdf f given by

$$f(x) = \mathrm{e}^{-\lambda}\, \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots\;.$$

X is said to have a **Poisson distribution**. The PGF of X is given by

$$\begin{aligned} G(z) &= \sum_{x=0}^{\infty} z^x\, \mathrm{e}^{-\lambda}\, \frac{\lambda^x}{x!} \\ &= \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{(z\lambda)^x}{x!} \\ &= \mathrm{e}^{-\lambda} \mathrm{e}^{z\lambda} = \mathrm{e}^{-\lambda(1-z)}\;. \end{aligned}$$

As this is finite for every z, the radius of convergence is here $R = \infty$.

Theorem 2.6. (Derivatives of a PGF). The k-th derivative of a PGF $\mathbb{E}z^X$ can be obtained by *differentiation under the expectation sign*:

$$\begin{aligned} \frac{\mathrm{d}^k}{\mathrm{d}z^k}\mathbb{E}z^X &= \mathbb{E}\frac{\mathrm{d}^k}{\mathrm{d}z^k} z^X \\ &= \mathbb{E}\left[X(X-1)\cdots(X-k+1)z^{X-k}\right] \quad \text{for } |z| < R\;, \end{aligned}$$

where $R \geqslant 1$ is the radius of convergence of the PGF.

☞ 381 *Proof.* The proof is deferred to Appendix B.2. □

Let $G(z)$ be the PGF of a random variable X. Thus, $G(z) = z^0\, \mathbb{P}(X = 0) + z^1\, \mathbb{P}(X = 1) + z^2\, \mathbb{P}(X = 2) + \cdots$. Substituting $z = 0$ gives, $G(0) = \mathbb{P}(X = 0)$. By Theorem 2.6 the derivative of G is

$$G'(z) = \mathbb{P}(X = 1) + 2z\, \mathbb{P}(X = 2) + 3z^2\, \mathbb{P}(X = 3) + \cdots\;.$$

In particular, $G'(0) = \mathbb{P}(X = 1)$. By differentiating $G'(z)$, we see that the second derivative of G at 0 is $G''(0) = 2\,\mathbb{P}(X = 2)$. Repeating this procedure gives the following corollary to Theorem 2.6.

Corollary 2.1. (Probabilities from PGFs). Let X be a non-negative integer-valued random variable with PGF $G(z)$. Then,

$$\mathbb{P}(X = k) = \frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}z^k}G(0) \;.$$

The PGF thus uniquely determines the discrete pdf. Another consequence of Theorem 2.6 is that expectations, variances, and moments can be easily found from the PGF.

Corollary 2.2. (Moments from PGFs). Let X be a non-negative integer-valued random variable with PGF $G(z)$ and k-th derivative $G^{(k)}(z)$. Then,

$$\lim_{\substack{z\to 1\\ |z|<1}} \frac{\mathrm{d}^k}{\mathrm{d}z^k}G(z) = \mathbb{E}\left[X(X-1)\cdots(X-k+1)\right] \;. \tag{2.10}$$

In particular, if the expectation and variance of X are finite, then $\mathbb{E}X = G'(1)$ and $\mathrm{Var}(X) = G''(1) + G'(1) - (G'(1))^2$.

Proof. The proof is deferred to Appendix B.2. □ ☞ 381

Definition 2.9. (Moment Generating Function). The **moment generating function** (MGF) of a random variable X is the function $M : \mathbb{R} \to [0, \infty]$ given by

$$M(s) = \mathbb{E}\,\mathrm{e}^{sX} \;.$$

In particular, for a discrete random variable with pdf f,

$$M(s) = \sum_x \mathrm{e}^{sx}\, f(x) \;,$$

and for a continuous random variable with pdf f,

$$M(s) = \int_{-\infty}^{\infty} \mathrm{e}^{sx}\, f(x)\,\mathrm{d}x \;.$$

Note that $M(s)$ can be infinite for certain values of s. We sometimes write M_X to stress the role of X.

Similar to the PGF, the MGF has the **uniqueness property**: two MGFs are the same if and only if their corresponding cdfs are the same. In addition, the integer moments of X can be computed from the derivatives of M, as summarized in the next theorem. The proof is similar to that of Theorem 2.6 ☞ 382 and Corollary 2.2 and is given in Appendix B.3.

Theorem 2.7. (Moments from MGFs). If the MGF is finite in an open interval containing 0, then all moments $\mathbb{E}X^n$, $n = 0, 1, \ldots$ are finite and satisfy

$$\mathbb{E}X^n = M^{(n)}(0) \ ,$$

where $M^{(n)}(0)$ is the n-th derivative of M evaluated at 0.

Note that under the conditions of Theorem 2.7, the variance of X can be obtained from the moment generating function as

$$\mathrm{Var}(X) = M''(0) - (M'(0))^2 \ .$$

2.5 Common Discrete Distributions

In this section we give a number of common discrete distributions and list some of their properties. Note that the discrete pdf of each of these distributions, denoted f, depends on one or more parameters; so in fact we are dealing with *families* of distributions.

2.5.1 Bernoulli Distribution

Definition 2.10. (Bernoulli Distribution). A random variable X is said to have a **Bernoulli** distribution with success probability p if X can only assume the values 0 and 1, with probabilities

$$f(0) = \mathbb{P}(X = 0) = 1 - p \qquad \text{and} \qquad f(1) = \mathbb{P}(X = 1) = p \ .$$

We write $X \sim \mathsf{Ber}(p)$.

The Bernoulli distribution is the most fundamental of all probability distributions. It models a single coin toss experiment. Three important properties of the Bernoulli are summarized in the following theorem.

Theorem 2.8. (Properties of the Bernoulli Distribution). Let $X \sim \mathsf{Ber}(p)$. Then,

1. $\mathbb{E}X = p$,
2. $\mathrm{Var}(X) = p(1-p)$,
3. the PGF is $G(z) = 1 - p + zp$.

Proof. The expectation and the variance of X can be obtained by direct computation:

$$\mathbb{E}X = 0 \times \mathbb{P}(X = 0) + 1 \times \mathbb{P}(X = 1) = 0 \times (1-p) + 1 \times p = p$$

and

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{E}X - (\mathbb{E}X)^2 = p - p^2 = p(1-p) \ ,$$

where we have used the fact that in this case $X^2 = X$. Finally, the PGF is given by $G(z) = z^0(1-p) + z^1 p = 1 - p + zp$. □

2.5.2 Binomial Distribution

Definition 2.11. (Binomial Distribution). A random variable X is said to have a **binomial** distribution with parameters n and p if X has pdf

$$f(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n \ . \qquad (2.11)$$

We write $X \sim \mathsf{Bin}(n, p)$.

From Example 2.2 we see that X can be interpreted as the total number of Heads in n successive coin flip experiments, with probability of Heads equal to p. An example of the graph of the pdf is given in Figure 2.6. Theorem 2.9 lists some important properties of the binomial distribution. ☞ 17

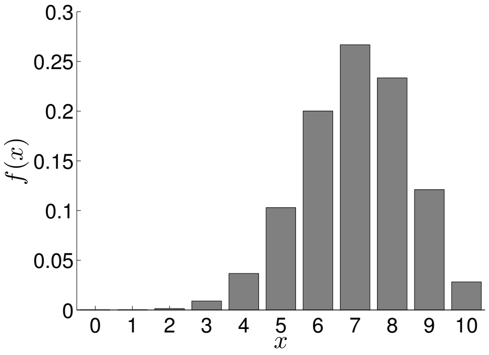

Fig. 2.6 The pdf of the $\mathsf{Bin}(10, 0.7)$-distribution.

Theorem 2.9. (Properties of the Binomial Distribution). Let $X \sim \mathsf{Bin}(n, p)$. Then,

1. $\mathbb{E}X = np$,
2. $\mathrm{Var}(X) = np(1 - p)$,
3. the PGF is $G(z) = (1 - p + zp)^n$.

Proof. Using Newton's binomial formula:

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k} \ ,$$

we see that

$$G(z) = \sum_{k=0}^{n} z^k \binom{n}{k} p^k\, (1 - p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (z\,p)^k (1 - p)^{n-k} = (1 - p + zp)^n \ .$$

☞ 35 From Corollary 2.2 we obtain the expectation and variance via $G'(1) = np$ and $G''(1) + G'(1) - (G'(1))^2 = (n - 1)np^2 + np - n^2p^2 = np(1 - p)$. □

2.5.3 Geometric Distribution

Definition 2.12. (Geometric Distribution). A random variable X is said to have a **geometric** distribution with parameter p if X has pdf

$$f(x) = \mathbb{P}(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, 3, \ldots . \tag{2.12}$$

We write $X \sim \mathsf{Geom}(p)$.

From Example 1.13 we see that X can be interpreted as the number of tosses needed until the first Heads occurs in a sequence of coin tosses, with the probability of Heads equal to p. An example of the graph of the pdf is given in Figure 2.7. Theorem 2.10 summarizes some properties of the geometric distribution. ☞ 18

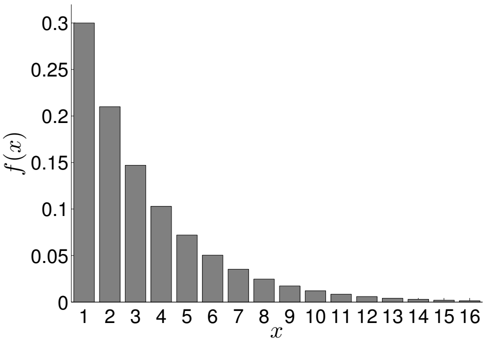

Fig. 2.7 The pdf of the $\mathsf{Geom}(0.3)$-distribution.

Theorem 2.10. (Properties of the Geometric Distribution). Let $X \sim \mathsf{Geom}(p)$. Then,

1. $\mathbb{E}X = 1/p$,
2. $\mathrm{Var}(X) = (1-p)/p^2$,
3. the PGF is

$$G(z) = \frac{z\,p}{1 - z\,(1-p)} \;, \quad |z| < \frac{1}{1-p} \;. \tag{2.13}$$

Proof. The PGF of X follows from

$$G(z) = \sum_{x=1}^{\infty} z^x p(1-p)^{x-1} = z\,p \sum_{k=0}^{\infty} (z(1-p))^k = \frac{z\,p}{1 - z\,(1-p)} \;,$$

using the well-known result for *geometric sums*: $1 + a + a^2 + \cdots = (1-a)^{-1}$, for $|a| < 1$. By Corollary 2.2 the expectation is therefore ☞ 35

$$\mathbb{E}X = G'(1) = \frac{1}{p} \;.$$

By differentiating the PGF twice we find the variance:

$$\mathrm{Var}(X) = G''(1) + G'(1) - (G''(1))^2 = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2} \;. \qquad \square$$

One property of the geometric distribution that deserves extra attention is the **memoryless property**. Consider again the coin toss experiment. Suppose we have tossed the coin k times without a success (Heads). What is the probability that we need more than x additional tosses before getting a success? The answer is, obviously, the same as the probability that we require more than x tosses if we start from scratch, that is, $\mathbb{P}(X > x) = (1-p)^x$, irrespective of k. The fact that we have already had k failures does not make the event of getting a success in the next trial(s) any more likely. In other words, the coin does not have a memory of what happened — hence the name memoryless property.

Theorem 2.11. (Memoryless Property). Let $X \sim \mathsf{Geom}(p)$. Then for any $x, k = 1, 2, \ldots$,

$$\mathbb{P}(X > k + x \,|\, X > k) = \mathbb{P}(X > x) \ .$$

☞ 12 *Proof.* By the definition of conditional probability,

$$\mathbb{P}(X > k + x \,|\, X > k) = \frac{\mathbb{P}(\{X > k + x\} \cap \{X > k\})}{\mathbb{P}(X > k)} \ .$$

The event $\{X > k + x\}$ is a subset of $\{X > k\}$, hence their intersection is $\{X > k + x\}$. Moreover, the probabilities of the events $\{X > k + x\}$ and $\{X > k\}$ are $(1-p)^{k+x}$ and $(1-p)^k$, respectively. Therefore,

$$\mathbb{P}(X > k + x \,|\, X > k) = \frac{(1-p)^{k+x}}{(1-p)^k} = (1-p)^x = \mathbb{P}(X > x) \ ,$$

as required. □

2.5.4 Poisson Distribution

Definition 2.13. (Poisson Distribution). A random variable X is said to have a **Poisson** distribution with parameter $\lambda > 0$ if X has pdf

$$f(x) = \mathbb{P}(X = x) = \frac{\lambda^x}{x!}\,\mathrm{e}^{-\lambda}, \quad x = 0, 1, 2, \ldots . \qquad (2.14)$$

We write $X \sim \mathsf{Poi}(\lambda)$.

The Poisson distribution may be viewed as the limit of the $\mathsf{Bin}(n, \lambda/n)$ distribution. Namely, if $X_n \sim \mathsf{Bin}(n, \lambda/n)$, then

$$\begin{aligned}\mathbb{P}(X_n = x) &= \binom{n}{x}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!}\,\frac{n\times(n-1)\times\cdots\times(n-x+1)}{n\times n\times\cdots\times n}\,\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} .\end{aligned}$$

As $n \to \infty$ the second and fourth factors converge to 1, and the third factor to $\mathrm{e}^{-\lambda}$ (this is one of the defining properties of the exponential function). Hence, we have

$$\lim_{n\to\infty} \mathbb{P}(X_n = x) = \frac{\lambda^x}{x!}\,\mathrm{e}^{-\lambda}.$$

An example of the graph of the Poisson pdf is given in Figure 2.8. Theorem 2.12 summarizes some properties of the Poisson distribution.

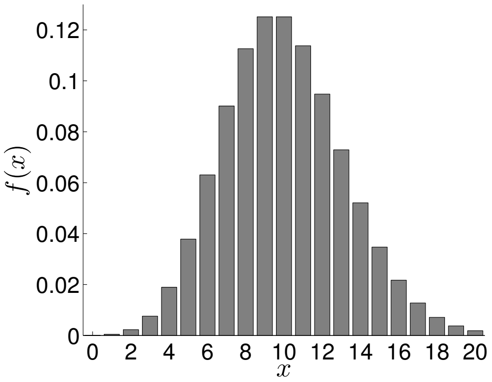

Fig. 2.8 The pdf of the Poi(10)-distribution.

Theorem 2.12. (Properties of the Poisson Distribution). Let $X \sim \mathsf{Poi}(\lambda)$. Then,

1. $\mathbb{E}X = \lambda$,
2. $\mathrm{Var}(X) = \lambda$,
3. the PGF is $G(z) = \mathrm{e}^{-\lambda(1-z)}$.

Proof. The PGF was derived in Example 2.6. It follows from Corollary 2.2 that $\mathbb{E}X = G'(1) = \lambda$ and ☞ 34

$$\mathrm{Var}(X) = G''(1) + G'(1) - (G'(1))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda .$$

Thus, the parameter λ can be interpreted as both the expectation and variance of X. □

2.6 Common Continuous Distributions

In this section we give a number of common continuous distributions and list some of their properties. Note that the pdf of each of these distributions depends on one or more parameters; so, as in the previous section, we are dealing with *families* of distributions.

2.6.1 Uniform Distribution

Definition 2.14. (Uniform Distribution). A random variable X is said to have a **uniform** distribution on the interval $[a, b]$ if its pdf is given by

$$f(x) = \frac{1}{b-a}, \quad a \leqslant x \leqslant b\ .$$

We write $X \sim \mathsf{U}[a, b]$ (and $X \sim \mathsf{U}(a, b)$ for a uniform random variable on an open interval (a, b)).

The random variable $X \sim \mathsf{U}[a, b]$ can model a randomly chosen point from the interval $[a, b]$, where each choice is equally likely. A graph of the pdf is given in Figure 2.9.

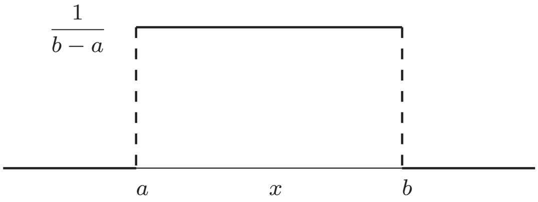

Fig. 2.9 The pdf of the uniform distribution on $[a, b]$.

Theorem 2.13. (Properties of the Uniform Distribution). Let $X \sim \mathsf{U}[a, b]$. Then,

1. $\mathbb{E}X = (a + b)/2$,
2. $\text{Var}(X) = (b - a)^2/12$.

Proof. We have

$$\mathbb{E}X = \int_a^b \frac{x}{b-a}\, \mathrm{d}x = \frac{1}{b-a}\left[\frac{b^2 - a^2}{2}\right] = \frac{a+b}{2}$$

and

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 &= \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{b^3-a^3}{3(b-a)} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}\,.\end{aligned}$$

□

2.6.2 Exponential Distribution

Definition 2.15. (Exponential Distribution). A random variable X is said to have an **exponential** distribution with parameter λ if its pdf is given by

$$f(x) = \lambda\,\mathrm{e}^{-\lambda\,x}, \quad x \geqslant 0\,. \tag{2.15}$$

We write $X \sim \mathsf{Exp}(\lambda)$.

The exponential distribution can be viewed as a continuous version of the geometric distribution. Graphs of the pdf for various values of λ are given in Figure 2.10. Theorem 2.14 summarizes some properties of the exponential distribution.

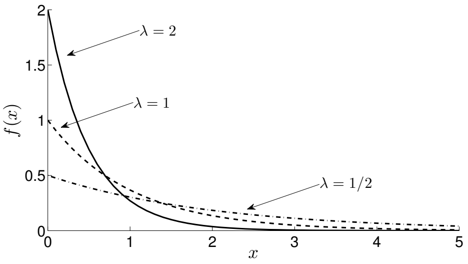

Fig. 2.10 The pdf of the $\mathsf{Exp}(\lambda)$-distribution for various λ.

Theorem 2.14. (Properties of the Exponential Distribution). Let $X \sim \mathsf{Exp}(\lambda)$. Then,

1. $\mathbb{E}X = 1/\lambda$,

2. $\mathrm{Var}(X) = 1/\lambda^2$,
3. The MGF of X is $M(s) = \lambda/(\lambda - s)$, $s < \lambda$,
4. the cdf of X is $F(x) = 1 - \mathrm{e}^{-\lambda x}$, $x \geqslant 0$,
5. the **memoryless property** holds: for any $s, t > 0$,

$$\mathbb{P}(X > s + t \,|\, X > s) = \mathbb{P}(X > t) \; . \tag{2.16}$$

Proof. 3. The moment generating function is given by

$$\begin{aligned} M(s) &= \int_0^\infty \mathrm{e}^{sx} \lambda \mathrm{e}^{-\lambda x} \mathrm{d}x = \lambda \int_0^\infty \mathrm{e}^{-(\lambda - s)x} \, \mathrm{d}x = \lambda \left[\frac{-\mathrm{e}^{-(\lambda - s)x}}{\lambda - s} \right]_0^\infty \\ &= \frac{\lambda}{\lambda - s}, \quad s < \lambda \quad (\text{and } M(s) = \infty \text{ for } s \geqslant \lambda). \end{aligned}$$

☞ 36 1. From Theorem 2.7, we obtain

$$\mathbb{E}X = M'(0) = \left. \frac{\lambda}{(\lambda - s)^2} \right|_{s=0} = \frac{1}{\lambda} \; .$$

2. Similarly, the second moment is $\mathbb{E}X^2 = M''(0) = \frac{2\lambda}{(\lambda - s)^3}\Big|_{s=0} = 2/\lambda^2$, so that the variance is

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \; .$$

4. The cdf of X is given by

$$F(x) = \mathbb{P}(X \leqslant x) = \int_0^x \lambda \mathrm{e}^{-\lambda u} \mathrm{d}u = \left[-\mathrm{e}^{-\lambda u} \right]_0^x = 1 - \mathrm{e}^{-\lambda x}, \quad x \geqslant 0 \; .$$

Note that the tail probability $\mathbb{P}(X > x)$ is exponentially decaying:

$$\mathbb{P}(X > x) = \mathrm{e}^{-\lambda x}, \quad x \geqslant 0 \; .$$

5. Similar to the proof of the memoryless property for the geometric distri-
☞ 40 bution (Theorem 2.11), we have

$$\begin{aligned} \mathbb{P}(X > s + t \,|\, X > s) &= \frac{\mathbb{P}(X > s + t, \, X > s)}{\mathbb{P}(X > s)} = \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)} \\ &= \frac{\mathrm{e}^{-\lambda(t+s)}}{\mathrm{e}^{-\lambda s}} = \mathrm{e}^{-\lambda t} = \mathbb{P}(X > t) \; . \end{aligned}$$

□

The memoryless property can be interpreted as a “non-aging” property. For example, when X denotes the lifetime of a machine then, given the fact that the machine is alive at time s, the remaining lifetime of the machine, $X - s$, has the same exponential distribution as a completely new machine. In other words, the machine has no memory of its age and does not deteriorate (although it will break down eventually).

2.6.3 Normal (Gaussian) Distribution

In this section we introduce the most important distribution in the study of statistics: the normal (or Gaussian) distribution. Additional properties of this distribution will be given in Section 3.6. ☞ 79

Definition 2.16. (Normal Distribution). A random variable X is said to have a **normal** distribution with parameters μ and σ^2 if its pdf is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}\,. \tag{2.17}$$

We write $X \sim \mathsf{N}(\mu, \sigma^2)$.

The parameters μ and σ^2 turn out to be the expectation and variance of the distribution, respectively. If $\mu = 0$ and $\sigma = 1$ then

$$f(x) = \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-x^2/2},$$

and the distribution is known as the **standard normal** distribution. The cdf of the standard normal distribution is often denoted by $\varPhi$ and its pdf by φ. In Figure 2.11 the pdf of the $\mathsf{N}(\mu, \sigma^2)$ distribution for various μ and σ^2 is plotted.

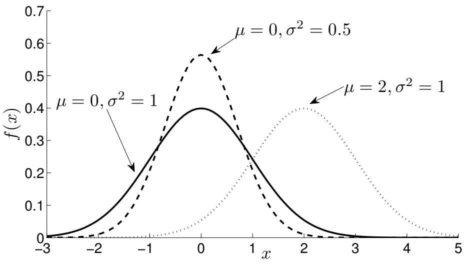

Fig. 2.11 The pdf of the $\mathsf{N}(\mu, \sigma^2)$ distribution for various μ and σ^2.

We next consider some important properties of the normal distribution.

Theorem 2.15. (Standardization). Let $X \sim \mathsf{N}(\mu, \sigma^2)$ and define $Z = (X - \mu)/\sigma$. Then Z has a standard normal distribution.

Proof. The cdf of Z is given by

$$\begin{aligned}\mathbb{P}(Z \leqslant z) &= \mathbb{P}((X-\mu)/\sigma \leqslant z) = \mathbb{P}(X \leqslant \mu + \sigma z) \\ &= \int_{-\infty}^{\mu+\sigma z} \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \,\mathrm{d}x = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-y^2/2} \mathrm{d}y = \Phi(z) \;,\end{aligned}$$

where we make a change of variable $y = (x - \mu)/\sigma$ in the fourth equation. Hence, $Z \sim \mathsf{N}(0, 1)$. □

The rescaling procedure in Theorem 2.15 is called **standardization**. It follows from Theorem 2.15 that any $X \sim \mathsf{N}(\mu, \sigma^2)$ can be written as

$$X = \mu + \sigma Z, \qquad \text{where } \; Z \sim \mathsf{N}(0,1) \;.$$

In other words, any normal random variable can be viewed as an **affine transformation** — that is, a linear transformation plus a constant — of a standard normal random variable.

Next we prove the earlier claim that the parameters μ and σ^2 are respectively the expectation and variance of the distribution.

Theorem 2.16. (Expectation and Variance for the Normal Distribution). If $X \sim \mathsf{N}(\mu, \sigma^2)$, then $\mathbb{E}X = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Proof. Since the pdf is symmetric around μ and $\mathbb{E}X < \infty$, it follows that $\mathbb{E}X = \mu$. To show that the variance of X is σ^2, we first write $X = \mu + \sigma Z$, where $Z \sim \mathsf{N}(0, 1)$. Then, $\mathrm{Var}(X) = \mathrm{Var}(\mu + \sigma Z) = \sigma^2 \mathrm{Var}(Z)$. Hence, it suffices to show that $\mathrm{Var}(Z) = 1$. Now, since the expectation of Z is 0, we have

$$\mathrm{Var}(Z) = \mathbb{E}Z^2 = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-z^2/2} \,\mathrm{d}z = \int_{-\infty}^{\infty} z \times \frac{z}{\sqrt{2\pi}}\, \mathrm{e}^{-z^2/2} \,\mathrm{d}z \;.$$

We apply integration by parts to the last integral to find

$$\mathbb{E}Z^2 = \left[-\frac{z}{\sqrt{2\pi}}\, \mathrm{e}^{-z^2/2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-z^2/2} \,\mathrm{d}z = 1 \;,$$

since the last integrand is the pdf of the standard normal distribution. □

Theorem 2.17. (MGF for the Normal Distribution). The MGF of $X \sim \mathsf{N}(\mu, \sigma^2)$ is

$$\mathbb{E}\mathrm{e}^{sX} = \mathrm{e}^{s\mu + s^2\sigma^2/2}, \quad s \in \mathbb{R} . \tag{2.18}$$

Proof. Write $X = \mu + \sigma Z$, where $Z \sim \mathsf{N}(0,1)$. We have

$$\mathbb{E}\mathrm{e}^{sZ} = \int_{-\infty}^{\infty} \mathrm{e}^{sz} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \, \mathrm{d}z = \mathrm{e}^{s^2/2} \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(z-s)^2/2}}_{\text{pdf of } \mathsf{N}(s,1)} \, \mathrm{d}z = \mathrm{e}^{s^2/2} ,$$

so that $\mathbb{E}\mathrm{e}^{sX} = \mathbb{E}\mathrm{e}^{s(\mu+\sigma Z)} = \mathrm{e}^{s\mu}\, \mathbb{E}\mathrm{e}^{s\sigma Z} = \mathrm{e}^{s\mu}\mathrm{e}^{\sigma^2 s^2/2} = \mathrm{e}^{s\mu + \sigma^2 s^2/2}$. □

2.6.4 Gamma and χ^2 Distribution

Definition 2.17. (Gamma Distribution). A random variable X is said to have a **gamma** distribution with **shape** parameter $\alpha > 0$ and **scale** parameter $\lambda > 0$ if its pdf is given by

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} \mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}, \; x \geqslant 0 , \tag{2.19}$$

where Γ is the gamma function. We write $X \sim \mathsf{Gamma}(\alpha, \lambda)$.

The **gamma function** $\Gamma(\alpha)$ is an important special function in mathematics, defined by

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} \, \mathrm{e}^{-u} \, \mathrm{d}u . \tag{2.20}$$

We mention a few properties of the Γ function.

1. $\Gamma(\alpha + 1) = \alpha\, \Gamma(\alpha)$, for $\alpha \in \mathbb{R}_+$.
2. $\Gamma(n) = (n-1)!$ for $n = 1, 2, \ldots$..
3. $\Gamma(1/2) = \sqrt{\pi}$.

Two special cases of the $\mathsf{Gamma}(\alpha, \lambda)$ distribution are worth mentioning. Firstly, the $\mathsf{Gamma}(1, \lambda)$ distribution is simply the $\mathsf{Exp}(\lambda)$ distribution. Secondly, the $\mathsf{Gamma}(n/2,\ 1/2)$ distribution, where $n \in \{1, 2, \ldots\}$, is called the **chi-squared** distribution with n **degrees of freedom**. We write $X \sim \chi^2_n$. A graph of the pdf of the χ^2_n distribution for various n is given in Figure 2.12.

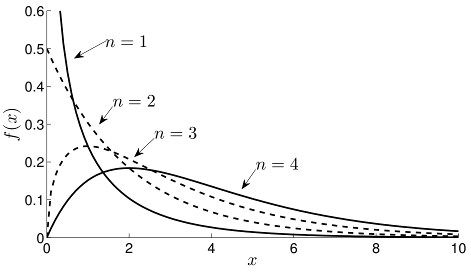

Fig. 2.12 The pdf of the χ^2_n distribution for various degrees of freedom n.

The following theorem summarizes some properties of the gamma distribution.

Theorem 2.18. (Properties of the Gamma Distribution). Let $X \sim \mathsf{Gamma}(\alpha, \lambda)$. Then,

1. $\mathbb{E}X = \alpha/\lambda$,
2. $\mathrm{Var}(X) = \alpha/\lambda^2$,
3. the MGF is $M(s) = [\lambda/(\lambda - s)]^\alpha, s < \lambda$ (and ∞ otherwise).

Proof. 3. For $s < \lambda$, the MGF of X at s is given by

$$\begin{aligned} M(s) = \mathbb{E}\,\mathrm{e}^{sX} &= \int_0^\infty \frac{\mathrm{e}^{-\lambda x}\,\lambda^\alpha\,x^{\alpha-1}}{\Gamma(\alpha)}\,\mathrm{e}^{sx}\,\mathrm{d}x \\ &= \left(\frac{\lambda}{\lambda - s}\right)^\alpha \int_0^\infty \underbrace{\frac{\mathrm{e}^{-(\lambda-s)x}\,(\lambda-s)^\alpha\,x^{\alpha-1}}{\Gamma(\alpha)}}_{\text{pdf of } \mathsf{Gamma}(\alpha,\lambda-s)}\,\mathrm{d}x \\ &= \left(\frac{\lambda}{\lambda - s}\right)^\alpha . \end{aligned} \tag{2.21}$$

☞ 36 1. Consequently, by Theorem 2.7,

$$\mathbb{E}X = M'(0) = \frac{\alpha}{\lambda}\left(\frac{\lambda}{\lambda - s}\right)^{\alpha+1}\Bigg|_{s=0} = \frac{\alpha}{\lambda}.$$

2. Similarly, $\mathrm{Var}(X) = M''(0) - (M'(0))^2 = (\alpha+1)\alpha/\lambda^2 - (\alpha/\lambda)^2 = \alpha/\lambda^2$.

2.6.5 F Distribution

Definition 2.18. (F Distribution). Let m and n be strictly positive integers. A random variable X is said to have an **F** distribution with **degrees of freedom** m and n if its pdf is given by

$$f(x) = \frac{\Gamma(\frac{m+n}{2})\,(m/n)^{m/2}x^{(m-2)/2}}{\Gamma(\frac{m}{2})\,\Gamma(\frac{n}{2})\,[1+(m/n)x]^{(m+n)/2}}, \quad x \geqslant 0\,, \tag{2.22}$$

where Γ denotes the gamma function. We write $X \sim \mathsf{F}(m,n)$.

The F distribution plays an important role in classical statistics, through Theorem 3.11. A graph of the pdf of the $\mathsf{F}(m,n)$ distribution for various m and n is given in Figure 2.13. ☞ 85

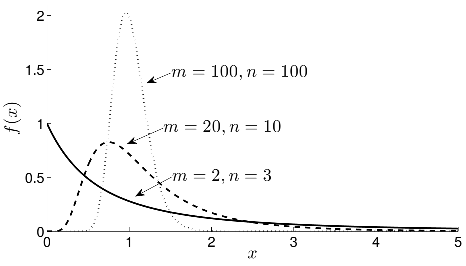

Fig. 2.13 The pdf of the $\mathsf{F}(m,n)$ distribution for various degrees of freedom m and n.

2.6.6 Student's t Distribution

Definition 2.19. (Student's *t* Distribution). A random variable X is said to have a **Student's *t*** distribution with parameter $\nu > 0$ if its pdf is given by

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \quad x \in \mathbb{R}\,, \tag{2.23}$$

where Γ denotes the gamma function. We write $X \sim \mathsf{t}_\nu$. For integer values the parameter ν is referred to as the **degrees of freedom** of the distribution.

A graph of the pdf of the t_ν distribution for various ν is given in Figure 2.14. Note that the pdf is symmetric. Moreover, it can be shown that the pdf of the t_ν distribution converges to the pdf of the $\mathsf{N}(0,1)$ distribution as $\nu \to \infty$. The t_1 distribution is called the **Cauchy distribution**.

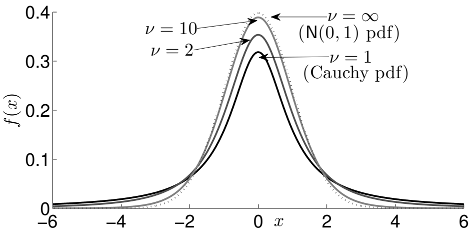

Fig. 2.14 The pdfs of t_1 (Cauchy), t_2, t_{10} and $\mathsf{t}_\infty(\mathsf{N}(0,1))$ distributions.

For completeness we mention that if $X \sim \mathsf{t}_\nu$, then

$$\mathbb{E}X = 0 \quad (\nu > 1) \qquad \text{and} \qquad \text{Var}(X) = \frac{\nu}{\nu-2}, \quad (\nu > 2)\ .$$

The t and F distributions are related in the following way.

Theorem 2.19. (Relationship between the t and F Distribution). For integer $n \geqslant 1$, if $X \sim \mathsf{t}_n$, then $X^2 \sim \mathsf{F}(1,n)$.

Proof. Let $Z = X^2$. We can express the cdf of Z in terms of the cdf of X. Namely, for every $z > 0$ we have

$$F_Z(z) = \mathbb{P}(X^2 \leqslant z) = \mathbb{P}(-\sqrt{z} \leqslant X \leqslant \sqrt{z}) = F_X(\sqrt{z}) - F_X(-\sqrt{z})\ .$$

Differentiating with respect to z gives the following relation between the two pdfs:

$$f_Z(z) = f_X(\sqrt{z})\frac{1}{2\sqrt{z}} + f_X(-\sqrt{z})\frac{1}{2\sqrt{z}} = f_X(\sqrt{z})\frac{1}{\sqrt{z}}\ ,$$

using the symmetry of the t distribution. Substituting (2.23) into the last equation yields

$$f_Z(z) = c(n)\,\frac{z^{-1/2}}{(1+z/n)^{(n+1)/2}}, \quad z > 0$$

for some constant $c(n)$. The only pdf of this form is that of the $\mathsf{F}(1,n)$ distribution. □

2.7 Generating Random Variables

This section shows how to generate random variables on a computer. We first discuss a modern uniform random generator and then introduce two general methods for drawing from an arbitrary one-dimensional distribution: the inverse-transform method and the acceptance–rejection method.

2.7.1 Generating Uniform Random Variables

The MATLAB `rand` function simulates the drawing of a uniform random number on the interval $(0, 1)$ by generating *pseudo*-random numbers; that is, numbers that, although not actually random (because the computer is a deterministic device), behave for all intended purposes as truly random. The following algorithm [L'Ecuyer, 1999] uses simple recurrences to produce high-quality pseudo-random numbers, in the sense that the numbers pass all currently known statistical tests for randomness and uniformity.

Algorithm 2.1. (Combined Multiple-Recursive Generator).

1. Suppose N random numbers are required. Define $m_1 = 2^{32} - 209$ and $m_2 = 2^{32} - 22853$.
2. Initialize a vector $(X_{-2}, X_{-1}, X_0) = (12345, 12345, 12345)$ and a vector $(Y_{-2}, Y_{-1}, Y_0) = (12345, 12345, 12345)$.
3. For $t = 1$ to N let

$$X_t = (1403580\, X_{t-2} - 810728\, X_{t-3}) \bmod m_1 ,$$

$$Y_t = (527612\, Y_{t-1} - 1370589\, Y_{t-3}) \bmod m_2 ,$$

and output the t-th random number as

$$U_t = \begin{cases} \dfrac{X_t - Y_t + m_1}{m_1 + 1} & \text{if } X_t \leqslant Y_t , \\ \dfrac{X_t - Y_t}{m_1 + 1} & \text{if } X_t > Y_t . \end{cases}$$

Here, $x \bmod m$ means the remainder of x when divided by m. The initialization in Step 2 determines the initial state — the so-called **seed** — of the random number stream. Restarting the stream from the same seed produces the same sequence.

Algorithm 2.1 is implemented as a core MATLAB uniform random number generator from Version 7. Currently the default generator in MATLAB is the *Mersenne twister*, which also passes (most) statistical tests, and tends to be a little faster. However, it is considerably more difficult to implement. A typical usage of MATLAB's uniform random number generator is as follows.

```
>>rng(1,'combRecursive') % use the CMRG with seed 1
>>rand(1,5)   % draw 5 random numbers

ans =
    0.4957    0.2243    0.2073    0.6823    0.6799

>>rng(1234)  % set the seed to 1234
>>rand(1,5)

ans =
    0.2830    0.2493    0.3600    0.9499    0.8071

>>rng(1234)  % reset the seed to 1234

>>rand(1,5)

ans =
    0.2830    0.2493    0.3600    0.9499    0.8071
```

2.7.2 Inverse-Transform Method

Once we have a method for drawing a uniform random number, we can, in principle, simulate a random variable X from *any* cdf F by using the following algorithm.

Algorithm 2.2. (Inverse-Transform Method).

```
1. Generate $U$ from $\mathsf{U}(0,1)$.
2. Return $X = F^{-1}(U)$, where $F^{-1}$ is the inverse function of $F$.
```

Figure 2.15 illustrates the inverse-transform method. We see that the random variable $X = F^{-1}(U)$ has cdf F, since

$$\mathbb{P}(X \leqslant x) = \mathbb{P}(F^{-1}(U) \leqslant x) = \mathbb{P}(U \leqslant F(x)) = F(x) \ . \tag{2.24}$$

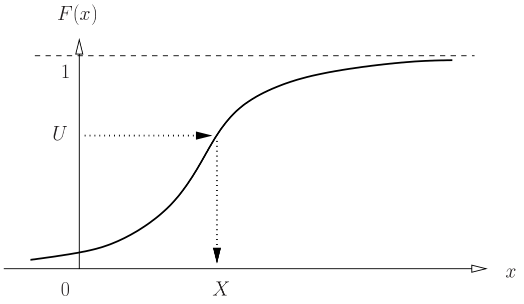

Fig. 2.15 The inverse-transform method.

Example 2.7 (Generating Uniformly on a Unit Disk). Suppose we wish to draw a random point (X, Y) uniformly on the unit disk; see Figure 2.16. In polar coordinates we have $X = R\cos\Theta$ and $Y = R\sin\Theta$, where Θ has a $\mathsf{U}(0, 2\pi)$ distribution. The cdf of R is given by

$$F(r) = \mathbb{P}(R \leqslant r) = \frac{\pi r^2}{\pi} = r^2, \quad 0 < r < 1\ .$$

Its inverse is $F^{-1}(u) = \sqrt{u}$, $0 < u < 1$. We can thus generate R via the inverse transform method as $R = \sqrt{U_1}$, where $U_1 \sim \mathsf{U}(0, 1)$. In addition, we can simulate Θ as $\Theta = 2\pi U_2$, where $U_2 \sim \mathsf{U}(0, 1)$. Note that U_1 and U_2 should be independent draws from $\mathsf{U}(0, 1)$.

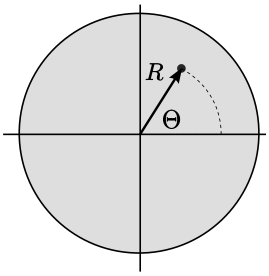

Fig. 2.16 Draw a point (X, Y) uniformly on the unit disk.

The inverse-transform method holds for general cdfs F. Note that F for discrete random variables is a step function, as illustrated in Figure 2.17. The algorithm for generating a random variable X from a discrete distribution that takes values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$ is thus as follows.

Algorithm 2.3. (Discrete Inverse-Transform Method).

```
1. Generate $U \sim \mathsf{U}(0,1)$.
2. Find the smallest positive integer $k$ such that $F(x_k) \geqslant U$ and return
   $X = x_k$.
```

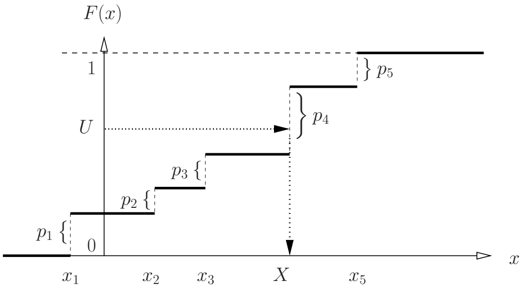

Fig. 2.17 The inverse-transform method for a discrete random variable.

Drawing one of the numbers $1, \ldots, n$ according to a probability vector $(p_1, \ldots, p_n)$ can be done in one line of Matlab code:

```
    min(find(cumsum(p)> rand));
```

Here `p` is the vector of probabilities, such as $(0.3, 0.2, 0.5)$, `cumsum` gives the cumulative vector, e.g., $(0.3, 0.5, 1)$, `find`$(\cdots)$ finds the indices i such that the cumulative probability is greater than some random number `rand`, and `min` takes the smallest of these indices.

2.7.3 Acceptance–Rejection Method

The inverse-transform method may not always be easy to implement, in particular when the inverse cdf is difficult to compute. In that case the **acceptance–rejection** method may prove to be useful. The idea of this method is depicted in Figure 2.18. Suppose we wish to sample from a pdf f. Let g be another pdf such that for some constant $C \geqslant 1$ we have that $Cg(x) \geqslant f(x)$ for all x. It is assumed that it is easy to sample from g; for example, via the inverse-transform method.

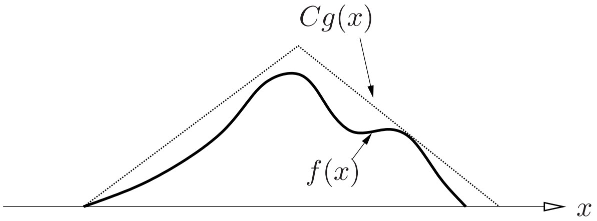

Fig. 2.18 Illustration of the acceptance–rejection method.

It is intuitively clear that if a random point (X, Y) is *uniformly* distributed under the graph of f — that is, on the set $\{(x, y) : 0 \leqslant y \leqslant f(x)\}$ — then X must have pdf f. To construct such a point, let us first draw a random point (Z, V) by drawing Z from g and then drawing V uniformly on $[0, Cg(Z)]$. The point (Z, V) is uniformly distributed under the graph of Cg. If we keep drawing such a point (Z, V) *until it lies under the graph of* f, then the resulting point (X, Y) must be uniformly distributed under the graph of f and hence the X coordinate must have pdf f. This leads to the following algorithm.

Algorithm 2.4. (Acceptance–Rejection Method).

```
1. Generate $Z \sim g$.
2. Generate $Y \sim \mathsf{U}(0, C\,g(Z))$.
3. If $Y \leqslant f(Z)$ return $X = Z$; otherwise, repeat from Step 1.
```

Example 2.8 (Generating from the Standard Normal Distribution). To sample from the standard normal pdf via the inverse-transform method requires knowledge of the inverse of the corresponding cdf, which involves numerical integration. Instead, we can use acceptance–rejection. First, observe that the standard normal pdf is symmetric around 0. Hence, if we can generate a random variable X from the **positive normal** pdf (see Figure 2.19),

$$f(x) = \sqrt{\frac{2}{\pi}}\, \mathrm{e}^{-x^2/2}, \qquad x \geqslant 0\,, \tag{2.25}$$

then we can generate a standard normal random variable by multiplying X with 1 or -1, each with probability 1/2. We can bound $f(x)$ by $C\,g(x)$, where $g(x) = \mathrm{e}^{-x}$ is the pdf of the $\mathsf{Exp}(1)$ distribution. The smallest constant C such that $f(x) \leqslant Cg(x)$ is $\sqrt{2\mathrm{e}/\pi}$.

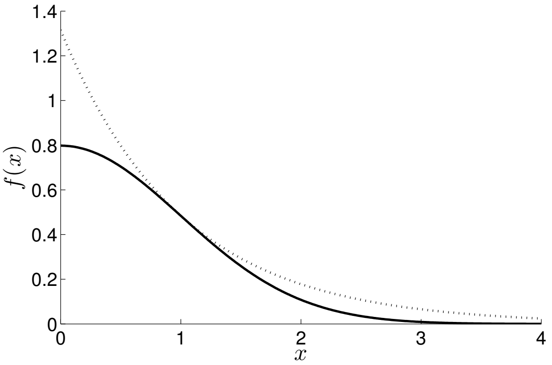

Fig. 2.19 Bounding the positive normal density (solid curve) via an $\mathsf{Exp}(1)$ pdf (times $C \approx 1.3155$).

Drawing from the $\mathsf{Exp}(1)$ distribution can be easily done via the inverse-transform method, noting that the corresponding cdf is the function $1 - \mathrm{e}^{-x}, x \geqslant 0$, whose inverse is the function $-\ln(1-u)$, $u \in (0,1)$. This gives the following specification of Algorithm 2.4, where f and C are defined above.

Algorithm 2.5. ($\mathsf{N}(0,1)$ Generator).

1. Draw $U_1 \sim \mathsf{U}(0,1)$, and let $Z = -\ln U_1$.
2. Draw $U_2 \sim \mathsf{U}(0,1)$, and let $Y = U_2\, C\, \mathrm{e}^{-Z}$.
3. If $Y \leqslant f(Z)$, let $X = Z$ and continue with Step 4. Otherwise, repeat from Step 1.
4. Draw $U_3 \sim \mathsf{U}(0,1)$ and return $\widetilde{X} = X\,(2\,\mathrm{I}_{\{U_3<1/2\}} - 1)$ as a standard normal random variable.

In Step 1, we have used the fact that if $U \sim \mathsf{U}(0,1)$ then also $1 - U \sim \mathsf{U}(0,1)$. In Step 4, $\mathrm{I}_{\{U_3<1/2\}}$ denotes the **indicator** of the event $\{U_3 < 1/2\}$; which is 1 if $U_3 < 1/2$ and 0 otherwise. An alternative generation method ☞ 79 is given in Algorithm 3.2. In MATLAB normal random variable generation is implemented via the `randn` function.

2.8 Problems

2.1. Two fair dice are thrown and the smallest of the face values, M say, is noted.

☞ 27 a. Give the discrete pdf of M in table form, as in Table 2.1.

b. What is the probability that M is at least 3?
c. Calculate the expectation and variance of M.

2.2. A continuous random variable X has cdf

$$F(x) = \begin{cases} 0, & x < 0 \\ x^2/5, & 0 \leqslant x \leqslant 1 \\ \frac{1}{5}\left(-x^2 + 6x - 4\right), & 1 < x \leqslant 3 \\ 1, & x > 3\;. \end{cases}$$

a. Find the corresponding pdf and plot its graph.
b. Calculate the following probabilities.

i. $\mathbb{P}(X \leqslant 2)$
ii. $\mathbb{P}(1 < X \leqslant 2)$
iii. $\mathbb{P}(1 \leqslant X \leqslant 2)$.
iv. $\mathbb{P}(X > 1/2)$.

c. Show that $\mathbb{E}X = 22/15$.

2.3. In this book most random variables are either discrete or continuous; that is, they have either a discrete or continuous pdf. It is also possible to define random variables that have a mix of discrete and continuous characteristics. A simple example is a random variable X with cdf

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - c\,\mathrm{e}^{-x}, & x \geqslant 0 \end{cases}$$

for some fixed $0 < c < 1$.

a. Sketch the cdf F.
b. Find the following probabilities.

i. $\mathbb{P}(0 \leqslant X \leqslant x)$, $x \geqslant 0$.
ii. $\mathbb{P}(0 < X \leqslant x)$, $x \geqslant 0$.
iii. $\mathbb{P}(X = x)$, $x \geqslant 0$.

c. Describe how the inverse-transform method can be used to draw samples from this distribution.

2.4. Let X be a positive random variable with cdf F. Prove that

$$\mathbb{E}X = \int_0^\infty (1 - F(x))\,\mathrm{d}x\;. \qquad (2.26)$$

2.5. Let X be a random variable that can possibly take values $-\infty$ and ∞ with probabilities $\mathbb{P}(X = -\infty) = a$ and $\mathbb{P}(X = \infty) = b$, respectively. Show that the corresponding cdf F satisfies $\lim_{x\to-\infty} F(x) = a$ and $\lim_{x\to\infty} F(x) = 1 - b$.

2.6. Suppose that in a large population the fraction of left-handers is 12%. We select at random 100 people from this population. Let X be the number of left-handers among the selected people. What is the distribution of X? What is the probability that at most 7 of the selected people are left-handed?

2.7. Let $X \sim \mathsf{Geom}(p)$. Show that

$$\mathbb{P}(X > k) = (1-p)^k .$$

2.8. Find the moment generating function (MGF) of $X \sim \mathsf{U}[a,b]$.

2.9. Let $X = a + (b-a)U$, where $U \sim \mathsf{U}[0,1]$. Prove that $X \sim \mathsf{U}[a,b]$. Use this to provide a more elegant proof of Theorem 2.13.

☞ 42

2.10. Show that the exponential distribution is the *only* continuous (positive) distribution that possesses the memoryless property. Hint: show that the memoryless property implies that the tail probability $g(x) = \mathbb{P}(X > x)$ satisfies $g(x+y) = g(x)g(y)$.

2.11. Let $X \sim \mathsf{Exp}(2)$. Calculate the following quantities.

a. $\mathbb{P}(-1 \leqslant X \leqslant 1)$.
b. $\mathbb{P}(X > 4)$.
c. $\mathbb{P}(X > 4 \,|\, X > 2)$.
d. $\mathbb{E}X^2$.

2.12. What is the expectation of a random variable X with the following discrete pdf on the set of integer numbers, excluding 0:

$$f(x) = \frac{3}{\pi^2}\,\frac{1}{x^2}, \quad x \in \mathbb{Z} \setminus \{0\}\ .$$

What is the pdf of the absolute value $|X|$ and what is its expectation?

2.13. A random variable X is said to have a **discrete uniform distribution** on the set $\{a, a+1, \ldots, b\}$ if

$$\mathbb{P}(X = x) = \frac{1}{b-a+1}, \quad x = a, a+1, \ldots, b\ .$$

a. What is the expectation of X?
b. Show that $\mathrm{Var}(X) = (b-a)(b-a+2)/12$.
c. Find the probability generating function (PGF) of X.
d. Describe a simple way to generate X using a uniform number generator.

2.14. Let X and Y be random variables. Prove that if $X \leqslant Y$, then $\mathbb{E}X \leqslant \mathbb{E}Y$.

2.15. A continuous random variable is said to have a **logistic** distribution if its pdf is given by

$$f(x) = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2}, \quad x \in \mathbb{R} \,. \tag{2.27}$$

a. Plot the graph of the pdf.
b. Show that $\mathbb{P}(X > x) = 1/(1+\mathrm{e}^{x})$ for all x.
c. Write an algorithm based on the inverse-transform method to generate random variables from this distribution.

2.16. An electrical component has a lifetime (in years) that is distributed according to an exponential distribution with expectation 3. What is the probability that the component is still functioning after 4.5 years, given that it still works after 4 years? Answer the same question for the case where the component's lifetime is normally distributed with the same expected value and variance as before.

2.17. Consider the pdf given by

$$f(x) = \begin{cases} 4\,\mathrm{e}^{-4(x-1)}, & x \geqslant 1\,, \\ 0, & x < 1\,. \end{cases}$$

a. If X is distributed according to this pdf f, what is its expectation?
b. Specify how one can generate a random variable $X \sim f$ using a uniform random number generator.

2.18. Let $X \sim \mathsf{N}(4, 9)$.

a. Plot the graph of the pdf.
b. Express the following probabilities in terms of the cdf Φ of the standard normal distribution.

   i. $\mathbb{P}(X \leqslant 3)$.
   ii. $\mathbb{P}(X > 4)$.
   iii. $\mathbb{P}(-1 \leqslant X \leqslant 5)$.

c. Find $\mathbb{E}[2X + 1]$.
d. Calculate $\mathbb{E}X^2$.

2.19. Let Φ be the cdf of $X \sim \mathsf{N}(0,1)$. The integral

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}u^2}\, \mathrm{d}u$$

needs to be evaluated numerically. In Matlab there are several ways to do this.

1. If the *Statistics Toolbox* is available, the cdf can be evaluated via the functions `normcdf` or `cdf`. The inverse cdf can be evaluated using `norminv` or `icdf`. See also their replacements `cumdf` and `icumdf` in Appendix A.9. ☞ 377

2. Or one could use the built-in **error function** `erf`, defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-u^2}\, \mathrm{d}u\,, \quad x \in \mathbb{R}\,.$$

   The inverse of the error function, erf^{-1} is implemented in MATLAB as `erfinv`.
3. A third alternative is to use numerical integration (quadrature) via the `quad` function. For example, `quad(@f,0,1)` integrates a MATLAB function `f.m` on the interval $[0, 1]$.

a. Show that $\Phi(x) = (\operatorname{erf}(x/\sqrt{2}) + 1)/2$.
b. Evaluate $\Phi(x)$ for $x = 1, 2$, and 3 via (a) the error function and (b) numerical integration of the pdf, using the fact that $\Phi(0) = 1/2$.
c. Show that the inverse of Φ is given by

$$\Phi^{-1}(y) = \sqrt{2}\,\operatorname{erf}^{-1}(2y-1)\,, \quad 0 < y < 1\,.$$

2.20. Based on MATLAB's `rand` and `randn` functions *only*, implement algorithms that generate random variables from the following distributions.

a. $\mathsf{U}[2, 3]$.
b. $\mathsf{N}(3, 9)$.
c. $\mathsf{Exp}(4)$.
d. $\mathsf{Bin}(10, 1/2)$.
e. $\mathsf{Geom}(1/6)$.

2.21. The **Weibull** distribution $\mathsf{Weib}(\alpha, \lambda)$ has cdf

$$F(x) = 1 - \mathrm{e}^{-(\lambda x)^{\alpha}}, \quad x \geqslant 0\,. \tag{2.28}$$

It can be viewed as a generalization of the exponential distribution. Write a MATLAB program that draws 1000 samples from the $\mathsf{Weib}(2, 1)$ distribution using the inverse-transform method. Give a histogram of the sample.

2.22. Consider the pdf

$$f(x) = c\,\mathrm{e}^{-x} x(1-x), \quad 0 \leqslant x \leqslant 1\,.$$

a. Show that $c = \mathrm{e}/(3 - \mathrm{e})$.
b. Devise an acceptance–rejection algorithm to generate random variables that are distributed according to f.
c. Implement the algorithm in MATLAB.

2.23. Implement two different algorithms to draw 100 uniformly generated points on the unit disk: one based on Example 2.7 and the other using (two-dimensional) acceptance–rejection.

☞ 53

Chapter 3
Joint Distributions

Often a random experiment is described via more than one random variable. Here are some examples.

1. We randomly select $n = 10$ people and observe their heights. Let $X_1, \ldots, X_n$ be the individual heights.
2. We toss a coin repeatedly. Let $X_i = 1$ if the i-th toss is Heads and $X_i = 0$ otherwise. The experiment is thus described by the sequence $X_1, X_2, \ldots$ of Bernoulli random variables.
3. We randomly select a person from a large population and measure his/her weight X and height Y.

How can we specify the behavior of the random variables above? We should not just specify the pdf of the individual random variables, but also say something about the interaction (or lack thereof) between the random variables. For example, in the third experiment above if the height Y is large, then most likely X is large as well. In contrast, in the first two experiments it is reasonable to assume that the random variables are "independent" in some way; that is, information about one of the random variables does not give extra information about the others. What we need to specify is the **joint distribution** of the random variables. The theory below for multiple random variables follows a similar path to that of a single random variable described in Sections 2.1–2.3. ☞ 23

Let $X_1, \ldots, X_n$ be random variables describing some random experiment. We can accumulate the $\{X_i\}$ into a **random vector** $\mathbf{X} = (X_1, \ldots, X_n)$ (row vector) or $\mathbf{X} = (X_1, \ldots, X_n)^\top$ (column vector). Recall that the distribution of a *single* random variable X is completely specified by its cumulative distribution function. For *multiple* random variables we have the following generalization.

Definition 3.1. (Joint Cdf). The **joint cdf** of $X_1, \ldots, X_n$ is the function $F : \mathbb{R}^n \to [0, 1]$ defined by

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n) .$$

Notice that we have used the abbreviation $\mathbb{P}(\{X_1 \leqslant x_1\} \cap \cdots \cap \{X_n \leqslant x_n\}) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n)$ to denote the probability of the intersection of events. We will use this abbreviation throughout the book.

As in the univariate (that is, single-variable) case we distinguish between *discrete* and *continuous* distributions.

3.1 Discrete Joint Distributions

Example 3.1 (Dice Experiment). In a box there are three dice. Die 1 is an ordinary die; die 2 has no 6 face, but instead two 5 faces; die 3 has no 5 face, but instead two 6 faces. The experiment consists of selecting a die at random followed by a toss with that die. Let X be the die number that is selected and let Y be the face value of that die. The probabilities $\mathbb{P}(X = x, Y = y)$ in Table 3.1 specify the joint distribution of X and Y. Note that it is more convenient to specify the joint probabilities $\mathbb{P}(X = x, Y = y)$ than the joint cumulative probabilities $\mathbb{P}(X \leqslant x, Y \leqslant y)$. The latter can be found, however, from the former by applying the sum rule. For example, $\mathbb{P}(X \leqslant 2, Y \leqslant 3) = \mathbb{P}(X = 1, Y = 1) + \cdots + \mathbb{P}(X = 2, Y = 3) = 6/18 = 1/3$. Moreover, by that same sum rule, the distribution of X is found by summing the $\mathbb{P}(X = x, Y = y)$ over all values of y — giving the last column of Table 3.1. Similarly, the distribution of Y is given by the column totals in the last row of the table.

Table 3.1 The joint distribution of X (die number) and Y (face value).

		y						
		1	2	3	4	5	6	$\sum$
	1	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{3}$
x	2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{9}$	0	$\frac{1}{3}$
	3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	$\frac{1}{9}$	$\frac{1}{3}$
	$\sum$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

In general, for discrete random variables $X_1, \ldots, X_n$ the joint distribution is easiest to specify via the joint pdf.

Definition 3.2. (Discrete Joint Pdf). The **joint pdf** f of discrete random variables $X_1, \ldots, X_n$ is given by the function

$$f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)\ .$$

We sometimes write $f_{X_1,\ldots,X_n}$ instead of f to show that this is the pdf of the random variables $X_1, \ldots, X_n$. Or, if $\mathbf{X} = (X_1, \ldots, X_n)$ is the corresponding random vector, we can write $f_{\mathbf{X}}$ instead.

If the joint pdf f is known, we can calculate the probability of any event $\{\mathbf{X} \in B\}$, B in $\mathbb{R}^n$, via the sum rule as

$$\mathbb{P}(\mathbf{X} \in B) = \sum_{\mathbf{x} \in B} f(\mathbf{x}) .$$

Compare this with (2.2). In particular, as explained in Example 3.1, we can find the pdf of X_i — often referred to as a **marginal** pdf, to distinguish it from the joint pdf — by summing the joint pdf over all possible values of the other variables: ☞ 27

$$\mathbb{P}(X_i = x) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, x_n) . \tag{3.1}$$

The converse is not true: from the marginal distributions one cannot in general reconstruct the joint distribution. For example, in Example 3.1 we cannot reconstruct the inside of the two-dimensional table if only given the column and row totals.

However, there is one important exception, namely when the random variables are *independent*. We have so far only defined what independence is for *events*. We can define random variables $X_1, \ldots, X_n$ to be independent if events $\{X_1 \in B_1\}, \ldots, \{X_n \in B_n\}$ are independent for any choice of sets $\{B_i\}$. Intuitively, this means that any information about one of the random variables does not affect our knowledge about the others. ☞ 17

Definition 3.3. (Independence). Random variables $X_1, \ldots, X_n$ are called **independent** if for all events $\{X_i \in B_i\}$ with $B_i \subset \mathbb{R}$, $i = 1, \ldots, n$

$$\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \cdots \mathbb{P}(X_n \in B_n) . \tag{3.2}$$

A direct consequence of the above definition is the following important theorem.

Theorem 3.1. (Independence and Product Rule). Random variables $X_1, \ldots, X_n$ with joint pdf f are independent if and only if

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \tag{3.3}$$

for all $x_1, \ldots, x_n$, where $\{f_{X_i}\}$ are the marginal pdfs.

Proof. The theorem is true in both the discrete and continuous case. We only show the discrete case, where (3.3) is a special case of (3.2). It follows that (3.3) is a *necessary* condition for independence. To see that it is also a *sufficient* condition, let $\mathbf{X} = (X_1, \ldots, X_n)$ and observe that

$$\begin{aligned}\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) &= \mathbb{P}(\mathbf{X} \in \underbrace{B_1 \times \cdots \times B_n}_{A}) = \sum_{\mathbf{x} \in A} f(\mathbf{x}) \\ &= \sum_{\mathbf{x} \in A} f_{X_1}(x_1) \cdots f_{X_n}(x_n) = \sum_{x_1 \in B_1} f_{X_1}(x_1) \cdots \sum_{x_n \in B_n} f_{X_n}(x_n) \\ &= \mathbb{P}(X_1 \in B_1) \cdots \mathbb{P}(X_n \in B_n) \; .\end{aligned}$$

Here $A = B_1 \times \cdots \times B_n$ denotes the Cartesian product of $B_1, \ldots, B_n$. □

Example 3.2 (Dice Experiment Continued). We repeat the experiment in Example 3.1 with three ordinary fair dice. Since the events $\{X = x\}$ and $\{Y = y\}$ are now independent, each entry in the pdf table is $\frac{1}{3} \times \frac{1}{6}$. Clearly in the first experiment not *all* events $\{X = x\}$ and $\{Y = y\}$ are independent.

Remark 3.1. An *infinite* sequence $X_1, X_2, \ldots$ of random variables is said to be *independent* if for any finite choice of positive integers $i_1, i_2, \ldots, i_n$ (none of them the same) the random variables $X_{i_1}, \ldots, X_{i_n}$ are independent. Many statistical models involve random variables $X_1, X_2, \ldots$ that are **independent and identically distributed**, abbreviated as **iid**. We will use this abbreviation throughout this book and write the corresponding model as

$$X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Dist} \text{ (or } f \text{ or } F) \; ,$$

where Dist is the common distribution with pdf f and cdf F.

Example 3.3 (Bernoulli Process). Consider the experiment where we toss a biased coin n times, with probability p of Heads. We can model this experiment in the following way. For $i = 1, \ldots, n$ let X_i be the result of the i-th toss: $\{X_i = 1\}$ means Heads (or success), $\{X_i = 0\}$ means Tails (or failure). Also, let

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0), \quad i = 1, 2, \ldots, n \; .$$

Finally, assume that $X_1, \ldots, X_n$ are *independent.* The sequence

$$X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Ber}(p)$$

is called a **Bernoulli process** with success probability p. Let $X = X_1 + \cdots + X_n$ be the total number of successes in n trials (tosses of the coin). Denote by B_k the set of all binary vectors $\mathbf{x} = (x_1, \ldots, x_n)$ such that $\sum_{i=1}^n x_i = k$. Note that B_k has $\binom{n}{k}$ elements. We have for every $k = 0, \ldots, n$,

$$\begin{aligned}\mathbb{P}(X = k) &= \sum_{\mathbf{x} \in B_k} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \\ &= \sum_{\mathbf{x} \in B_k} \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) = \sum_{\mathbf{x} \in B_k} p^k (1-p)^{n-k} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \;.\end{aligned}$$

In other words, $X \sim \mathsf{Bin}(n, p)$. Compare this with Example 2.2. ☞ 24

For the joint pdf of *dependent* discrete random variables we can write, as a consequence of the product rule (1.5), ☞ 14

$$\begin{aligned}f(x_1, \ldots, x_n) &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \\ &= \mathbb{P}(X_1 = x_1)\, \mathbb{P}(X_2 = x_2 \,|\, X_1 = x_1) \times \cdots \\ &\quad \cdots \times \mathbb{P}(X_n = x_n \,|\, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \;,\end{aligned}$$

assuming that all probabilities $\mathbb{P}(X = x_1), \ldots, \mathbb{P}(X_1 = x_1, \ldots, X_{n-1} = x_{n-1})$ are nonzero. The function which maps, *for a fixed* x_1, each variable x_2 to the conditional probability

$$\mathbb{P}(X_2 = x_2 \,|\, X_1 = x_1) = \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2)}{\mathbb{P}(X_1 = x_1)} \tag{3.4}$$

is called the **conditional pdf** of X_2 given $X_1 = x_1$. We write it as $f_{X_2 \,|\, X_1}(x_2 \,|\, x_1)$. Similarly, the function $x_n \mapsto \mathbb{P}(X_n = x_n \,|\, X_1 = x_1, \ldots, X_{n-1} = x_{n-1})$ is the conditional pdf of X_n given $X_1 = x_1, \ldots, X_{n-1} = x_{n-1}$, which is written as $f_{X_n \,|\, X_1, \ldots, X_{n-1}}(x_n \,|\, x_1, \ldots, x_{n-1})$.

Example 3.4 (Generating Uniformly on a Triangle). We uniformly select a point (X, Y) from the triangle $T = \{(x, y) : x, y \in \{1, \ldots, 6\}, y \leqslant x\}$ in Figure 3.1.

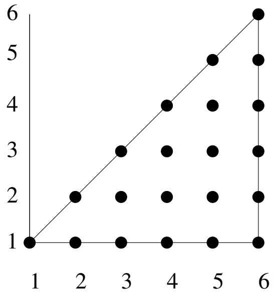

Fig. 3.1 Uniformly select a point from the triangle.

Because each of the 21 points is equally likely to be selected, the joint pdf is constant on T:

$$f(x, y) = \frac{1}{21}, \quad (x, y) \in T \;.$$

The pdf of X is found by summing $f(x,y)$ over all y. Hence,

$$f_X(x) = \frac{x}{21}, \quad x \in \{1,\dots,6\} .$$

Similarly,

$$f_Y(y) = \frac{7-y}{21}, \quad y \in \{1,\dots,6\} .$$

For a fixed $x \in \{1,\dots,6\}$ the conditional pdf of Y given $X = x$ is

$$f_{Y|X}(y \,|\, x) = \frac{f(x,y)}{f_X(x)} = \frac{1/21}{x/21} = \frac{1}{x}, \quad y \in \{1,\dots,x\} ,$$

which simply means that, given $X = x$, Y has a discrete uniform distribution on $\{1,\dots,x\}$.

3.1.1 Multinomial Distribution

An important discrete joint distribution is the multinomial distribution. It can be viewed as a generalization of the binomial distribution. We give the definition and then an example of how this distribution arises in applications.

Definition 3.4. (Multinomial Distribution). A random vector $(X_1, X_2, \dots, X_k)$ is said to have a **multinomial** distribution with parameters n and $p_1, p_2, \dots, p_k$ (positive and summing up to 1), if

$$\mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1!\, x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} , \qquad (3.5)$$

for all $x_1, \dots, x_k \in \{0, 1, \dots, n\}$ such that $x_1 + x_2 + \cdots + x_k = n$. We write $(X_1, \dots, X_k) \sim \mathsf{Mnom}(n, p_1, \dots, p_k)$.

Example 3.5 (Urn Problem). We independently throw n balls into k urns, such that each ball is thrown in urn i with probability p_i, $i = 1, \dots, k$; see Figure 3.2.

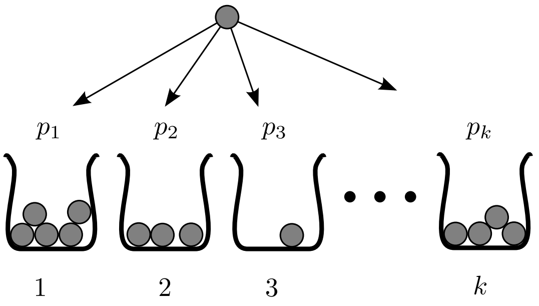

Fig. 3.2 Throwing n balls into k urns with probabilities $p_1, \dots, p_k$. The random configuration of balls has a multinomial distribution.

Let X_i be the total number of balls in urn i, $i = 1, \dots, k$. We show that $(X_1, \dots, X_k) \sim \mathsf{Mnom}(n, p_1, \dots, p_k)$. Let $x_1, \dots, x_k$ be integers between 0 and n that sum up to n. The probability that the *first* x_1 balls fall in the first urn, the *next* x_2 balls fall in the second urn, etc., is

$$p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} .$$

To find the probability that there are x_1 balls in the first urn, x_2 in the second, and so on, we have to multiply the probability above with the number of ways in which we can fill the urns with $x_1, x_2, \dots, x_k$ balls, i.e., $n!/(x_1!\, x_2! \cdots x_k!)$. This gives (3.5).

Remark 3.2. Note that for the *binomial* distribution there are only *two* possible urns. Also, note that for each $i = 1, \dots, k$, $X_i \sim \mathsf{Bin}(n, p_i)$.

3.2 Continuous Joint Distributions

Joint distributions for continuous random variables are usually defined via their joint pdf. The theoretical development below follows very similar lines to both the univariate continuous case in Section 2.2.2 and the multivariate ☞ 28
discrete case in Section 3.1. ☞ 62

Definition 3.5. (Continuous Joint Pdf). Continuous random variables $X_1, \dots, X_n$ are said to have a **joint pdf** f if

$$\mathbb{P}(a_1 < X_1 \leqslant b_1, \dots, a_n < X_n \leqslant b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \dots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

for all $a_1, \dots, b_n$.

This implies, similar to the univariate case in (2.3), that the probability ☞ 28
of any event pertaining to $\mathbf{X} = (X_1, \dots, X_n)$ — say event $\{\mathbf{X} \in B\}$, where B is some subset of $\mathbb{R}^n$ — can be found by *integration*:

$$\mathbb{P}(\mathbf{X} \in B) = \int_B f(x_1, \dots, x_n)\ \mathrm{d}x_1 \dots \mathrm{d}x_n\ . \tag{3.6}$$

As in (2.5) we can interpret $f(x_1, \dots, x_n)$ as the *density* of the probability ☞ 29
distribution at $(x_1, \dots, x_n)$. For example, in the two-dimensional case, for small $h > 0$,

$$\mathbb{P}(x_1 \leqslant X_1 \leqslant x_1 + h,\; x_2 \leqslant X_2 \leqslant x_2 + h)$$
$$= \int_{x_1}^{x_1+h} \int_{x_2}^{x_2+h} f(u,v)\, \mathrm{d}u\, \mathrm{d}v \approx h^2 \, f(x_1, x_2)\ .$$

Similar to the discrete multivariate case in (3.1), the marginal pdfs can be recovered from the joint pdf by integrating out the other variables:

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_{i-1}\, \mathrm{d}x_{i+1}\, \ldots \mathrm{d}x_n\ .$$

We illustrate this for the two-dimensional case. We have

$$F_{X_1}(x) = \mathbb{P}(X_1 \leqslant x, X_2 \leqslant \infty) = \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_2 \right) \mathrm{d}x_1\ .$$

By differentiating the last integral with respect to x, we obtain

$$f_{X_1}(x) = \int_{-\infty}^{\infty} f(x, x_2)\, \mathrm{d}x_2\ .$$

It is not possible, in general, to reconstruct the joint pdf from the marginal pdfs. An exception is when the random variables are *independent*; see Definition 3.3. By modifying the arguments in the proof of Theorem 3.3 to the continuous case — basically replacing sums with integrals — it is not difficult to see that the theorem also holds in the continuous case. In particular, continuous random variables $X_1, \ldots, X_n$ are independent if and only if their joint pdf, f say, is the product of the marginal pdfs:

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \tag{3.7}$$

for all $x_1, \ldots, x_n$. Independence for an infinite sequence of random variables
☞ 64 is discussed in Remark 3.1.

Example 3.6 (Generating a General iid Sample). Consider the sequence of numbers produced by a uniform random number generator such as Matlab's `rand` function. A mathematical model for the output stream is: $U_1, U_2, \ldots,$ are independent and $\mathsf{U}(0,1)$-distributed; that is,

$$U_1, U_2, \ldots \overset{\text{iid}}{\sim} \mathsf{U}(0,1)\ .$$

☞ 52 Using the inverse-transform method it follows that for any cdf F,

$$F^{-1}(U_1), F^{-1}(U_2), \ldots \overset{\text{iid}}{\sim} F\ .$$

Example 3.7 (Quotient of Two Independent Random Variables). Let X and Y be independent continuous random variables, with $Y > 0$. What is the pdf of the quotient $U = X/Y$ in terms of the pdfs of X and Y? Consider first the cdf of U. We have

$$\begin{aligned}F_U(u) &= \mathbb{P}(U \leqslant u) = \mathbb{P}(X/Y \leqslant u) = \mathbb{P}(X \leqslant Yu) \\ &= \int_0^\infty \int_{-\infty}^{yu} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{u} \int_0^\infty y f_X(yz) f_Y(y)\,\mathrm{d}y\,\mathrm{d}z\ ,\end{aligned}$$

where we have used the change of variable $z = x/y$ and changed the order of integration in the last equation. It follows that the pdf is given by

$$f_U(u) = \frac{\mathrm{d}}{\mathrm{d}u} F_U(u) = \int_0^\infty y f_X(yu)\, f_Y(y)\,\mathrm{d}y\ . \tag{3.8}$$

As a particular example, suppose that X and V both have a standard normal distribution. Note that X/V has the same distribution as $U = X/Y$, where $Y = |V| > 0$ has a *positive normal* distribution. It follows from (3.8) that ☞ 55

$$\begin{aligned}f_U(u) &= \int_0^\infty y \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}y^2u^2} \frac{2}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}y^2}\,\mathrm{d}y \\ &= \int_0^\infty y \frac{1}{\pi}\,\mathrm{e}^{-\frac{1}{2}y^2(1+u^2)}\,\mathrm{d}y = \frac{1}{\pi}\frac{1}{1+u^2}, \qquad u \in \mathbb{R}\ .\end{aligned}$$

This is the pdf of the *Cauchy* distribution. ☞ 50

Definition 3.6. (Conditional Pdf). Let X and Y have joint pdf f and suppose $f_X(x) > 0$. The **conditional pdf** of Y given $X = x$ is defined as

$$f_{Y|X}(y\,|\,x) = \frac{f(x,y)}{f_X(x)} \quad \text{for all } y\ . \tag{3.9}$$

For the discrete case, this is just a rewrite of (3.4). For the continuous case, the interpretation is that $f_{Y|X}(y\,|\,x)$ is the density corresponding to the cdf $F_{Y|X}(y\,|\,x)$ defined by the limit

$$F_{Y|X}(y\,|\,x) = \lim_{h\downarrow 0} \mathbb{P}(Y \leqslant y\,|\,x \leqslant X \leqslant x{+}h) = \lim_{h\downarrow 0} \frac{\mathbb{P}(Y \leqslant y,\ x \leqslant X \leqslant x+h)}{\mathbb{P}(x \leqslant X \leqslant x+h)}\ .$$

In many statistical situations, the conditional and marginal pdfs are known and (3.9) is used to find the joint pdf via

$$f(x,y) = f_X(x)\, f_{Y|X}(y\,|\,x)\ ,$$

or, more generally for the n-dimensional case:

$$\begin{aligned}&f(x_1,\ldots,x_n) = \\ &\qquad f_{X_1}(x_1)\, f_{X_2|X_1}(x_2\,|\,x_1) \cdots f_{X_n|X_1,\ldots,X_{n-1}}(x_n\,|\,x_1,\ldots,x_{n-1})\ ,\end{aligned} \tag{3.10}$$

which in the discrete case is just a rephrasing of the *product rule* in terms ☞ 14 of probability densities. For independent random variables (3.10) reduces to (3.7). Equation (3.10) also shows how one could sequentially generate a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ according to a pdf f, provided that it is possible to generate random variables from the successive conditional distributions, as summarized in the following algorithm.

Algorithm 3.1. (Dependent Random Variable Generation).

```
1. Generate $X_1$ from pdf $f_{X_1}$. Set $t = 1$.
2. While $t < n$, given $X_1 = x_1, \ldots, X_t = x_t$, generate $X_{t+1}$ from the
   conditional pdf $f_{X_{t+1} \mid X_1,\ldots,X_t}(x_t \mid x_1, \ldots, x_t)$ and set $t = t + 1$.
3. Return $\mathbf{X} = (X_1, \ldots, X_n)$.
```

Example 3.8 (Non-Uniform Distribution on Triangle). We select a point (X, Y) from the triangle $(0, 0)$-$(1, 0)$-$(1, 1)$ in such a way that X has a uniform distribution on $(0, 1)$ and the conditional distribution of Y given $X = x$ is uniform on $(0, x)$. Figure 3.3 shows the result of 1000 independent draws from the joint pdf $f(x, y) = f_X(x)\, f_{Y \mid X}(y \mid x)$, generated via Algorithm 3.1. It is clear that the points are not uniformly distributed over the triangle.

```
%nutriang.m
N = 1000;
x = rand(N,1);
y = rand(N,1).*x;
plot(x,y,'.')
```

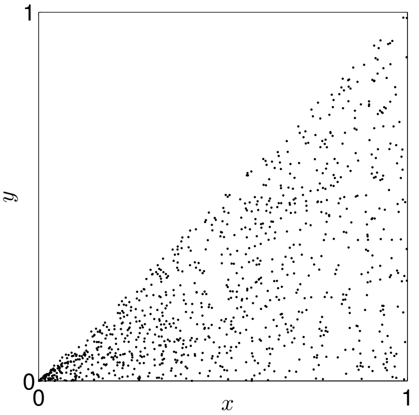

Fig. 3.3 1000 realizations from the joint density $f(x, y)$, generated using the MATLAB program on the left, which implements Algorithm 3.1.

Random variable X has a uniform distribution on $(0, 1)$; hence, its pdf is $f_X(x) = 1$ on $x \in (0, 1)$. For any fixed $x \in (0, 1)$, the conditional distribution of Y given $X = x$ is uniform on the interval $(0, x)$, which means that

$$f_{Y \mid X}(y \mid x) = \frac{1}{x}, \quad 0 < y < x \; .$$

It follows that the joint pdf is given by

$$f(x,y) = f_X(x)\, f_{Y|X}(y\,|\,x) = \frac{1}{x}, \quad 0 < x < 1, \quad 0 < y < x\ .$$

From the joint pdf we can obtain the pdf of Y as

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x = \int_y^1 \frac{1}{x}\,\mathrm{d}x = -\ln y, \quad 0 < y < 1\ .$$

Finally, for any fixed $y \in (0,1)$ the conditional pdf of X given $Y = y$ is

$$f_{X|Y}(x\,|\,y) = \frac{f(x,y)}{f_Y(y)} = \frac{-1}{x \ln y}, \quad y < x < 1\ .$$

3.3 Mixed Joint Distributions

So far we have only considered joint distributions in which the random variables are all discrete or all continuous. The theory can be extended to mixed cases in a straightforward way. For example, the joint pdf of a discrete variable X and a continuous variable Y is defined as the function $f(x,y)$ such that for all events $\{(X,Y) \in A\}$, where $A \subset \mathbb{R}^2$,

$$\mathbb{P}((X,Y) \in A) = \sum_x \int \mathrm{I}_{\{(x,y)\in A\}}\, f(x,y)\,\mathrm{d}y\ ,$$

where I denotes the indicator. The pdf is often specified via (3.10).

Example 3.9 (Beta Distribution). Let $\Theta \sim \mathsf{U}(0,1)$ and $(X\,|\,\Theta = \theta) \sim \mathsf{Bin}(n,\theta)$. Using (3.10), the joint pdf of X and Θ is given by

$$f(x,\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad \theta \in (0,1),\ \ x = 0,1,\ldots,n\ .$$

By integrating out θ, we find the pdf of X:

$$f_X(x) = \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \mathrm{d}\theta = \binom{n}{x} B(x+1, n-x+1)\ ,$$

where B is the **beta function**, defined as

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\ , \tag{3.11}$$

and Γ is the gamma function in (2.20). The conditional pdf of Θ given $X = x$, where $x \in \{0,\ldots,n\}$, is ☞ 47

$$f_{\Theta|X}(\theta\,|\,x) = \frac{f(\theta,x)}{f_X(x)} = \frac{\theta^x(1-\theta)^{n-x}}{B(x+1,n-x+1)}, \quad \theta \in (0,1)\ .$$

The continuous distribution with pdf

$$f(x;\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \quad x \in (0,1) \tag{3.12}$$

is called the **beta distribution** with parameters α and β. Both parameters are assumed to be strictly positive. We write $\mathsf{Beta}(\alpha,\beta)$ for this distribution. For this example we have thus $(\Theta \,|\, X = x) \sim \mathsf{Beta}(x+1, n-x+1)$.

3.4 Expectations for Joint Distributions

☞ 31 Similar to the univariate case in Theorem 2.2, the expected value of a real-valued function h of $(X_1,\ldots,X_n) \sim f$ is a weighted average of all values that $h(X_1,\ldots,X_n)$ can take. Specifically, in the continuous case,

$$\mathbb{E}h(X_1,\ldots,X_n) = \int \cdots \int h(x_1,\ldots,x_n)\, f(x_1,\ldots,x_n)\,\mathrm{d}x_1 \ldots \mathrm{d}x_n\ . \tag{3.13}$$

In the discrete case replace the integrals above with sums.

Two important special cases are the expectation of the *sum* (or more generally affine transformations) of random variables and the *product* of random variables.

Theorem 3.2. (Properties of the Expectation). Let $X_1,\ldots,X_n$ be random variables with expectations $\mu_1,\ldots,\mu_n$. Then,

$$\mathbb{E}[a + b_1X_1 + b_2X_2 + \cdots + b_nX_n] = a + b_1\mu_1 + \cdots + b_n\mu_n \tag{3.14}$$

for all constants a, $b_1,\ldots,b_n$. Also, for *independent* random variables,

$$\mathbb{E}[X_1X_2\cdots X_n] = \mu_1\,\mu_2\cdots\mu_n\ . \tag{3.15}$$

Proof. We show it for the continuous case with two variables only. The general case follows by analogy and, for the discrete case, by replacing integrals with sums. Let X_1 and X_2 be continuous random variables with joint pdf f. Then, by (3.13),

$$\begin{aligned}
\mathbb{E}[a + b_1X_1 + b_2X_2] &= \iint (a + b_1x_1 + b_2x_2)\, f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= a + b_1 \iint x_1 f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2\ + b_2 \iint x_2 f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= a + b_1 \int x_1 \left(\int f(x_1,x_2)\,\mathrm{d}x_2 \right) \mathrm{d}x_1 + b_2 \int x_2 \left(\int f(x_1,x_2)\,\mathrm{d}x_1 \right) \mathrm{d}x_2 \\
&= a + b_1 \int x_1 f_{X_1}(x_1)\,\mathrm{d}x_1 + b_2 \int x_2 f_{X_2}(x_2)\,\mathrm{d}x_2 = a + b_1\mu_1 + b_2\mu_2\ .
\end{aligned}$$

Next, assume that X_1 and X_2 are independent, so that $f(x_1, x_2) = f_{X_1}(x_1) \times f_{X_2}(x_2)$. Then,

$$\begin{aligned}\mathbb{E}[X_1\,X_2] &= \iint x_1\,x_2\,f_{X_1}(x_1) f_{X_2}(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= \int x_1 f_{X_1}(x_1)\,\mathrm{d}x_1 \times \int x_2 f_{X_2}(x_2)\,\mathrm{d}x_2 = \mu_1\,\mu_2\ .\end{aligned}$$

□

Definition 3.7. (Covariance). The **covariance** of two random variables X and Y with expectations $\mathbb{E}X = \mu_X$ and $\mathbb{E}Y = \mu_Y$ is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]\ .$$

The covariance is a measure of the amount of linear dependency between two random variables. A scaled version of the covariance is given by the **correlation coefficient**:

$$\varrho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\,\sigma_Y}\ , \tag{3.16}$$

where $\sigma_X^2 = \mathrm{Var}(X)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$. The correlation coefficient always lies between -1 and 1; see Problem 3.16. ☞ 92

For easy reference Theorem 3.3 lists some important properties of the variance and covariance.

Theorem 3.3. (Properties of the Variance and Covariance). For random variables X, Y and Z, and constants a and b, we have

1. $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
2. $\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$.
3. $\mathrm{Cov}(X, Y) = \mathbb{E}XY - \mathbb{E}X\,\mathbb{E}Y$.
4. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
5. $\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z)$.
6. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
7. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$.
8. If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Proof. For simplicity of notation we write $\mathbb{E}Z = \mu_Z$ for a generic random variable Z. Properties 1 and 2 were already shown in Theorem 2.4.

☞ 33

3. $\text{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[X\,Y - X\,\mu_Y - Y\,\mu_X + \mu_X\,\mu_Y] = \mathbb{E}[X\,Y] - \mu_X\,\mu_Y$.
4. $\text{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[(Y-\mu_Y)(X-\mu_X)] = \text{Cov}(Y,X)$.
5. $\text{Cov}(aX+bY,Z) = \mathbb{E}[(aX+bY)Z] - \mathbb{E}[aX+bY]\,\mathbb{E}Z = a\,\mathbb{E}[XZ] - a\,\mathbb{E}X\mathbb{E}Z + b\,\mathbb{E}[YZ] - b\,\mathbb{E}Y\mathbb{E}Z = a\,\text{Cov}(X,Z) + b\,\text{Cov}(Y,Z)$.
6. $\text{Cov}(X,X) = \mathbb{E}[(X-\mu_X)(X-\mu_X)] = \mathbb{E}[(X-\mu_X)^2] = \text{Var}(X)$.
7. By Property 6, $\text{Var}(X+Y) = \text{Cov}(X+Y,X+Y)$. By Property 5, $\text{Cov}(X+Y,X+Y) = \text{Cov}(X,X)+\text{Cov}(Y,Y)+\text{Cov}(X,Y)+\text{Cov}(Y,X) = \text{Var}(X)+\text{Var}(Y)+2\,\text{Cov}(X,Y)$, where in the last equation Properties 4 and 6 are used.
8. If X and Y are independent, then $\mathbb{E}[X\,Y] = \mu_X\,\mu_Y$. Therefore, $\text{Cov}(X,Y) = 0$ follows immediately from Property 3.

As a consequence of Properties 2 and 7, we have the following general result for the variance of affine transformations of random variables.

Corollary 3.1. (Variance of an Affine Transformation). Let $X_1,\ldots,X_n$ be random variables with variances $\sigma_1^2,\ldots,\sigma_n^2$. Then,

$$\text{Var}\left(a + \sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2\,\sigma_i^2 + 2\sum_{i<j} b_i b_j \text{Cov}(X_i,X_j) \qquad (3.17)$$

for any choice of constants a and $b_1,\ldots,b_n$. In particular, for *independent* random variables $X_1,\ldots,X_n$,

$$\text{Var}(a + b_1X_1 + \cdots + b_nX_n) = b_1^2\sigma_1^2 + \cdots + b_n^2\sigma_n^2\,. \qquad (3.18)$$

Let $\mathbf{X} = (X_1,\ldots,X_n)^\top$ be a random column vector. Sometimes it is convenient to write the expectations and covariances in vector notation.

Definition 3.8. (Expectation Vector and Covariance Matrix). For any random column vector $\mathbf{X}$ we define the **expectation vector** as the vector of expectations

$$\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)^\top = (\mathbb{E}X_1,\ldots,\mathbb{E}X_n)^\top\,.$$

The **covariance matrix** Σ is defined as the matrix whose (i,j)-th element is

$$\text{Cov}(X_i,X_j) = \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)]\,.$$

If we define the expectation of a matrix to be the matrix of expectations, then we can write the covariance matrix succinctly as

$$\Sigma = \mathbb{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top\right] \ .$$

Definition 3.9. (Conditional Expectation). The **conditional expectation** of Y given $X = x$, denoted $\mathbb{E}[Y \,|\, X = x]$, is the expectation corresponding to the conditional pdf $f_{Y|X}(y \,|\, x)$. That is, in the continuous case,

$$\mathbb{E}[Y \,|\, X = x] = \int y \, f_{Y|X}(y \,|\, x) \, \mathrm{d}y \ .$$

In the discrete case replace the integral with a sum.

Note that $\mathbb{E}[Y \,|\, X = x]$ is a function of x, say $h(x)$. The corresponding random variable $h(X)$ is written as $\mathbb{E}[Y \,|\, X]$. The expectation of $\mathbb{E}[Y \,|\, X]$ is, in the continuous case,

$$\begin{aligned}\mathbb{E}\mathbb{E}[Y \,|\, X] &= \int \mathbb{E}[Y \,|\, X = x] f_X(x) \, \mathrm{d}x = \int\int y \frac{f(x,y)}{f_X(x)} f_X(x) \, \mathrm{d}y \, \mathrm{d}x \\ &= \int y \, f_Y(y) \, \mathrm{d}y = \mathbb{E}Y \ .\end{aligned} \tag{3.19}$$

This “stacking” of (conditional) expectations is sometimes referred to as the **tower property**.

Example 3.10 (Non-Uniform Distribution on Triangle Continued). In Example 3.8 the conditional expectation of Y given $X = x$, with $0 < x < 1$, is

$$\mathbb{E}[Y \,|\, X = x] = \frac{1}{2} x \ ,$$

because conditioned on $X = x$, Y is uniformly distributed on the interval $(0, x)$. Using the tower property we find

$$\mathbb{E}Y = \frac{1}{2}\mathbb{E}X = \frac{1}{4} \ .$$

3.5 Functions of Random Variables

Suppose $X_1, \ldots, X_n$ are measurements of a random experiment. What can be said about the distribution of a *function* of the data, say $Z = g(X_1, \ldots, X_n)$, when the joint distribution of $X_1, \ldots, X_n$ is known?

Example 3.11 (Pdf of an Affine Transformation). Let X be a continuous random variable with pdf f_X and let $Z = a + bX$, where $b \neq 0$. We wish to determine the pdf f_Z of Z. Suppose that $b > 0$. We have for any z

$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}\big(X \leqslant (z-a)/b\big) = F_X\big((z-a)/b\big) .$$

Differentiating this with respect to z gives $f_Z(z) = f_X\big((z-a)/b\big)/b$. For $b < 0$ we similarly obtain $f_Z(z) = f_X\big((z-a)/b\big)/(-b)$. Thus, in general,

$$f_Z(z) = \frac{1}{|b|}\, f_X\left(\frac{z-a}{b}\right) . \tag{3.20}$$

Example 3.12 (Pdf of a Monotone Transformation). Generalizing the previous example, suppose that $Z = g(X)$ for some strictly increasing function g. To find the pdf of Z from that of X we first write

$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}\left(X \leqslant g^{-1}(z)\right) = F_X\left(g^{-1}(z)\right) ,$$

where g^{-1} is the inverse of g. Differentiating with respect to z now gives

$$f_Z(z) = f_X(g^{-1}(z))\,\frac{\mathrm{d}}{\mathrm{d}z} g^{-1}(z) = \frac{f_X(g^{-1}(z))}{g'(g^{-1}(z))} . \tag{3.21}$$

For strictly decreasing functions, g' needs to be replaced with its negative value.

3.5.1 Linear Transformations

Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ be a column vector in $\mathbb{R}^n$ and B an $m \times n$ matrix. The mapping $\mathbf{x} \mapsto \mathbf{z}$, with $\mathbf{z} = B\mathbf{x}$, is called a **linear transformation**. Now consider a *random* vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$, and let

$$\mathbf{Z} = B\mathbf{X} .$$

Then $\mathbf{Z}$ is a random vector in $\mathbb{R}^m$. In principle, if we know the joint distribution of $\mathbf{X}$, then we can derive the joint distribution of $\mathbf{Z}$. Let us first see how the expectation vector and covariance matrix are transformed.

Theorem 3.4. (Expectation and Covariance Under a Linear Transformation). If $\mathbf{X}$ has expectation vector $\boldsymbol{\mu}_\mathbf{X}$ and covariance matrix $\varSigma_\mathbf{X}$, then the expectation vector and covariance matrix of $\mathbf{Z} = B\mathbf{X}$ are given by

$$\boldsymbol{\mu}_\mathbf{Z} = B\boldsymbol{\mu}_\mathbf{X} \tag{3.22}$$

and

$$\varSigma_\mathbf{Z} = B\ \varSigma_\mathbf{X}\ B^\top . \tag{3.23}$$

Proof. We have $\boldsymbol{\mu}_{\mathbf{Z}} = \mathbb{E}\mathbf{Z} = \mathbb{E}B\mathbf{X} = B\,\mathbb{E}\mathbf{X} = B\boldsymbol{\mu}_{\mathbf{X}}$ and

$$\begin{aligned}\Sigma_{\mathbf{Z}} &= \mathbb{E}[(\mathbf{Z}-\boldsymbol{\mu}_{\mathbf{Z}})(\mathbf{Z}-\boldsymbol{\mu}_{\mathbf{Z}})^\top] = \mathbb{E}[B(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(B(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}}))^\top] \\ &= B\,\mathbb{E}[(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})^\top]B^\top \\ &= B\;\Sigma_{\mathbf{X}}\;B^\top\,.\end{aligned}$$

□

Suppose that B is an *invertible* $n \times n$ matrix. If $\mathbf{X}$ has a joint pdf $f_{\mathbf{X}}$, what is the joint density $f_{\mathbf{Z}}$ of $\mathbf{Z}$? Let us consider the continuous case. For any fixed $\mathbf{x}$, let $\mathbf{z} = B\mathbf{x}$. Hence, $\mathbf{x} = B^{-1}\mathbf{z}$. Consider the n-dimensional cube $C = [z_1, z_1 + h] \times \cdots \times [z_n, z_n + h]$. Then, by definition of the joint density for $\mathbf{Z}$, we have

$$\mathbb{P}(\mathbf{Z} \in C) \approx h^n\, f_{\mathbf{Z}}(\mathbf{z})\,.$$

Let D be the image of C under B^{-1} — that is, the parallelepiped of all points $\mathbf{x}$ such that $B\mathbf{x} \in C$; see Figure 3.4.

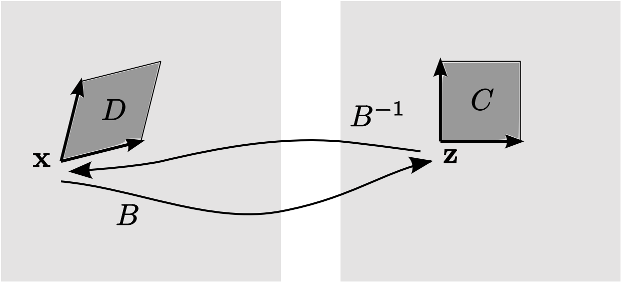

Fig. 3.4 Linear transformation.

A basic result from linear algebra is that any matrix B linearly transforms an n-dimensional rectangle with volume V into an n-dimensional parallelepiped with volume $V\,|B|$, where $|B| = |\det(B)|$. Thus, in addition to the above expression for $\mathbb{P}(\mathbf{Z} \in C)$ we also have

$$\mathbb{P}(\mathbf{Z} \in C) = \mathbb{P}(\mathbf{X} \in D) \approx h^n |B^{-1}|\, f_{\mathbf{X}}(\mathbf{x}) = h^n |B|^{-1}\, f_{\mathbf{X}}(\mathbf{x})\,.$$

Equating these two expressions for $\mathbb{P}(\mathbf{Z} \in C)$ and letting h go to 0, we obtain

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{f_{\mathbf{X}}(B^{-1}\mathbf{z})}{|B|}, \quad \mathbf{z} \in \mathbb{R}^n. \tag{3.24}$$

3.5.2 General Transformations

We can apply similar reasoning as in the previous subsection to deal with general transformations $\mathbf{x} \mapsto \mathbf{g}(\mathbf{x})$, written out as

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_n(\mathbf{x}) \end{pmatrix} .$$

For a fixed $\mathbf{x}$, let $\mathbf{z} = \mathbf{g}(\mathbf{x})$. Suppose $\mathbf{g}$ is invertible; hence, $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{z})$. Any infinitesimal n-dimensional rectangle at $\mathbf{x}$ with volume V is transformed into an n-dimensional parallelepiped at $\mathbf{z}$ with volume $V\,|J_{\mathbf{g}}(\mathbf{x})|$, where $J_{\mathbf{g}}(\mathbf{x})$ is the *matrix of Jacobi* at $\mathbf{x}$ of the transformation $\mathbf{g}$; that is, ☞ 379

$$J_{\mathbf{g}}(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix} .$$

Now consider a random column vector $\mathbf{Z} = \mathbf{g}(\mathbf{X})$. Let C be a small cube around $\mathbf{z}$ with volume h^n. Let D be the image of C under $\mathbf{g}^{-1}$. Then, as in the linear case,

$$h^n\, f_{\mathbf{Z}}(\mathbf{z}) \approx \mathbb{P}(\mathbf{Z} \in C) \approx h^n |J_{\mathbf{g}^{-1}}(\mathbf{z})|\, f_{\mathbf{X}}(\mathbf{x}) .$$

Hence, we have the following result.

Theorem 3.5. (Transformation Rule). Let $\mathbf{X}$ be a continuous n-dimensional random vector with pdf $f_{\mathbf{X}}$ and $\mathbf{g}$ a function from $\mathbb{R}^n$ to $\mathbb{R}^n$ with inverse $\mathbf{g}^{-1}$. Then, $\mathbf{Z} = \mathbf{g}(\mathbf{X})$ has pdf

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{z}))\, |J_{\mathbf{g}^{-1}}(\mathbf{z})|, \quad \mathbf{z} \in \mathbb{R}^n. \tag{3.25}$$

Remark 3.3. Note that $|J_{\mathbf{g}^{-1}}(\mathbf{z})| = 1/|J_{\mathbf{g}}(\mathbf{x})|$.

Example 3.13 (Box-Muller Method). The joint distribution of $X, Y \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$ is

$$f_{X,Y}(x,y) = \frac{1}{2\pi} \mathrm{e}^{-\frac{1}{2}(x^2+y^2)}, \quad (x,y) \in \mathbb{R}^2 .$$

In polar coordinates we have

$$X = R\cos\Theta \qquad \text{and} \qquad Y = R\sin\Theta\ , \tag{3.26}$$

where $R \geqslant 0$ and $\Theta \in (0, 2\pi)$. What is the joint pdf of R and Θ? Consider the inverse transformation $\mathbf{g}^{-1}$, defined by

$$\begin{pmatrix} r \\ \theta \end{pmatrix} \xmapsto{\mathbf{g}^{-1}} \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} .$$

The corresponding matrix of Jacobi is

$$J_{\mathbf{g}^{-1}}(r,\theta) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} ,$$

which has determinant r. Since $x^2 + y^2 = r^2(\cos^2\theta + \sin^2\theta) = r^2$, it follows that

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(x,y)\, r = \frac{1}{2\pi} \mathrm{e}^{-\frac{1}{2}r^2}\, r, \quad \theta \in (0, 2\pi), \quad r \geqslant 0 .$$

By integrating out θ and r, respectively, we find $f_R(r) = r\,\mathrm{e}^{-r^2/2}$ and $f_\Theta(\theta) = 1/(2\pi)$. Since $f_{R,\Theta}$ is the product of f_R and f_Θ, the random variables R and Θ are independent. This shows how X and Y could be generated: independently generate $R \sim f_R$ and $\Theta \sim \mathsf{U}(0, 2\pi)$ and return X and Y via (3.26). Generation from f_R can be done via the inverse-transform method. ☞ 52 In particular, R has the same distribution as $\sqrt{-2\ln U}$ with $U \sim \mathsf{U}(0,1)$. This leads to the following method for generating standard normal random variables.

Algorithm 3.2. (Box–Muller Method).

1. Generate $U_1, U_2 \overset{\text{iid}}{\sim} \mathsf{U}(0,1)$.
2. Return two independent standard normal variables, X and Y, via

$$\begin{aligned} X &= \sqrt{-2\ln U_1}\, \cos(2\pi U_2) , \\ Y &= \sqrt{-2\ln U_1}\, \sin(2\pi U_2) . \end{aligned} \tag{3.27}$$

3.6 Multivariate Normal Distribution

It is helpful to view a normally distributed random variable as an affine transformation of a standard normal random variable. In particular, if Z has a standard normal distribution, then $X = \mu + \sigma Z$ has a $\mathsf{N}(\mu, \sigma^2)$ distribution; see Theorem 2.15. ☞ 46

We now generalize this to n dimensions. Let $Z_1, \ldots, Z_n$ be independent and standard normal random variables. The joint pdf of $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$

is given by

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}z_i^2} = (2\pi)^{-\frac{n}{2}}\, \mathrm{e}^{-\frac{1}{2}\mathbf{z}^\top \mathbf{z}}, \quad \mathbf{z} \in \mathbb{R}^n. \tag{3.28}$$

We write $\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, I)$, where I is the identity matrix. Consider the affine transformation (that is, a linear transformation plus a constant vector)

$$\mathbf{X} = \boldsymbol{\mu} + B\,\mathbf{Z} \tag{3.29}$$

for some $m \times n$ matrix B and m-dimensional vector $\boldsymbol{\mu}$. Note that, by Theorem 3.4, $\mathbf{X}$ has expectation vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma = BB^\top$.

Definition 3.10. (Multivariate Normal Distribution). A random vector $\mathbf{X}$ is said to have a **multivariate normal** or **multivariate Gaussian** distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ if it can be written as $\mathbf{X} = \boldsymbol{\mu} + B\,\mathbf{Z}$, where $\mathbf{Z} \sim \mathsf{N}(\mathbf{0}, I)$ and $BB^\top = \Sigma$. We write $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$.

Suppose that B is an invertible $n \times n$ matrix. Then, by (3.24), the density of $\mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|B|\sqrt{(2\pi)^n}}\, \mathrm{e}^{-\frac{1}{2}(B^{-1}\mathbf{y})^\top B^{-1}\mathbf{y}} = \frac{1}{|B|\sqrt{(2\pi)^n}}\, \mathrm{e}^{-\frac{1}{2}\mathbf{y}^\top (B^{-1})^\top B^{-1}\mathbf{y}}\,.$$

We have $|B| = \sqrt{|\Sigma|}$ and $(B^{-1})^\top B^{-1} = (B^\top)^{-1}B^{-1} = (BB^\top)^{-1} = \Sigma^{-1}$, so that

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\, \mathrm{e}^{-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1}\mathbf{y}}\,.$$

Because $\mathbf{X}$ is obtained from $\mathbf{Y}$ by simply adding a constant vector $\boldsymbol{\mu}$, we have $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{x} - \boldsymbol{\mu})$ and therefore

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\, \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \mathbb{R}^n\,. \tag{3.30}$$

Figure 3.5 shows the pdfs of two bivariate (that is two-dimensional) normal distributions. In both cases the mean vector is $\boldsymbol{\mu} = (0,0)^\top$ and the variances (the diagonal elements of Σ) are 1. The correlation coefficients (or, equivalently here, the covariances) are respectively $\varrho = 0$ and $\varrho = 0.8$.

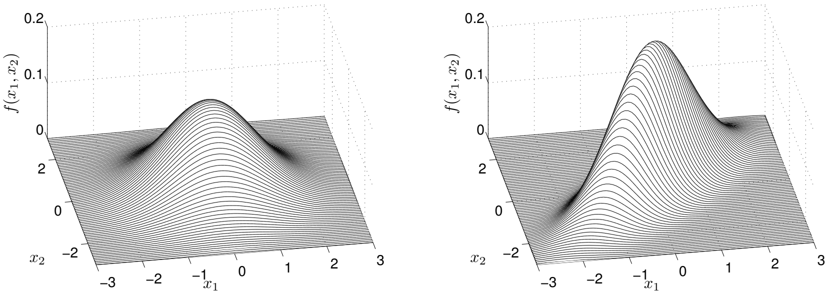

Fig. 3.5 Pdfs of bivariate normal distributions with means zero, variances 1, and correlation coefficients 0 (left) and 0.8 (right).

Conversely, given a covariance matrix $\Sigma = (\sigma_{ij})$, there exists a unique lower triangular matrix B such that $\Sigma = BB^\top$. In MATLAB, the function `chol` accomplishes this so-called **Cholesky factorization**. Note that it is important to use the option `'lower'` when calling this function, as MATLAB produces an upper triangular matrix by default. Once the Cholesky factorization is determined, it is easy to sample from a multivariate normal distribution.

Algorithm 3.3. (Normal Random Vector Generation). To generate N independent draws from a $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$ distribution of dimension n carry out the following steps.

1. Determine the lower Cholesky factorization $\Sigma = BB^\top$.
2. Generate $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ by drawing $Z_1, \ldots, Z_n \sim_{\text{iid}} \mathsf{N}(0, 1)$.
3. Output $\mathbf{X} = \boldsymbol{\mu} + B\mathbf{Z}$.
4. Repeat Steps 2 and 3 independently N times.

Example 3.14 (Generating from a Bivariate Normal Distribution). The MATLAB code below draws 1000 samples from the two pdfs in Figure 3.5. The resulting point clouds are given in Figure 3.6.

```
%bivnorm.m
N = 1000; rho = 0.8;
Sigma = [1 rho; rho 1];
B=chol(Sigma,'lower');
x=B*randn(2,N);
plot(x(1,:),x(2,:),'.')
```

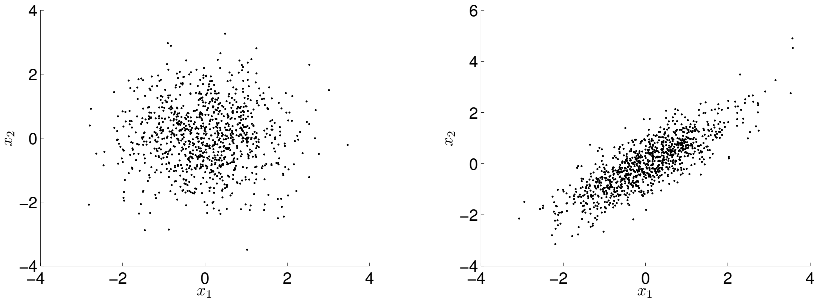

Fig. 3.6 1000 realizations of bivariate normal distributions with means zero, variances 1, and correlation coefficients 0 (left) and 0.8 (right).

The following theorem states that any affine combination of independent multivariate normal random variables is again multivariate normal.

Theorem 3.6. (Affine Transformation of Normal Random Vectors). Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_r$ be independent m_i-dimensional normal random vectors, with $\mathbf{X}_i \sim \mathsf{N}(\boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, \ldots, r$. Then, for any $n \times 1$ vector $\mathbf{a}$ and $n \times m_i$ matrices $B_1, \ldots, B_r$,

$$\mathbf{a} + \sum_{i=1}^{r} B_i\, \mathbf{X}_i \sim \mathsf{N}\left(\mathbf{a} + \sum_{i=1}^{r} B_i\, \boldsymbol{\mu}_i,\ \sum_{i=1}^{r} B_i\, \Sigma_i\, B_i^\top\right). \tag{3.31}$$

Proof. Denote the n-dimensional random vector in the left-hand side of (3.31) by $\mathbf{Y}$. By Definition 3.10, each $\mathbf{X}_i$ can be written as $\boldsymbol{\mu}_i + A_i\mathbf{Z}_i$, where the $\{\mathbf{Z}_i\}$ are independent (because the $\{\mathbf{X}_i\}$ are independent), so that

$$\mathbf{Y} = \mathbf{a} + \sum_{i=1}^{r} B_i\, (\boldsymbol{\mu}_i + A_i\mathbf{Z}_i) = \mathbf{a} + \sum_{i=1}^{r} B_i\, \boldsymbol{\mu}_i + \sum_{i=1}^{r} B_i A_i \mathbf{Z}_i\ ,$$

which is an affine combination of independent standard normal random vectors. Hence, $\mathbf{Y}$ is multivariate normal. Its expectation vector and covariance matrix can be found easily from Theorem 3.4. □

The next theorem shows that the distribution of a subvector of a multivariate normal random vector is again normal.

Theorem 3.7. (Marginal Distributions of Normal Random Vectors). Let $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ be an n-dimensional normal random vector. Decompose $\mathbf{X}$, $\boldsymbol{\mu}$, and Σ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_p \\ \mathbf{X}_q \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_p & \Sigma_r \\ \Sigma_r^\top & \Sigma_q \end{pmatrix}, \tag{3.32}$$

where Σ_p is the upper left $p \times p$ corner of Σ and Σ_q is the lower right $q \times q$ corner of Σ. Then, $\mathbf{X}_p \sim \mathsf{N}(\boldsymbol{\mu}_p, \Sigma_p)$.

Proof. Let $BB^\top$ be the lower Cholesky factorization of Σ. We can write

$$\begin{pmatrix} \mathbf{X}_p \\ \mathbf{X}_q \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q \end{pmatrix} + \underbrace{\begin{pmatrix} B_p & O \\ C_r & C_q \end{pmatrix}}_{B} \begin{pmatrix} \mathbf{Z}_p \\ \mathbf{Z}_q \end{pmatrix}, \tag{3.33}$$

where $\mathbf{Z}_p$ and $\mathbf{Z}_q$ are independent p- and q-dimensional standard normal random vectors. In particular, $\mathbf{X}_p = \boldsymbol{\mu}_p + B_p\mathbf{Z}_p$, which means that $\mathbf{X}_p \sim \mathsf{N}(\boldsymbol{\mu}_p, \Sigma_p)$, since $B_pB_p^\top = \Sigma_p$. □

By relabeling the elements of $\mathbf{X}$ we see that Theorem 3.7 implies that *any* subvector of $\mathbf{X}$ has a multivariate normal distribution. For example, $\mathbf{X}_q \sim \mathsf{N}(\boldsymbol{\mu}_q, \Sigma_q)$.

Not only the marginal distributions of a normal random vector are normal but also its *conditional distributions*.

Theorem 3.8. (Conditional Distributions of Normal Random Vectors). Let $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ be an n-dimensional normal random vector with $\det(\Sigma) > 0$. If $\mathbf{X}$ is decomposed as in (3.32), then

$$(\mathbf{X}_q \,|\, \mathbf{X}_p = \mathbf{x}_p) \sim \mathsf{N}(\boldsymbol{\mu}_q + \Sigma_r^\top \Sigma_p^{-1}(\mathbf{x}_p - \boldsymbol{\mu}_p),\, \Sigma_q - \Sigma_r^\top \Sigma_p^{-1} \Sigma_r). \tag{3.34}$$

As a consequence, $\mathbf{X}_p$ and $\mathbf{X}_q$ are *independent* if and only if they are *uncorrelated*; that is, if $\Sigma_r = O$ (zero matrix).

Proof. From (3.33) we see that

$$(\mathbf{X}_q \,|\, \mathbf{X}_p = \mathbf{x}_p) = \boldsymbol{\mu}_q + C_r\, B_p^{-1}(\mathbf{x}_p - \boldsymbol{\mu}_p) + C_q\mathbf{Z}_q\,,$$

where $\mathbf{Z}_q$ is a q-dimensional multivariate standard normal random vector. It follows that $\mathbf{X}_q$ conditional on $\mathbf{X}_p = \mathbf{x}_p$ has a $\mathsf{N}(\boldsymbol{\mu}_q + C_r\, B_p^{-1}(\mathbf{x}_p - \boldsymbol{\mu}_p), C_qC_q^\top)$

distribution. The proof of (3.34) is completed by observing that $\Sigma_r^\top \Sigma_p^{-1} = C_r B_p^\top (B_p^\top)^{-1} B_p^{-1} = C_r \, B_p^{-1}$, and

$$\Sigma_q - \Sigma_r^\top \Sigma_p^{-1} \Sigma_r = C_r C_r^\top + C_q C_q^\top - C_r B_p^{-1} \underbrace{\Sigma_r}_{B_p C_r^\top} = C_q C_q^\top \; .$$

If $\mathbf{X}_p$ and $\mathbf{X}_q$ are independent, then they are obviously uncorrelated, as $\Sigma_r = \mathbb{E}[(\mathbf{X}_p - \boldsymbol{\mu}_p)(\mathbf{X}_q - \boldsymbol{\mu}_q)^\top] = \mathbb{E}(\mathbf{X}_p - \boldsymbol{\mu}_p)\,\mathbb{E}(\mathbf{X}_q - \boldsymbol{\mu}_q)^\top = O$. Conversely, if $\Sigma_r = O$, then by (3.34) the conditional distribution of $\mathbf{X}_q$ given $\mathbf{X}_p$ is the same as the unconditional distribution of $\mathbf{X}_q$; that is, $\mathsf{N}(\boldsymbol{\mu}_q, \Sigma_q)$. In other words, $\mathbf{X}_q$ is independent of $\mathbf{X}_p$. □

Theorem 3.9. (Relationship between Normal and χ^2 Distributions). If $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ is an n-dimensional normal random with vector with $\det(\Sigma) > 0$, then

$$(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_n^2 \; . \tag{3.35}$$

Proof. Let $BB^\top$ be the Cholesky factorization of Σ, where B is invertible. Since $\mathbf{X}$ can be written as $\boldsymbol{\mu} + B\mathbf{Z}$, where $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ is a vector of independent standard normal random variables, we have

$$(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = (\mathbf{X} - \boldsymbol{\mu})^\top (BB^\top)^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Z}^\top \mathbf{Z} = \sum_{i=1}^n Z_i^2 \; .$$

The moment generating function of $Y = \sum_{i=1}^n Z_i^2$ is given by

$$\mathbb{E}\,\mathrm{e}^{tY} = \mathbb{E}\,\mathrm{e}^{t(Z_1^2 + \cdots + Z_n^2)} = \mathbb{E}\,[\mathrm{e}^{tZ_1^2} \cdots \mathrm{e}^{tZ_n^2}] = \left(\mathbb{E}\,\mathrm{e}^{tZ^2}\right)^n \; ,$$

where $Z \sim \mathsf{N}(0,1)$. The moment generating function of Z^2 is

$$\mathbb{E}\,\mathrm{e}^{tZ^2} = \int_{-\infty}^{\infty} \mathrm{e}^{tz^2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}(1-2t)z^2} \mathrm{d}z = \frac{1}{\sqrt{1-2t}} \; ,$$

so that

$$\mathbb{E}\mathrm{e}^{tY} = \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{n}{2}}, \quad t < \frac{1}{2} \; ,$$

which is the moment generating function of the $\mathsf{Gamma}(n/2, 1/2)$ distribution; that is, the χ_n^2 distribution — see Theorem 2.18. □ ☞ 48

A consequence of Theorem 3.9 is that if $\mathbf{X} = (X_1, \ldots, X_n)^\top$ is n-dimensional standard normal, then the squared length $\|\mathbf{X}\|^2 = X_1^2 + \cdots + X_n^2$ has a χ_n^2 distribution. If instead $X_i \sim \mathsf{N}(\mu_i, 1)$, $i = 1, \ldots$, then $\|\mathbf{X}\|^2$ is said to have a **noncentral $\boldsymbol{\chi_n^2}$ distribution**. This distribution depends on the $\{\mu_i\}$ only through the norm $\|\boldsymbol{\mu}\|$; see Problem 3.22. We write $\|\mathbf{X}\|^2 \sim \chi_n^2(\theta)$, where $\theta = \|\boldsymbol{\mu}\|$ is the **noncentrality parameter**.

Such distributions frequently occur in statistics when considering *projections* of multivariate normal random variables. The proof of the following theorem can be found in Appendix B.4. ☞ 383

Theorem 3.10. (Relationship between Normal and Noncentral $\boldsymbol{\chi^2}$ Distributions). Let $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, I)$ be an n-dimensional normal random vector and let $\mathscr{V}_k$ and $\mathscr{V}_m$ be linear subspaces of dimensions k and m, respectively, with $k < m \leqslant n$. Let $\mathbf{X}_k$ and $\mathbf{X}_m$ be orthogonal projections of $\mathbf{X}$ onto $\mathscr{V}_k$ and $\mathscr{V}_m$, and let $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_m$ be the corresponding projections of $\boldsymbol{\mu}$. Then, the following holds.

1. The random vectors $\mathbf{X}_k$, $\mathbf{X}_m - \mathbf{X}_k$, and $\mathbf{X} - \mathbf{X}_m$ are independent.
2. $\|\mathbf{X}_k\|^2 \sim \chi_k^2(\|\boldsymbol{\mu}_k\|)$, $\|\mathbf{X}_m - \mathbf{X}_k\|^2 \sim \chi_{m-k}^2(\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_k\|)$, and $\|\mathbf{X} - \mathbf{X}_m\|^2 \sim \chi_{n-m}^2(\|\boldsymbol{\mu} - \boldsymbol{\mu}_m\|)$.

Theorem 3.10 is frequently used in the statistical analysis of *normal linear models*; see Section 5.3.1. In typical situations $\boldsymbol{\mu}$ lies in the subspace $\mathscr{V}_m$ or even $\mathscr{V}_k$ — in which case $\|\mathbf{X}_m - \mathbf{X}_k\|^2 \sim \chi_{m-k}^2$ and $\|\mathbf{X} - \mathbf{X}_m\|^2 \sim \chi_{n-m}^2$, independently. The (scaled) quotient then turns out to have an F distribution — a consquence of the following theorem. ☞ 137

Theorem 3.11. (Relationship between $\boldsymbol{\chi^2}$ and $\boldsymbol{F}$ Distributions). Let $U \sim \chi_m^2$ and $V \sim \chi_n^2$ be independent. Then,

$$\frac{U/m}{V/n} \sim \mathsf{F}(m, n) \; .$$

Proof. For notational simplicity, let $c = m/2$ and $d = n/2$. It follows from Example 3.7 that the pdf of $W = U/V$ is given by ☞ 68

$$
\begin{aligned}
f_W(w) &= \int_0^\infty f_U(wv)\, v\, f_V(v)\, \mathrm{d}v \\
&= \int_0^\infty \frac{(wv)^{c-1}\, \mathrm{e}^{-wv/2}}{\Gamma(c)\, 2^c}\, v\, \frac{v^{d-1} \mathrm{e}^{-v/2}}{\Gamma(d)\, 2^d}\, \mathrm{d}v \\
&= \frac{w^{c-1}}{\Gamma(c)\, \Gamma(d)\, 2^{c+d}} \int_0^\infty v^{c+d-1}\, \mathrm{e}^{-(1+w)v/2}\, \mathrm{d}v \\
&= \frac{\Gamma(c+d)}{\Gamma(c)\, \Gamma(d)}\, \frac{w^{c-1}}{(1+w)^{c+d}}\,,
\end{aligned}
$$

where the last equality follows from the fact that the integrand is equal to $\Gamma(\alpha)\lambda^\alpha$ times the density of the $\mathsf{Gamma}(\alpha, \lambda)$ distribution with $\alpha = c + d$ and $\lambda = (1 + w)/2$. The proof is completed by observing that the density of $Z = \frac{n}{m}\frac{U}{V}$ is given by

$$f_Z(z) = f_W(z\, m/n)\, m/n\ .$$

□

Corollary 3.2. (Relationship between Normal, χ^2, and t Distributions). Let $Z \sim \mathsf{N}(0, 1)$ and $V \sim \chi^2_n$ be independent. Then,

$$\frac{Z}{\sqrt{V/n}} \sim \mathsf{t}_n\ .$$

Proof. Let $T = Z/\sqrt{V/n}$. Because $Z^2 \sim \chi^2_1$, we have by Theorem 3.11 that ☞ 50 $T^2 \sim \mathsf{F}(1, n)$. The result follows now from Theorem 2.19 and the symmetry around 0 of the pdf of T. □

3.7 Limit Theorems

Two main results in probability are the *law of large numbers* and *the central limit theorem.* Both are limit theorems involving sums of independent random variables. In particular, consider a sequence $X_1, X_2, \ldots$ of iid random variables with finite expectation μ and finite variance σ^2. For each n define $S_n = X_1 + \cdots + X_n$. What can we say about the (random) sequence ☞ 72 of sums $S_1, S_2, \ldots$ or averages $S_1, S_2/2, S_3/3, \ldots$? By (3.14) and (3.18) we have $\mathbb{E}[S_n/n] = \mu$ and $\mathrm{Var}(S_n/n) = \sigma^2/n$. Hence, as n increases the variance of the (random) average S_n/n goes to 0. Informally, this means that (S_n/n) tends to the constant μ, as $n \to \infty$. This makes intuitive sense, but the important point is that the mathematical theory *confirms* our intuition in this respect. Here is a more precise statement.

Theorem 3.12. (Weak Law of Large Numbers). If $X_1, \ldots, X_n$ are iid with finite expectation μ and finite variance σ^2, then for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(|S_n/n - \mu| > \varepsilon\right) = 0 \;.$$

Proof. Let $Y = (S_n/n - \mu)^2$ and $\delta = \varepsilon^2$. We have

$$\begin{aligned}\mathrm{Var}(S_n/n) &= \mathbb{E}Y = \mathbb{E}[Y \mathrm{I}_{\{Y>\delta\}}] + \mathbb{E}[Y \mathrm{I}_{\{Y\leqslant\delta\}}] \geqslant \mathbb{E}[\delta\, \mathrm{I}_{\{Y>\delta\}}] + 0 \\ &= \delta\, \mathbb{P}(Y > \delta) = \varepsilon^2\, \mathbb{P}(|S_n/n - \mu| > \varepsilon) \;.\end{aligned}$$

Rearranging gives

$$\mathbb{P}(|S_n/n - \mu| > \varepsilon) \leqslant \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{\sigma^2}{n\,\varepsilon^2} \;.$$

The proof is concluded by observing that $\sigma^2/(n\varepsilon^2)$ goes to 0 as $n \to \infty$. □

Remark 3.4. In Theorem 3.12 the qualifier "weak" is used to distinguish the result from the *strong* law of large numbers, which states that

$$\mathbb{P}(\lim_{n\to\infty} S_n/n = \mu) = 1 \;.$$

In terms of a computer simulation this means that the probability of drawing a sequence for which the sequence of averages fails to converge to μ is zero. The strong law implies the weak law, but is more difficult to prove in its full generality; see, for example, [Feller, 1970].

The central limit theorem describes the approximate distribution of S_n (or S_n/n), and it applies to both continuous and discrete random variables. Loosely, it states that

> *the sum of a large number of iid random variables approximately has a normal distribution.*

Specifically, the random variable S_n has a distribution that is approximately normal, with expectation $n\mu$ and variance $n\sigma^2$. A more precise statement is given next.

Theorem 3.13. (Central Limit Theorem). If $X_1, \dots, X_n$ are iid with finite expectation μ and finite variance σ^2, then for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant x\right) = \varPhi(x) ,$$

where $\varPhi$ is the cdf of the standard normal distribution.

Proof. (Sketch) A full proof is out of the scope of this book. However, the main ideas are not difficult. Without loss of generality assume $\mu = 0$ and $\sigma = 1$. This amounts to replacing X_n by $(X_n - \mu)/\sigma$. We also assume, for simplicity, that the moment generating function of X_i is finite in an open interval containing 0, so that we can use Theorem 2.7. We wish to show that the cdf of $S_n/\sqrt{n}$ converges to that of the standard normal distribution. It can be proved (and makes intuitive sense) that this is equivalent (up to some technical conditions) to demonstrating that the corresponding moment generating functions converge. That is, we wish to show that ☞ 36

$$\lim_{n\to\infty} \mathbb{E}\exp\left(t\frac{S_n}{\sqrt{n}}\right) = \mathrm{e}^{\frac{1}{2}t^2}, \;\; t \in \mathbb{R} ,$$

where the right-hand side is the moment generating function of the standard normal distribution. Because $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 = \mathrm{Var}(X_1) = 1$, we have by Theorem 2.7 that the moment generation function of X_1 has the following Taylor expansion: ☞ 381

$$M(t) \overset{\text{def}}{=} \mathbb{E}\,\mathrm{e}^{tX_1} = 1 + t\,\mathbb{E}X_1 + \frac{1}{2}t^2\,\mathbb{E}X_1^2 + o(t^2) = 1 + \frac{1}{2}\,t^2 + o(t^2) ,$$

where $o(t^2)$ is a function for which $\lim_{t\downarrow 0} o(t^2)/t^2 = 0$. Because the $\{X_i\}$ are iid, it follows that the moment generating function of $S_n/\sqrt{n}$ satisfies

$$\begin{aligned}\mathbb{E}\exp\left(t\frac{S_n}{\sqrt{n}}\right) &= \mathbb{E}\exp\left(\frac{t}{\sqrt{n}}(X_1 + \cdots + X_n)\right) = \prod_{i=1}^{n}\mathbb{E}\exp\left(\frac{t}{\sqrt{n}}X_i\right) \\ &= M^n\left(\frac{t}{\sqrt{n}}\right) = \left[1 + \frac{t^2}{2n} + o(t^2/n)\right]^n \longrightarrow \mathrm{e}^{\frac{1}{2}t^2}\end{aligned}$$

as $n \to \infty$. □

Figure 3.7 shows central limit theorem in action. The left part shows the pdfs of $S_1, \dots, S_4$ for the case where the $\{X_i\}$ have a $\mathsf{U}[0,1]$ distribution. The right part shows the same for the $\mathsf{Exp}(1)$ distribution. We clearly see convergence to a bell-shaped curve, characteristic of the normal distribution.

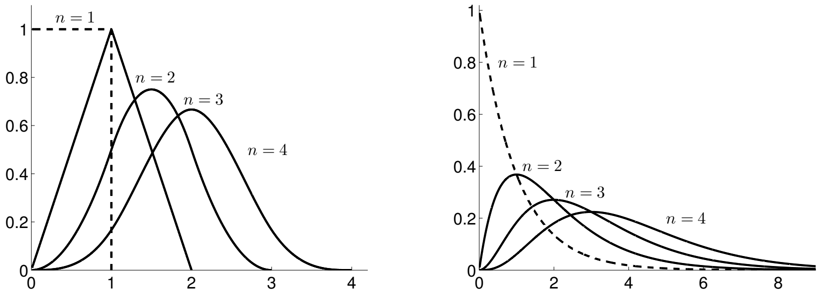

Fig. 3.7 Illustration of the central limit theorem for (left) the uniform distribution and (right) the exponential distribution.

Recall that a binomial random variable $X \sim \mathsf{Bin}(n, p)$ can be viewed as the sum of n iid $\mathsf{Ber}(p)$ random variables: $X = X_1 + \cdots + X_n$. As a direct consequence of the central limit theorem it follows that for large n $\mathbb{P}(X \leqslant k) \approx \mathbb{P}(Y \leqslant k)$, where $Y \sim \mathsf{N}(np, np(1-p))$. As a rule of thumb, this normal approximation to the binomial distribution is accurate if both np and $n(1-p)$ are larger than 5. ☞ 64

There is also a central limit theorem for random vectors. The multidimensional version is as follows.

Theorem 3.14. (Multivariate Central Limit Theorem). Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be iid random vectors with expectation vector $\boldsymbol{\mu}$ and covariance matrix Σ. For large n the random vector $\mathbf{X}_1 + \cdots + \mathbf{X}_n$ approximately has a $\mathsf{N}(n\boldsymbol{\mu}, n\Sigma)$ distribution.

A more precise formulation of the above theorem is that the average random vector $\mathbf{Z}_n = (\mathbf{X}_1 + \cdots + \mathbf{X}_n)/n$, when rescaled via $\sqrt{n}(\mathbf{Z}_n - \boldsymbol{\mu})$, converges in distribution to a random vector $\mathbf{K} \sim \mathsf{N}(\mathbf{0}, \Sigma)$ as $n \to \infty$. A useful consequence of this is given next.

Theorem 3.15. (Delta Method). Let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ be a sequence of random vectors such that $\sqrt{n}(\mathbf{Z}_n - \boldsymbol{\mu}) \to \mathbf{K} \sim \mathsf{N}(\mathbf{0}, \Sigma)$ as $n \to \infty$. Then, for any continuously differentiable function $\mathbf{g}$ of $\mathbf{Z}_n$,

$$\sqrt{n}(\mathbf{g}(\mathbf{Z}_n) - \mathbf{g}(\boldsymbol{\mu})) \to \mathbf{R} \sim \mathsf{N}(\mathbf{0}, J\Sigma J^\top) \ , \tag{3.36}$$

where $J = J(\boldsymbol{\mu}) = (\partial g_i(\boldsymbol{\mu})/\partial x_j)$ is the Jacobian matrix of $\mathbf{g}$ evaluated at $\boldsymbol{\mu}$.

Proof. (Sketch) A formal proof requires some deeper knowledge of statistical convergence, but the idea of the proof is quite straightforward. The key step is ☞ 381 to construct the first-order Taylor expansion (see Theorem B.1) of $\mathbf{g}$ around $\boldsymbol{\mu}$, which yields

$$\mathbf{g}(\mathbf{Z}_n) = \mathbf{g}(\boldsymbol{\mu}) + J(\boldsymbol{\mu})(\mathbf{Z}_n - \boldsymbol{\mu}) + \mathcal{O}(\|\mathbf{Z}_n - \boldsymbol{\mu}\|^2) \ .$$

As $n \to \infty$, the remainder term goes to 0, because $\mathbf{Z}_n \to \boldsymbol{\mu}$. Hence, the left-hand side of (3.36) is approximately $J\sqrt{n}(\mathbf{Z}_n - \boldsymbol{\mu})$. For large n this converges ☞ 77 to a random vector $\mathbf{R} = J\,\mathbf{K}$, where $\mathbf{K} \sim \mathsf{N}(\mathbf{0}, \Sigma)$. Finally, by Theorem 3.4 we have $\mathbf{R} \sim \mathsf{N}(\mathbf{0}, J\Sigma J^\top)$. □

Example 3.15 (Ratio Estimator). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid copies of a random vector (X, Y) with mean vector (μ_X, μ_Y) and covariance matrix Σ. Denoting the average of the $\{X_i\}$ and $\{Y_i\}$ by $\bar{X}$ and $\bar{Y}$ respectively, what can we say about the distribution of $\bar{X}/\bar{Y}$ for large n?

Let $\mathbf{Z}_n = (\bar{X}, \bar{Y})$ and $\boldsymbol{\mu} = (\mu_X, \mu_Y)$. By the multivariate central limit theorem $\mathbf{Z}_n$ has approximately a $\mathsf{N}(\boldsymbol{\mu}, \Sigma/n)$ distribution. More precisely, $\sqrt{n}(\mathbf{Z}_n - \boldsymbol{\mu})$ converges to a $\mathsf{N}(\mathbf{0}, \Sigma)$-distributed random vector.

We apply the delta method using the function $g(x, y) = x/y$, whose Jacobian matrix is

$$J(x, y) = \left(\frac{\partial g(x,y)}{\partial x}, \quad \frac{\partial g(x,y)}{\partial y} \right) = \left(\frac{1}{y}, \quad \frac{-x}{y^2} \right) \ .$$

It follows from (3.36) that $g(\bar{X}, \bar{Y}) = \bar{X}/\bar{Y}$ has approximately a normal distribution with expectation $g(\boldsymbol{\mu}) = \mu_X/\mu_Y$ and variance σ^2/n, where

$$\begin{aligned} \sigma^2 = J(\boldsymbol{\mu})\Sigma J^\top(\boldsymbol{\mu}) &= \left(\frac{1}{\mu_Y}, \quad \frac{-\mu_X}{\mu_Y^2} \right) \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{pmatrix} \begin{pmatrix} \frac{1}{\mu_Y} \\ \frac{-\mu_X}{\mu_Y^2} \end{pmatrix} \\ &= \left(\frac{\mu_X}{\mu_Y} \right)^2 \left(\frac{\mathrm{Var}(X)}{\mu_X^2} + \frac{\mathrm{Var}(Y)}{\mu_Y^2} - 2\frac{\mathrm{Cov}(X,Y)}{\mu_X\,\mu_Y} \right) \ . \end{aligned} \tag{3.37}$$

3.8 Problems

3.1. Let U and V be independent random variables with $\mathbb{P}(U = 1) = \mathbb{P}(V = 1) = 1/4$ and $\mathbb{P}(U = -1) = \mathbb{P}(V = -1) = 3/4$. Define $X = U/V$ and $Y = U + V$. Give the joint discrete pdf of X and Y in table form, as in ☞ 62 Table 3.1. Are X and Y independent?

3.2. Let $X_1, \ldots, X_4 \sim_{\text{iid}} \mathsf{Ber}(p)$.

a. Give the joint discrete pdf of $X_1, \ldots, X_4$.

b. Give the joint discrete pdf of $X_1, \ldots, X_4$ given $X_1 + \cdots + X_4 = 2$.

3.3. Three identical-looking urns each have 4 balls. Urn 1 has 1 red and 3 white balls, Urn 2 has 2 red and 2 white balls, and Urn 3 has 3 red and 1 white ball. We randomly select an urn with equal probability. Let X be the number of the urn. We then draw 2 balls from the selected urn. Let Y be the number of red balls drawn. Find the following discrete pdfs.

a. The pdf of X.
b. The conditional pdf of Y given $X = x$ for $x = 1, 2, 3$.
c. The joint pdf of X and Y.
d. The pdf of Y.
e. The conditional pdf of X given $Y = y$ for $y = 0, 1, 2$.

3.4. We randomly select a point (X, Y) from the triangle $\{(x, y) : x, y \in \{1, \ldots, 6\}, y \leqslant x\}$ (see Figure 3.1) in the following *non-uniform* way. First, select X discrete uniformly from $\{1, \ldots, 6\}$. Then, given $X = x$, select Y discrete uniformly from $\{1, \ldots, x\}$. Find the conditional distribution of X given $Y = 1$ and its corresponding conditional expectation.

☞ 65

3.5. We randomly and uniformly select a continuous random vector (X, Y) in the triangle $(0, 0)$–$(1, 0)$–$(1, 1)$; the same triangle as in Example 3.8.

☞ 70

a. Give the joint pdf of X and Y.
b. Calculate the pdf of Y and sketch its graph.
c. Specify the conditional pdf of Y given $X = x$ for any fixed $x \in (0, 1)$.
d. Determine $\mathbb{E}[Y \mid X = 1/2]$.

3.6. Let $X \sim \mathsf{U}[0, 1]$ and $Y \sim \mathsf{Exp}(1)$ be independent.

a. Determine the joint pdf of X and Y and sketch its graph.
b. Calculate $\mathbb{P}((X, Y) \in [0, 1] \times [0, 1])$,
c. Calculate $\mathbb{P}(X + Y < 1)$.

3.7. Let $X \sim \mathsf{Exp}(\lambda)$ and $Y \sim \mathsf{Exp}(\mu)$ be independent.

a. Show that $\min(X, Y)$ also has an exponential distribution, and determine its corresponding parameter.
b. Show that

$$\mathbb{P}(X < Y) = \frac{\lambda}{\lambda + \mu} .$$

3.8. Let $X \sim \mathsf{Exp}(1)$ and $(Y \mid X = x) \sim \mathsf{Exp}(x)$.

a. What is the joint pdf of X and Y?
b. What is the marginal pdf of Y?

3.9. Let $X \sim \mathsf{U}(-\pi/2, \pi/2)$. Show that $Y = \tan(X)$ has a Cauchy distribution.

☞ 69

3.10. Let $X \sim \mathsf{Exp}(3)$ and $Y = \ln(X)$. What is the pdf of Y?

3.11. We draw n numbers independently and uniformly from the interval [0,1] and denote their sum S_n.

a. Determine the pdf of S_2 and sketch its graph.
b. What is approximately the distribution of S_{20}?
c. Approximate the probability that the average of the 20 numbers is greater than 0.6.

3.12. A certain type of electrical component has an exponential lifetime distribution with an expected lifetime of 1/2 year. When the component fails it is immediately replaced by a second (new) component; when the second component fails, it is replaced by a third, etc. Suppose there are 10 such identical components. Let T be the time that the last of the components fails.

a. What is the expectation and variance of T?
b. Approximate, using the central limit theorem, the probability that T exceeds 6 years.
c. What is the exact distribution of T?

3.13. Let A be an invertible $n \times n$ matrix and let $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(0,1)$. Define $\mathbf{X} = (X_1, \ldots, X_n)^\top$ and let $(Z_1, \ldots, Z_n)^\top = A\mathbf{X}$. Show that $Z_1, \ldots, Z_n$ are iid standard normal only if $AA^\top = I$ (identity matrix); in other words, only if A is an *orthogonal* matrix. Can you find a geometric interpretation of this?

3.14. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Let $\bar{X} = (X_1 + \cdots + X_n)/n$. Calculate the correlation coefficient of X_1 and $\bar{X}$.

3.15. Suppose that $X_1, \ldots, X_6$ are iid with pdf

$$f(x) = \begin{cases} 3x^2, \, 0 \leqslant x \leqslant 1, \\ 0, \quad \text{elsewhere.} \end{cases}$$

a. What is the probability that all $\{X_i\}$ are greater than 1/2?
b. Find the probability that at least one of the $\{X_i\}$ is less than 1/2.

3.16. Let X and Y be random variables.

a. Express $\mathrm{Var}(-aX+Y)$, where a is a constant, in terms of $\mathrm{Var}(X), \mathrm{Var}(Y)$, and $\mathrm{Cov}(X,Y)$.
b. Take $a = \mathrm{Cov}(X,Y)/\mathrm{Var}(X)$. Using the fact that the variance in (a) is always non-negative, prove the following **Cauchy–Schwartz inequality**:

$$(\mathrm{Cov}(X,Y))^2 \leqslant \mathrm{Var}(X)\,\mathrm{Var}(Y) \; .$$

c. Show that, as a consequence, the correlation coefficient of X and Y must lie between -1 and 1.

3.17. Suppose X and Y are independent uniform random variables on [0,1]. Let $U = X/Y$ and $V = XY$, which means $X = \sqrt{UV}$ and $Y = \sqrt{V/U}$.

a. Sketch the two-dimensional region where the density of (U, V) is non-zero.
b. Find the matrix of Jacobi for the transformation $(x, y)^\top \mapsto (u, v)^\top$.
c. Show that its determinant is $2x/y = 2u$.
d. What is the joint pdf of U and V?
e. Show that the marginal pdf of U is

$$f_U(u) = \begin{cases} \frac{1}{2}, & 0 < u < 1 \\ \frac{1}{2u^2}, & u \geqslant 1 \end{cases} . \tag{3.38}$$

3.18. Let $X_1, \ldots, X_n$ be iid with mean μ and variance σ^2. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $Y = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

a. Show that
$$Y = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 .$$
b. Calculate $\mathbb{E}Y$.
c. Show that $\mathbb{E}Y \to \sigma^2$ as $n \to \infty$.

3.19. Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$, with $\{X_i\} \sim_{\text{iid}} \mathsf{N}(\mu, 1)$. Consider the orthogonal projection, denoted $\mathbf{X}_1$, of $\mathbf{X}$ onto the subspace spanned by $\mathbf{1} = (1, \ldots, 1)^\top$.

a. Show that $\mathbf{X}_1 = \bar{X}\mathbf{1}$.
b. Show that $\mathbf{X}_1$ and $\mathbf{X} - \mathbf{X}_1$ are independent.
c. Show that $\|\mathbf{X} - \mathbf{X}_1\|^2 = \sum_{i=1}^n (X_i - \bar{X})^2$ has a χ^2_{n-1} distribution.

Hint: apply Theorem 3.10.

3.20. Let $X_1, \ldots, X_6$ be the weights of six randomly chosen people. Assume each weight is $\mathsf{N}(75, 100)$ distributed (in kg). Let $W = X_1 + \cdots + X_6$ be the total weight of the group. Explain why the distribution of W is equal or not equal to $6X_1$.

3.21. Let $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ be independent. Show that $X + Y \sim \chi^2_{m+n}$. Hint: use moment generating functions.

3.22. Let $X \sim \mathsf{N}(\mu, 1)$. Show that the moment generation function of X^2 is

$$M(t) = \frac{\mathrm{e}^{\mu^2 t/(1-2t)}}{\sqrt{1-2t}} \qquad t < 1/2 .$$

Next, consider independent random variables $X_i \sim \mathsf{N}(\mu_i, 1)$, $i = 1, \ldots, n$. Use the result above to show that the distribution of $\|\mathbf{X}\|^2$ only depends on n and $\|\boldsymbol{\mu}\|$. Can you find a symmetry argument why this must be so?

3.23. A machine produces cylinders with a diameter which is normally distributed with mean 3.97 cm and standard deviation 0.03 cm. Another machine produces (independently of the first machine) shafts with a diameter which is normally distributed with mean 4.05 cm and standard deviation 0.02cm. What is the probability that a randomly chosen cylinder fits into a randomly chosen shaft?

3.24. A sieve with diameter d is used to separate a large number of blueberries into two classes: small and large. Suppose that the diameters of the blueberries are normally distributed with an expectation $\mu = 1$ (cm) and a standard deviation $\sigma = 0.1$ (cm).

a. Find the diameter of the sieve such that the proportion of large blueberries is 30%.
b. Suppose that the diameter is chosen such as in (a). What is the probability that out of 1000 blueberries, fewer than 280 end up in the "large" class?

3.25. Suppose X, Y, and Z are independent $\mathsf{N}(1, 2)$-distributed random variables. Let $U = X - 2Y + 3Z$ and $V = 2X - Y + Z$. Give the joint distribution of U and V.

3.26. For many of the above problems it is instructive to simulate the corresponding model on a computer in order to better understand the theory.

a. Generate 10^5 points (X, Y) from the model in Problem 3.6.
b. Compare the fraction of points falling in the unit square $[0, 1] \times [0, 1]$ with the theoretical probability in Problem 3.6 (b).
c. Do the same for the probability $\mathbb{P}(X + Y < 1)$.

3.27. Simulate 10^5 draws from $\mathsf{U}(-\pi/2, \pi/2)$ and transform these using the tangent function, as in Problem 3.9. Compare the histogram of the transformed values with the theoretical (Cauchy) pdf.

3.28. Simulate 10^5 independent draws of (U, V) in Problem 3.17. Verify with a histogram of the U-values that the pdf of U is of the form (3.38).

3.29. Consider the Matlab experiments in Example 3.14.

☞ 81 a. Carry out the experiments with $\varrho = 0.4, 0.7, 0.9, 0.99$, and -0.8, and observe how the outcomes change.
b. Plot the corresponding pdfs, as in Figure 3.6.
c. Give also the contour plots of the pdfs, for $\varrho = 0$ and $\varrho = 0.8$. Observe that the contours are *ellipses*.
d. Show that these ellipses are of the form

$$x_1^2 + 2\varrho\, x_1\, x_2 + x_2^2 = \text{constant} .$$

Index