SIMULATION AND THE MONTE CARLO METHOD
WILEY SERIES IN PROBABILITY AND STATISTICS
Established by Walter A. Shewhart and Samuel S. Wilks

Editors: David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice,
Geof H. Givens, Harvey Goldstein, Geert Molenberghs, David W. Scott,
Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg

Editors Emeriti: J. Stuart Hunter, Iain M. Johnstone, Joseph B. Kadane,
Jozef L. Teugels

The Wiley Series in Probability and Statistics is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at http://www.wiley.com/go/wsps
To the memory of

Reuven Y. Rubinstein (1938–2012)

Reuven Rubinstein passed away during the writing of this third edition. Reuven was one of the pioneers of Monte Carlo simulation and remained at the forefront of research in this area right up to the end of his life. In 2011 he received the highest honor given by INFORMS Simulation Society: The Lifetime Professional Achievement Award, where his achievements were summarized as follows:

Professor Rubinstein has been a pivotal figure in the theory and practice of simulation as we know it today. His career reflects a high level of creativity and contribution, with a willingness to explore new areas and an amazing ability to suggest surprising new avenues of research and to influence subsequent work.

May his contagious enthusiasm and curiosity live on through his books and the many people whom he inspired.

Dirk P. Kroese
CONTENTS

Preface xiii
Acknowledgments xvii

1 Preliminaries 1

1.1 Introduction 1
1.2 Random Experiments 1
1.3 Conditional Probability and Independence 2
1.4 Random Variables and Probability Distributions 4
1.5 Some Important Distributions 5
1.6 Expectation 6
1.7 Joint Distributions 7
1.8 Functions of Random Variables 11
  1.8.1 Linear Transformations 12
  1.8.2 General Transformations 13
1.9 Transforms 14
1.10 Jointly Normal Random Variables 15
1.11 Limit Theorems 16
1.12 Poisson Processes 17
1.13 Markov Processes 19
  1.13.1 Markov Chains 19
  1.13.2 Classification of States 21
### 6 Markov Chain Monte Carlo

6.1 Introduction 187
6.2 Metropolis–Hastings Algorithm 188
6.3 Hit-and-Run Sampler 193
6.4 Gibbs Sampler 194
6.5 Ising and Potts Models 197
   6.5.1 Ising Model 197
   6.5.2 Potts Model 198
6.6 Bayesian Statistics 200
6.7 Other Markov Samplers 202
   6.7.1 Slice Sampler 204
   6.7.2 Reversible Jump Sampler 205
6.8 Simulated Annealing 208
6.9 Perfect Sampling 212
Problems 214
References 219

### 7 Sensitivity Analysis and Monte Carlo Optimization

7.1 Introduction 221
7.2 Score Function Method for Sensitivity Analysis of DESS 224
7.3 Simulation-Based Optimization of DESS 231
   7.3.1 Stochastic Approximation 232
   7.3.2 Stochastic Counterpart Method 237
7.4 Sensitivity Analysis of DEDS 246
Problems 252
References 255

### 8 Cross-Entropy Method

8.1 Introduction 257
8.2 Estimation of Rare-Event Probabilities 258
   8.2.1 Root-Finding Problem 267
   8.2.2 Screening Method for Rare Events 268
   8.2.3 CE Method Combined with Sampling from the Zero-Variance Distribution 271
8.3 CE Method for Optimization 272
8.4 Max-Cut Problem 276
8.5 Partition Problem 282
  8.5.1 Empirical Computational Complexity 283
8.6 Traveling Salesman Problem 283
  8.6.1 Incomplete Graphs 288
  8.6.2 Node Placement 289
  8.6.3 Case Studies 290
8.7 Continuous Optimization 291
8.8 Noisy Optimization 292
8.9 MinxEnt Method 294
Problems 298
References 303

9 Splitting Method 307
  9.1 Introduction 307
  9.2 Counting Self-Avoiding Walks via Splitting 308
  9.3 Splitting with a Fixed Splitting Factor 310
  9.4 Splitting with a Fixed Effort 313
  9.5 Generalized Splitting 314
  9.6 Adaptive Splitting 318
  9.7 Application of Splitting to Network Reliability 321
  9.8 Applications to Counting 322
  9.9 Case Studies for Counting with Splitting 325
    9.9.1 Satisfiability (SAT) Problem 325
    9.9.2 Independent Sets 330
    9.9.3 Permanent and Counting Perfect Matchings 332
    9.9.4 Binary Contingency Tables 334
    9.9.5 Vertex Coloring 336
  9.10 Splitting as a Sampling Method 337
  9.11 Splitting for Optimization 340
    9.11.1 Continuous Optimization 343
Problems 344
References 348

10 Stochastic Enumeration Method 351
  10.1 Introduction 351
  10.2 Tree Search and Tree Counting 352
  10.3 Knuth’s Algorithm for Estimating the Cost of a Tree 355
  10.4 Stochastic Enumeration 357
    10.4.1 Combining SE with Oracles 359
  10.5 Application of SE to Counting 360
Since the publication in 2008 of the second edition of *Simulation and the Monte Carlo Method*, significant changes have taken place in the field of Monte Carlo simulation. This third edition gives a fully updated and comprehensive account of the major topics in Monte Carlo simulation.

The book is based on an undergraduate course on Monte Carlo methods given at the Israel Institute of Technology (Technion) and the University of Queensland for the last five years. It is aimed at a broad audience of students in engineering, physical and life sciences, statistics, computer science, mathematics, and simply anyone interested in using Monte Carlo simulation in their study or work. Our aim is to provide an accessible introduction to modern Monte Carlo methods, focusing on the main concepts, while providing a sound foundation for problem solving. For this reason most ideas are introduced and explained via concrete examples, algorithms, and experiments.

Although we assume that the reader has some basic mathematical background, such as an elementary course in probability and statistics, we nevertheless review the basic concepts of probability, stochastic processes, information theory, and convex optimization in Chapter 1.

In a typical stochastic simulation, randomness is introduced into simulation models via independent uniformly distributed random variables. These random variables are then used as building blocks to simulate more general stochastic systems. Chapter 2 deals with the generation of such random numbers, random variables, and stochastic processes.

Many real-world complex systems can be modeled as discrete-event systems. Examples of discrete-event systems include traffic systems, flexible manufacturing
systems, computer-communications systems, inventory systems, production lines, coherent lifetime systems, PERT networks, and flow networks. The behavior of such systems is identified via a sequence of discrete “events” that causes the system to change from one “state” to another. We discuss how to model such systems on a computer in Chapter 3.

Chapter 4 treats the statistical analysis of the output data from static and dynamic simulation models. The main difference is that the former do not evolve in time whereas the latter do. For dynamic models, we distinguish between finite-horizon and steady-state simulations. Two popular methods for estimating steady-state performance measures — the batch means and regenerative methods — are discussed as well.

Chapter 5 deals with variance reduction techniques in Monte Carlo simulation, such as antithetic and common random numbers, control random variables, conditional Monte Carlo, stratified sampling, and importance sampling. Using importance sampling, one can often achieve substantial (sometimes dramatic) variance reduction, in particular when estimating rare-event probabilities. While dealing with importance sampling, we present two alternative approaches, called the variance minimization and the cross-entropy methods. Special attention is paid to importance sampling algorithms in which paths are generated in a sequential manner. Further improvements of such algorithms are obtained by resampling successful paths, giving rise to sequential importance resampling algorithms. We illustrate their use via a nonlinear filtering example. In addition, this chapter contains two new importance sampling based methods, called the transform likelihood ratio method and the screening method for variance reduction. The former presents a simple, convenient and unifying way of constructing efficient importance sampling estimators, whereas the latter ensures lowering of the dimensionality of the importance sampling density. This is accomplished by identifying (screening out) the most important (bottleneck) parameters to be used in the importance sampling distribution. As results, the accuracy of the importance sampling estimator increases substantially.

Chapter 6 gives a concise treatment of the generic Markov chain Monte Carlo (MCMC) method for approximately generating samples from an arbitrary distribution. We discuss the classic Metropolis–Hastings algorithm and the Gibbs sampler. In the former, one simulates a Markov chain such that its stationary distribution coincides with the target distribution, while in the latter the underlying Markov chain is constructed on the basis of a sequence of conditional distributions. We also deal with applications of MCMC in Bayesian statistics, and explain how MCMC is used to sample from the Boltzmann distribution for the Ising and Potts models, which are extensively used in statistical mechanics. Moreover, we show how MCMC is used in the simulated annealing method to find the global minimum of a multi-extremal function. We also show that both the Metropolis–Hastings and Gibbs samplers can be viewed as special cases of a general MCMC algorithm and then present two more modifications, namely the slice and the reversible jump samplers.

Chapter 7 is on sensitivity analysis and Monte Carlo optimization of simulated systems. Because of their complexity, the performance evaluation of discrete-event systems is usually studied by simulation, and the simulation is often associated with the estimation of the performance function with respect to some controllable parameters. Sensitivity analysis is concerned with evaluating sensitivities (gradients, Hessians, etc.) of the performance function with respect to system parameters.
This provides guidance to operational decisions and to selecting system parameters that optimize the performance measures. Monte Carlo optimization deals with solving stochastic programs, that is, optimization problems where the objective function and some of the constraints are unknown and need to be obtained via simulation. We deal with sensitivity analysis and optimization of both static and dynamic models. We introduce the celebrated score function method for sensitivity analysis, and two alternative methods for Monte Carlo optimization, the so-called stochastic approximation and stochastic counterpart methods. In particular, in the latter method, we show how using a single simulation experiment one can approximate quite accurately the true unknown optimal solution of the original deterministic program.

Chapter 8 deals with the cross-entropy (CE) method, which was introduced by the first author in 1997 as an adaptive algorithm for rare-event estimation using a cross-entropy minimization technique. It was soon realized that the underlying ideas had a much wider range of application than just in rare-event simulation: they could be readily adapted to tackle quite general combinatorial and multiextremal optimization problems, including many problems associated with learning algorithms and neural computation. We provide a gradual introduction to the CE method, and show its elegance and versatility. In particular, we present a general CE algorithm for the estimation of rare-event probabilities and then slightly modify it for solving combinatorial optimization problems. We discuss applications of the CE method to several combinatorial optimization problems, such as the max-cut problem and the traveling salesman problem, and provide supportive numerical results on its effectiveness. Due to its versatility, tractability, and simplicity, the CE method has potentially a diverse range of applications, for example, in computational biology, DNA sequence alignment, graph theory, and scheduling. Over the last 10 years many hundreds of papers have been written on the theory and applications of CE. For more details see the site www.cemethod.org, our book *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning* (Springer, 2004), and search in the wikipedia under “cross-entropy method”. The chapter concludes with a discussion of the minimum cross-entropy (MinxEnt) optimization program.

Chapter 9 introduces the splitting method, which uses a sequential sampling plan to decompose a “difficult” problem into a sequence of “easy” problems. The method was originally designed for rare-event simulation, but it has developed into a highly versatile “particle MCMC” algorithm that can be used for rare-event estimation, optimization, and sampling. The chapter presents various splitting algorithms for dynamic and static simulation models, and demonstrates how they can be used to (1) estimate rare-event probabilities, (2) solve hard counting problems, (3) find solutions to challenging optimization problems, and (4) sample from complicated probability distributions. The chapter features a wide variety of case studies and numerical experiments, demonstrating the effectiveness of the method.

Many combinatorial problems can be formulated in terms of searching or counting the total cost of a tree. Chapter 10 presents a new Monte Carlo called stochastic enumeration (SE) that is well suited to solve such problems by generating random paths through the tree in a parallel fashion. The SE algorithm can be viewed as a sequential importance sampling method on a “hyper-tree” whose vertices are sets of vertices of the original tree. By combining SE with fast polynomial decision algorithms, we show how it can be used for counting #P-complete problems, such
as the number of satisfiability assignments, number of paths in a general network, and the number of perfect matchings in a graph. The usefulness of the method is illustrated via a suite of numerical examples.

The appendix features a variety of supplementary topics, including a brief introduction to exponential families, the discrete-time Kalman filter, and the Cholesky square root method. The computational complexity of randomized algorithms is also discussed. An extensive range of exercises is provided at the end of each chapter.

In addition to two brand-new chapters (Chapters 9 and 10), this third edition offers substantial updates on a range of topics. The material on random number generation has been extensively revised by including state-of-the-art combined multiple recursive generators and the Mersenne Twister. The material on stochastic process generation has been extended by including the simulation of Gaussian processes, Brownian motion, and diffusion processes. The variance reduction chapter now includes a discussion of the novel multi-level Monte Carlo method. Our treatment of sequential importance has been significantly modified by emphasizing the significance of importance resampling. This addition also prepares for the particle MCMC approach in the new splitting chapter. The cross-entropy chapter is further enhanced by adding new insights into likelihood ratio degeneration, leading to the single-level improved CE algorithm. Twenty-five more questions have been added, along with their solutions in the online solutions manual that accompanies this book. Finally, to facilitate their implementation, most algorithms have been (re-)written in pseudo-code with flow control.

Reuven Rubinstein and Dirk Kroese

Haifa and Brisbane
July 2016
ACKNOWLEDGMENTS

We thank all who contributed to this book. Robert Smith and Zelda Zabinski read and provided useful suggestions on Chapter 6. Alex Shapiro gave a detailed account of the complexity of stochastic programming problems (Section A.8.4). Pierre L’Ecuyer communicated the counterexample in Section 9.10 and Zdravko Botev suggested the multilevel Monte Carlo example in Section 5.6. Jim Spall kindly provided feedback on the second edition, which we have incorporated in the third edition. Josh Chan, Tom Taimre, and Zdravko Botev helped proofread the new material in this book, which is highly appreciated.

We are particularly grateful to the many undergraduate and graduate students at the Technion and the University of Queensland who helped make this book possible and whose valuable ideas and experiments were extremely encouraging and motivating. Qibin Duan, Morgan Grant, Robert Salomone, Rohan Shah, and Erli Wang read through the new material, solved the new exercises, and provided excellent feedback. Special thanks goes out to Slava Vaisman for his help on many computational problems that we encountered, and for the many fruitful discussions we had on the splitting and stochastic enumeration methods.

This book was supported by by the Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers, under grant number CE140100049.

RYR, DPK
CHAPTER 1

PRELIMINARIES

1.1 INTRODUCTION

The purpose of this chapter is to review some basic facts from probability, information theory, and optimization. In particular, Sections 1.2–1.11 summarize the main points from probability theory. Sections 1.12–1.14 describe various fundamental stochastic processes, such as Poisson, Markov, and Gaussian processes. Elements of information theory are given in Section 1.15, and Section 1.16 concludes with an outline of convex optimization theory.

1.2 RANDOM EXPERIMENTS

The basic notion in probability theory is that of a random experiment: an experiment whose outcome cannot be determined in advance. The most fundamental example is the experiment where a fair coin is tossed a number of times. For simplicity suppose that the coin is tossed three times. The sample space, denoted \( \Omega \), is the set of all possible outcomes of the experiment. In this case \( \Omega \) has eight possible outcomes:

\[ \Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} , \]

where, for example, HTH means that the first toss is heads, the second tails, and the third heads.
Subsets of the sample space are called \textit{events}. For example, the event \( A \) that the third toss is heads is
\[
A = \{HHH, HTH, THH, TTH\}.
\]
We say that event \( A \) \textit{occurs} if the outcome of the experiment is one of the elements in \( A \). Since events are sets, we can apply the usual set operations to them. For example, the event \( A \cup B \), called the \textit{union} of \( A \) and \( B \), is the event that \( A \) or \( B \) or both occur, and the event \( A \cap B \), called the \textit{intersection} of \( A \) and \( B \), is the event that \( A \) and \( B \) both occur. Similar notation holds for unions and intersections of more than two events. The event \( A^c \), called the \textit{complement} of \( A \), is the event that \( A \) does not occur. Two events \( A \) and \( B \) that have no outcomes in common, that is, their intersection is empty, are called \textit{disjoint} events. The main step is to specify the probability of each event.

\textbf{Definition 1.2.1 (Probability)} A \textit{probability} \( P \) is a rule that assigns a number \( 0 \leq P(A) \leq 1 \) to each event \( A \), such that \( P(\Omega) = 1 \), and such that for any sequence \( A_1, A_2, \ldots \) of disjoint events
\[
P\left( \bigcup_i A_i \right) = \sum_i P(A_i). \tag{1.1}
\]
Equation (1.1) is referred to as the \textit{sum rule} of probability. It states that if an event can happen in a number of different ways, but not simultaneously, the probability of that event is simply the sum of the probabilities of the comprising events.

For the fair coin toss experiment the probability of any event is easily given. Namely, because the coin is fair, each of the eight possible outcomes is equally likely, so that \( P(\{HHH\}) = \cdots = P(\{TTT\}) = 1/8 \). Since any event \( A \) is the union of the “elementary” events \( \{HHH\}, \ldots, \{TTT\} \), the sum rule implies that
\[
P(A) = \frac{|A|}{|\Omega|}, \tag{1.2}
\]
where \(|A|\) denotes the number of outcomes in \( A \) and \(|\Omega| = 8 \). More generally, if a random experiment has finitely many and equally likely outcomes, the probability is always of the form (1.2). In that case the calculation of probabilities reduces to counting.

\section*{1.3 CONDITIONAL PROBABILITY AND INDEPENDENCE}

How do probabilities change when we know that some event \( B \subset \Omega \) has occurred? Given that the outcome lies in \( B \), the event \( A \) will occur if and only if \( A \cap B \) occurs, and the relative chance of \( A \) occurring is therefore \( P(A \cap B)/P(B) \). This leads to the definition of the \textit{conditional probability} of \( A \) given \( B \):
\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{1.3}
\]
For example, suppose that we toss a fair coin three times. Let \( B \) be the event that the total number of heads is two. The conditional probability of the event \( A \) that the first toss is heads, given that \( B \) occurs, is \((2/8)/(3/8) = 2/3\).
Rewriting (1.3) and interchanging the role of $A$ and $B$ gives the relation $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B | A)$. This can be generalized easily to the product rule of probability, which states that for any sequence of events $A_1, A_2, \ldots, A_n$,

$$\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1) \mathbb{P}(A_2 | A_1) \mathbb{P}(A_3 | A_1 A_2) \cdots \mathbb{P}(A_n | A_1 \cdots A_{n-1}), \quad (1.4)$$

using the abbreviation $A_1 A_2 \cdots A_k \equiv A_1 \cap A_2 \cap \cdots \cap A_k$.

Suppose that $B_1, B_2, \ldots, B_n$ is a partition of $\Omega$. That is, $B_1, B_2, \ldots, B_n$ are disjoint and their union is $\Omega$. Then, by the sum rule, $\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \cap B_i)$ and hence, by the definition of conditional probability, we have the law of total probability:

$$\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A | B_i) \mathbb{P}(B_i). \quad (1.5)$$

Combining this with the definition of conditional probability gives Bayes’ rule:

$$\mathbb{P}(B_j | A) = \frac{\mathbb{P}(A | B_j) \mathbb{P}(B_j)}{\sum_{i=1}^n \mathbb{P}(A | B_i) \mathbb{P}(B_i)}. \quad (1.6)$$

Independence is of crucial importance in probability and statistics. Loosely speaking, it models the lack of information between events. Two events $A$ and $B$ are said to be independent if the knowledge that $B$ has occurred does not change the probability that $A$ occurs. That is, $A, B$ independent $\iff \mathbb{P}(A | B) = \mathbb{P}(A)$. Since $\mathbb{P}(A | B) = \mathbb{P}(A \cap B) / \mathbb{P}(B)$, an alternative definition of independence is

$$A, B \text{ independent } \iff \mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B).$$

This definition covers the case where $B = \emptyset$ (empty set). We can extend this definition to arbitrarily many events.

**Definition 1.3.1 (Independence)** The events $A_1, A_2, \ldots,$ are said to be independent if for any $k$ and any choice of distinct indexes $i_1, \ldots, i_k$,

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

**Remark 1.3.1** In most cases independence of events is a model assumption. That is, we assume that there exists a $\mathbb{P}$ such that certain events are independent.

**Example 1.1**

We toss a biased coin $n$ times. Let $p$ be the probability of heads (for a fair coin $p = 1/2$). Let $A_i$ denote the event that the $i$-th toss yields heads, $i = 1, \ldots, n$. Then $\mathbb{P}$ should be such that the events $A_1, \ldots, A_n$ are independent, and $\mathbb{P}(A_i) = p$ for all $i$. 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

$$\mathbb{P}(A_1 \cdots A_k A_{k+1}^c \cdots A_n^c) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_k) \mathbb{P}(A_{k+1}^c) \cdots \mathbb{P}(A_n^c) = p^k (1 - p)^{n-k}.$$
1.4 RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

Specifying a model for a random experiment via a complete description of $\Omega$ and $\mathbb{P}$ may not always be convenient or necessary. In practice, we are only interested in certain observations (i.e., numerical measurements) in the experiment. We incorporate these into our modeling process via the introduction of random variables, usually denoted by capital letters from the last part of the alphabet (e.g., $X, X_1, X_2, \ldots, Y, Z$).

**EXAMPLE 1.2**

We toss a biased coin $n$ times, with $p$ the probability of heads. Suppose that we are interested only in the number of heads, say $X$. Note that $X$ can take any of the values in $\{0, 1, \ldots, n\}$. The probability distribution of $X$ is given by the binomial formula

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n.$$  \hspace{1cm} (1.7)

Namely, by Example 1.1, each elementary event $\{HTH \cdots T\}$ with exactly $k$ heads and $n-k$ tails has probability $p^k (1 - p)^{n-k}$, and there are $\binom{n}{k}$ such events.

The probability distribution of a general random variable $X$ — identifying such probabilities as $\mathbb{P}(X = x), \mathbb{P}(a \leq X \leq b)$, and so on — is completely specified by the cumulative distribution function (cdf), defined by

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$  \hspace{1cm} (1.8)

A random variable $X$ is said to have a discrete distribution if there exists a positive function $f$ with total integral 1, such that for all $a, b$,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(u) \, du.$$  \hspace{1cm} (1.8)

**EXAMPLE 1.3**

Toss two fair dice and let $M$ be the largest face value showing. The pmf of $M$ is given by

<table>
<thead>
<tr>
<th>$m$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>\sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(m)$</td>
<td>$\frac{1}{36}$</td>
<td>$\frac{3}{36}$</td>
<td>$\frac{5}{36}$</td>
<td>$\frac{7}{36}$</td>
<td>$\frac{9}{36}$</td>
<td>$\frac{11}{36}$</td>
<td>1</td>
</tr>
</tbody>
</table>

For example, to get $M = 3$, either $(1, 3), (2, 3), (3, 3), (3, 2)$, or $(3, 1)$ has to be thrown, each of which happens with probability $1/36$.

A random variable $X$ is said to have a continuous distribution if there exists a positive function $f$ with total integral 1, such that for all $a, b$,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(u) \, du.$$  \hspace{1cm} (1.8)
The function $f$ is called the probability density function (pdf) of $X$. Note that in the continuous case the cdf is given by

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u) \, du,$$

and $f$ is the derivative of $F$. We can interpret $f(x)$ as the probability “density” at $X = x$ in the sense that

$$\mathbb{P}(x \leq X \leq x + h) = \int_{x}^{x+h} f(u) \, du \approx h \, f(x).$$

**Remark 1.4.1 (Probability Density)** Note that we have deliberately used the same symbol, $f$, for both pmf and pdf. This is because the pmf and pdf play very similar roles and can, in more advanced probability theory, both be viewed as particular instances of the general notion of probability density. To stress this viewpoint, we will call $f$ in both the discrete and continuous case the pdf or (probability) density (function).

### 1.5 SOME IMPORTANT DISTRIBUTIONS

Tables 1.1 and 1.2 list a number of important continuous and discrete distributions. We will use the notation $X \sim f$, $X \sim F$, or $X \sim \text{Dist}$ to signify that $X$ has a pdf $f$, a cdf $F$ or a distribution Dist. We sometimes write $f_X$ instead of $f$ to stress that the pdf refers to the random variable $X$. Note that in Table 1.1, $\Gamma$ is the gamma function: $\Gamma(\alpha) = \int_{0}^{\infty} e^{-x} x^{\alpha-1} \, dx$, $\alpha > 0$.

**Table 1.1: Commonly used continuous distributions.**

<table>
<thead>
<tr>
<th>Name</th>
<th>Notation</th>
<th>$f(x)$</th>
<th>$x \in$</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>$U[\alpha, \beta]$</td>
<td>$\frac{1}{\beta - \alpha}$</td>
<td>$[\alpha, \beta]$</td>
<td>$\alpha &lt; \beta$</td>
</tr>
<tr>
<td>Normal</td>
<td>$\mathcal{N}(\mu, \sigma^2)$</td>
<td>$\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2}$</td>
<td>$\mathbb{R}$</td>
<td>$\sigma &gt; 0$, $\mu \in \mathbb{R}$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$\Gamma(\alpha, \lambda)$</td>
<td>$\lambda^\alpha x^{\alpha-1} e^{-\lambda x}$</td>
<td>$\mathbb{R}_+$</td>
<td>$\alpha, \lambda &gt; 0$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\text{Exp}(\lambda)$</td>
<td>$\lambda e^{-\lambda x}$</td>
<td>$\mathbb{R}_+$</td>
<td>$\lambda &gt; 0$</td>
</tr>
<tr>
<td>Beta</td>
<td>$\text{Beta}(\alpha, \beta)$</td>
<td>$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1 - x)^{\beta-1}$</td>
<td>$[0, 1]$</td>
<td>$\alpha, \beta &gt; 0$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\text{Weib}(\alpha, \lambda)$</td>
<td>$\alpha \lambda (\lambda x)^{\alpha-1} e^{-\lambda x}$</td>
<td>$\mathbb{R}_+$</td>
<td>$\alpha, \lambda &gt; 0$</td>
</tr>
<tr>
<td>Pareto</td>
<td>$\text{Pareto}(\alpha, \lambda)$</td>
<td>$\alpha \lambda \left(1 + \lambda x\right)^{-(\alpha + 1)}$</td>
<td>$\mathbb{R}_+$</td>
<td>$\alpha, \lambda &gt; 0$</td>
</tr>
</tbody>
</table>
Table 1.2: Commonly used discrete distributions.

<table>
<thead>
<tr>
<th>Name</th>
<th>Notation</th>
<th>( f(x) )</th>
<th>( x \in )</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli</td>
<td>Ber((p))</td>
<td>( p^x(1-p)^{1-x} )</td>
<td>{0,1}</td>
<td>0 ( \leq p \leq 1 )</td>
</tr>
<tr>
<td>Binomial</td>
<td>Bin((n, p))</td>
<td>( \binom{n}{x} p^x(1-p)^{n-x} )</td>
<td>{0, 1, \ldots, n}</td>
<td>0 ( \leq p \leq 1 ), ( n \in \mathbb{N} )</td>
</tr>
<tr>
<td>Discrete uniform</td>
<td>DU{1, \ldots, n}</td>
<td>( \frac{1}{n} )</td>
<td>{1, \ldots, n}</td>
<td>( n \in {1,2,\ldots} )</td>
</tr>
<tr>
<td>Geometric</td>
<td>G((p))</td>
<td>( p(1-p)^{x-1} )</td>
<td>{1,2,\ldots}</td>
<td>0 ( \leq p \leq 1 )</td>
</tr>
<tr>
<td>Poisson</td>
<td>Poi((\lambda))</td>
<td>( e^{-\lambda} \frac{\lambda^x}{x!} )</td>
<td>( \mathbb{N} )</td>
<td>( \lambda &gt; 0 )</td>
</tr>
</tbody>
</table>

1.6 EXPECTATION

It is often useful to consider different kinds of numerical characteristics of a random variable. One such quantity is the expectation, which measures the mean value of the distribution.

**Definition 1.6.1 (Expectation)** Let \( X \) be a random variable with pdf \( f \). The *expectation* (or expected value or mean) of \( X \), denoted by \( \mathbb{E}[X] \) (or sometimes \( \mu \)), is defined by

\[
\mathbb{E}[X] = \begin{cases} 
\sum_{x} x \cdot f(x) & \text{discrete case,} \\
\int_{-\infty}^{\infty} x \cdot f(x) \, dx & \text{continuous case.}
\end{cases}
\]

If \( X \) is a random variable, then a function of \( X \), such as \( X^2 \) or \( \sin(X) \), is again a random variable. Moreover, the expected value of a function of \( X \) is simply a weighted average of the possible values that this function can take. That is, for any real function \( h \)

\[
\mathbb{E}[h(X)] = \begin{cases} 
\sum_{x} h(x) \cdot f(x) & \text{discrete case,} \\
\int_{-\infty}^{\infty} h(x) \cdot f(x) \, dx & \text{continuous case.}
\end{cases}
\]

Another useful quantity is the variance, which measures the spread or dispersion of the distribution.

**Definition 1.6.2 (Variance)** The *variance* of a random variable \( X \), denoted by \( \text{Var}(X) \) (or sometimes \( \sigma^2 \)), is defined by

\[
\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.
\]

The square root of the variance is called the *standard deviation*. Table 1.3 lists the expectations and variances for some well-known distributions.
Table 1.3: Expectations and variances for some well-known distributions.

<table>
<thead>
<tr>
<th>Dist.</th>
<th>( \mathbb{E}[X] )</th>
<th>( \text{Var}(X) )</th>
<th>Dist.</th>
<th>( \mathbb{E}[X] )</th>
<th>( \text{Var}(X) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin((n, p))</td>
<td>( np )</td>
<td>( np(1-p) )</td>
<td>Gamma((\alpha, \lambda))</td>
<td>( \frac{\alpha}{\lambda} )</td>
<td>( \frac{\alpha}{\lambda^2} )</td>
</tr>
<tr>
<td>G((p))</td>
<td>( \frac{1}{p} )</td>
<td>( \frac{1-p}{p^2} )</td>
<td>N((\mu, \sigma^2))</td>
<td>( \mu )</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>Poi((\lambda))</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>Beta((\alpha, \beta))</td>
<td>( \frac{\alpha}{\alpha+\beta} )</td>
<td>( \frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)} )</td>
</tr>
<tr>
<td>U((\alpha, \beta))</td>
<td>( \frac{\alpha + \beta}{2} )</td>
<td>( \frac{(\beta - \alpha)^2}{12} )</td>
<td>Weib((\alpha, \lambda))</td>
<td>( \frac{\Gamma(1/\alpha)}{\alpha \lambda} )</td>
<td>( \frac{2\Gamma(2/\alpha)}{\alpha} - \left( \frac{\Gamma(1/\alpha)}{\alpha \lambda} \right)^2 )</td>
</tr>
</tbody>
</table>

The mean and the variance do not give, in general, enough information to completely specify the distribution of a random variable. However, they may provide useful bounds. We discuss two such bounds. Suppose \( X \) can only take nonnegative values and has pdf \( f \). For any \( x > 0 \), we can write

\[
\mathbb{E}[X] = \int_0^x tf(t) dt + \int_x^\infty tf(t) dt \geq \int_x^\infty tf(t) dt \geq \int_x^\infty xf(t) dt = x \mathbb{P}(X \geq x),
\]

from which follows the Markov inequality: if \( X \geq 0 \), then for all \( x > 0 \),

\[
\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[X]}{x} . \tag{1.9}
\]

If we also know the variance of a random variable, we can give a tighter bound. Namely, for any random variable \( X \) with mean \( \mu \) and variance \( \sigma^2 \), we have

\[
\mathbb{P}(|X - \mu| \geq x) \leq \frac{\sigma^2}{x^2} . \tag{1.10}
\]

This is called the Chebyshev inequality. The proof is as follows: Let \( D^2 = (X - \mu)^2 \); then, by the Markov inequality (1.9) and the definition of the variance,

\[
\mathbb{P}(D^2 \geq x^2) \leq \frac{\sigma^2}{x^2} .
\]

Also, note that the event \( \{D^2 \geq x^2\} \) is equivalent to the event \( \{|X - \mu| \geq x\} \), so that (1.10) follows.

### 1.7 Joint Distributions

Often a random experiment is described by more than one random variable. The theory for multiple random variables is similar to that for a single random variable.
Let $X_1, \ldots, X_n$ be random variables describing some random experiment. We can accumulate these into a random vector $\mathbf{X} = (X_1, \ldots, X_n)$. More generally, a collection $\{X_t, t \in T\}$ of random variables is called a stochastic process. The set $T$ is called the parameter set or index set of the process. It may be discrete (e.g., $\mathbb{N}$ or $\{1, \ldots, 10\}$) or continuous (e.g., $\mathbb{R}_+ = [0, \infty)$ or $[1, 10]$). The set of possible values for the stochastic process is called the state space.

The joint distribution of $X_1, \ldots, X_n$ is specified by the joint cdf
\[ F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n). \]

The joint pdf $f$ is given, in the discrete case, by $f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$, and in the continuous case $f$ is such that $\mathbb{P}(X \in B) = \int_B f(x_1, \ldots, x_n) \, dx_1 \ldots dx_n$ for any (measurable) region $B$ in $\mathbb{R}^n$. The marginal pdfs can be recovered from the joint pdf by integration or summation. For example, in the case of a continuous random vector $(X, Y)$ with joint pdf $f$, the pdf $f_X$ of $X$ is found as
\[ f_X(x) = \int f(x, y) \, dy. \]

Suppose that $X$ and $Y$ are both discrete or both continuous, with joint pdf $f$, and suppose that $f_X(x) > 0$. Then the conditional pdf of $Y$ given $X = x$ is given by
\[ f_Y|X(y \mid x) = \frac{f(x, y)}{f_X(x)} \quad \text{for all } y. \]

The corresponding conditional expectation is (in the continuous case)
\[ \mathbb{E}[Y \mid X = x] = \int y f_Y|X(y \mid x) \, dy. \]

Note that $\mathbb{E}[Y \mid X = x]$ is a function of $x$, say $h(x)$. The corresponding random variable $h(X)$ is written as $\mathbb{E}[Y \mid X]$. It can be shown (see, for example, [3]) that its expectation is simply the expectation of $Y$, that is,
\[ \mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]. \quad (1.11) \]

When the conditional distribution of $Y$ given $X$ is identical to that of $Y$, $X$ and $Y$ are said to be independent. More precisely:

**Definition 1.7.1 (Independent Random Variables)** The random variables $X_1, \ldots, X_n$ are called independent if for all events $\{X_i \in A_i\}$ with $A_i \subset \mathbb{R}$, $i = 1, \ldots, n,$
\[ \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n). \]

A direct consequence of the definition above for independence is that random variables $X_1, \ldots, X_n$ with joint pdf $f$ (discrete or continuous) are independent if and only if
\[ f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \quad (1.12) \]
for all $x_1, \ldots, x_n$, where $\{f_{X_i}\}$ are the marginal pdfs.
EXAMPLE 1.4 Bernoulli Sequence

Consider the experiment where we flip 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

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

Last, assume that \( X_1, \ldots, X_n \) are independent. The sequence \( \{ X_i, i = 1, 2, \ldots \} \) is called a Bernoulli sequence or 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 \( \mathcal{B} \) the set of all binary vectors \( x = (x_1, \ldots, x_n) \) such that \( \sum_{i=1}^n x_i = k \). Note that \( \mathcal{B} \) has \( \binom{n}{k} \) elements. We now have

\[
\mathbb{P}(X = k) = \sum_{x \in \mathcal{B}} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \sum_{x \in \mathcal{B}} p^k (1 - p)^{n-k} = \binom{n}{k} p^k (1 - p)^{n-k}.
\]

In other words, \( X \sim \text{Bin}(n, p) \). Compare this with Example 1.2.

**Remark 1.7.1** An infinite sequence \( X_1, X_2, \ldots \) of random variables is called independent if for any finite choice of parameters \( i_1, i_2, \ldots, i_n \) (none of them the same) the random variables \( X_{i_1}, \ldots, X_{i_n} \) are independent. Many probabilistic 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.

Similar to the one-dimensional case, the expected value of any real-valued function \( h \) of \( X_1, \ldots, X_n \) is a weighted average of all values that this function can take. Specifically, in the continuous case,

\[
\mathbb{E}[h(X_1, \ldots, X_n)] = \int \cdots \int h(x_1, \ldots, x_n) f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n.
\]

As a direct consequence of the definitions of expectation and independence, we have

\[
\mathbb{E}[a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n] = a + b_1 \mu_1 + \cdots + b_n \mu_n
\]

(1.13)

for any sequence of random variables \( X_1, X_2, \ldots, X_n \) with expectations \( \mu_1, \mu_2, \ldots, \mu_n \), where \( a, b_1, b_2, \ldots, b_n \) are constants. Similarly, for independent random variables, we have

\[
\mathbb{E}[X_1 X_2 \cdots X_n] = \mu_1 \mu_2 \cdots \mu_n.
\]

The covariance of two random variables \( X \) and \( Y \) with expectations \( \mathbb{E}[X] = \mu_X \) and \( \mathbb{E}[Y] = \mu_Y \), respectively, is defined as

\[
\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].
\]
This is a measure for the amount of linear dependency between the variables. A scaled version of the covariance is given by the correlation coefficient,

\[ \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}, \]

where \( \sigma_X^2 = \text{Var}(X) \) and \( \sigma_Y^2 = \text{Var}(Y) \). It can be shown that the correlation coefficient always lies between \(-1\) and \(1\); see Problem 1.13.

For easy reference, Table 1.4 lists some important properties of the variance and covariance. The proofs follow directly from the definitions of covariance and variance and the properties of the expectation.

### Table 1.4: Properties of variance and covariance.

<table>
<thead>
<tr>
<th></th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 )</td>
</tr>
<tr>
<td>2</td>
<td>( \text{Var}(aX + b) = a^2 \text{Var}(X) )</td>
</tr>
<tr>
<td>3</td>
<td>( \text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] )</td>
</tr>
<tr>
<td>4</td>
<td>( \text{Cov}(X, Y) = \text{Cov}(Y, X) )</td>
</tr>
<tr>
<td>5</td>
<td>( \text{Cov}(aX + bY, Z) = a \text{Cov}(X, Z) + b \text{Cov}(Y, Z) )</td>
</tr>
<tr>
<td>6</td>
<td>( \text{Cov}(X, X) = \text{Var}(X) )</td>
</tr>
<tr>
<td>7</td>
<td>( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) )</td>
</tr>
<tr>
<td>8</td>
<td>( X ) and ( Y ) independent. ( \implies \text{Cov}(X, Y) = 0 )</td>
</tr>
</tbody>
</table>

As a consequence of properties 2 and 7, for any sequence of \textit{independent} random variables \( X_1, \ldots, X_n \) with variances \( \sigma_1^2, \ldots, \sigma_n^2 \),

\[
\text{Var}(a + b_1X_1 + b_2X_2 + \cdots + b_nX_n) = b_1^2 \sigma_1^2 + \cdots + b_n^2 \sigma_n^2 \tag{1.14}
\]

for any choice of constants \( a \) and \( b_1, \ldots, b_n \).

For random vectors, such as \( \mathbf{X} = (X_1, \ldots, X_n) \top \), it is convenient to write the expectations and covariances in vector notation.

**Definition 1.7.2 (Expectation Vector and Covariance Matrix)** For any random vector \( \mathbf{X} \), we define the \textit{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 \textit{covariance matrix} \( \Sigma \) 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 vector (matrix) to be the vector (matrix) of expectations, then we can write

\[ \boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] \]