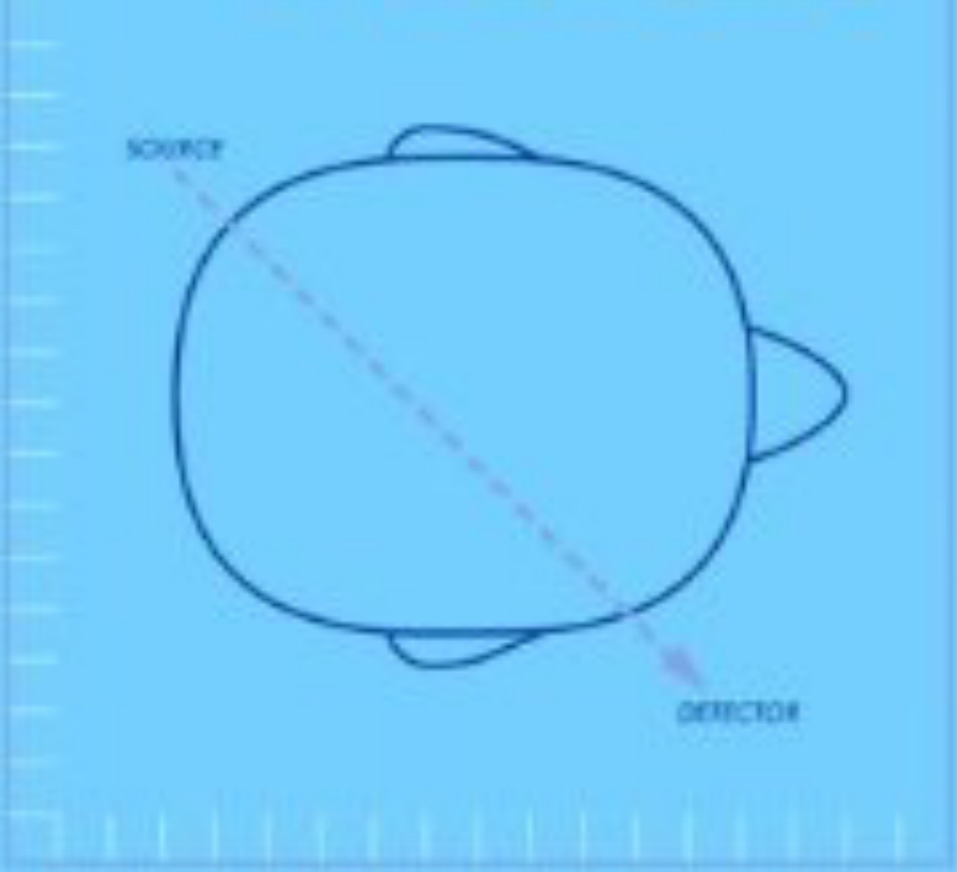
SPRINGER TEXTS IN STATISTICS

Applied Probability



Kenneth Lange

Springer

Kenneth Lange

Applied Probability

Second Edition

Springer

Prof. Kenneth Lange
University of California, Los Angeles
Departments of Biomathematics,
Human Genetics, and Statistics
90095-1766 Los Angeles
California
USA
klange@ucla.edu

ISSN 1431-875X
ISBN 978-1-4419-7164-7 e-ISBN 978-1-4419-7165-4
DOI 10.1007/978-1-4419-7165-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010933109

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface to the Second Edition

My original intent in writing *Applied Probability* was to strike a balance between theory and applications. Theory divorced from applications runs the risk of alienating many potential practitioners of the art of stochastic modeling. Applications without a clear statement of relevant theory drift in a sea of confusion. To a lesser degree I was also motivated by a desire to promote the nascent field of computational probability. Current students of the mathematical sciences are more computer savvy than ever. Putting the right computational tools in their hands is bound to advance probability and the broader good of science.

The second edition of *Applied Probability* remains true to these aims. I have added two new chapters on asymptotic and numerical methods and an appendix that separates some of the more delicate mathematical theory from the steady flow of examples in the main text. In addition to these major changes, there is now a much more extensive list of exercises. Some of these are trivial, but others will challenge even the best students. Finally, many errors, both large and small, have been corrected.

Chapter 4 on combinatorics includes new sections on bijections, Catalan numbers, and Faà di Bruno's formula. The proof of the inclusion-exclusion formula has been clarified. Chapter 7 on Markov chains contains new material on rates of convergence to equilibrium in reversible finite-state chains. This discussion draws on students' previous exposure to eigenvalues and eigenvectors in linear algebra. Chapter 9 on branching processes features a new section on basic reproduction numbers. Here the idea is to devise easy algebraic tests for deciding when a process is subcritical, critical, or supercritical. Chapter 11 on diffusion processes gives better coverage of

Brownian motion. The last two sections of the chapter have been moved to the new Chapter 13 on numerical methods. The orphan material on convergent sequences of random variables in Chapter 1 has been moved to the new Chapter 12 on asymptotic methods.

Once again I would like to thank the students of my UCLA biomathematics classes for their help. Particularly noteworthy are David Alexander, Kristin Ayers, Forrest Crawford, Kate Crespi, Gabriela Cybis, Lewis Lee, Sarah Nowak, John Ranola, Mary Sehl, Tongtong Wu, and Jin Zhou. I owe an especially heavy debt to Hua Zhou, my former postdoctoral fellow, for suggesting many problems and lecturing in my absence. I also thank my editor, John Kimmel, for his kind support. Finally, I am glad to report that my mother, to whom both editions of this book are dedicated, is alive and well. If I can spread even a fraction of the cheer she has spread, then I will be able to look back over a life well lived.

Preface

Despite the fears of university mathematics departments, mathematics education is growing rather than declining. But the truth of the matter is that the increases are occurring outside departments of mathematics. Engineers, computer scientists, physicists, chemists, economists, statisticians, biologists, and even philosophers teach and learn a great deal of mathematics. The teaching is not always terribly rigorous, but it tends to be better motivated and better adapted to the needs of students. In my own experience teaching students of biostatistics and mathematical biology, I attempt to convey both the beauty and utility of probability. This is a tall order, partially because probability theory has its own vocabulary and habits of thought. The axiomatic presentation of advanced probability typically proceeds via measure theory. This approach has the advantage of rigor, but it inevitably misses most of the interesting applications, and many applied scientists rebel against the onslaught of technicalities. In the current book, I endeavor to achieve a balance between theory and applications in a rather short compass. While the combination of brevity and balance sacrifices many of the proofs of a rigorous course, it is still consistent with supplying students with many of the relevant theoretical tools. In my opinion, it is better to present the mathematical facts without proof rather than omit them altogether.

In the preface to his lovely recent textbook [209], David Williams writes, "Probability and Statistics used to be married; then they separated; then they got divorced; now they hardly see each other." Although this split is doubtless irreversible, at least we ought to be concerned with properly bringing up their children, applied probability and computational statistics.

If we fail, then science as a whole will suffer. You see before you my attempt to give applied probability the attention it deserves. My other recent book [122] covers computational statistics and aspects of computational probability glossed over here.

This graduate-level textbook presupposes knowledge of multivariate calculus, linear algebra, and ordinary differential equations. In probability theory, students should be comfortable with elementary combinatorics, generating functions, probability densities and distributions, expectations, and conditioning arguments. My intended audience includes graduate students in applied mathematics, biostatistics, computational biology, computer science, physics, and statistics. Because of the diversity of needs, instructors are encouraged to exercise their own judgment in deciding what chapters and topics to cover.

Chapter 1 reviews elementary probability while striving to give a brief survey of relevant results from measure theory. Poorly prepared students should supplement this material with outside reading. Well-prepared students can skim Chapter 1 until they reach the less well-known material of the final two sections. Section 1.8 develops properties of the multivariate normal distribution of special interest to students in biostatistics and statistics. This material is applied to optimization theory in Section 3.3 and to diffusion processes in Chapter 11.

We get down to serious business in Chapter 2, which is an extended essay on calculating expectations. Students often complain that probability is nothing more than a bag of tricks. For better or worse, they are confronted here with some of those tricks. Readers may want to skip the final two sections of the chapter on surface area distributions on a first pass through the book.

Chapter 3 touches on advanced topics from convexity, inequalities, and optimization. Besides the obvious applications to computational statistics, part of the motivation for this material is its applicability in calculating bounds on probabilities and moments.

Combinatorics has the odd reputation of being difficult in spite of relying on elementary methods. Chapters 4 and 5 are my stab at making the subject accessible and interesting. There is no doubt in my mind of combinatorics' practical importance. More and more we live in a world dominated by discrete bits of information. The stress on algorithms in Chapter 5 is intended to appeal to computer scientists.

Chapters 6 through 11 cover core material on stochastic processes that I have taught to students in mathematical biology over a span of many years. If supplemented with appropriate sections from Chapters 1 and 2, there is sufficient material here for a traditional semester-long course in stochastic processes. Although my examples are weighted toward biology, particularly genetics, I have tried to achieve variety. The fortunes of this book doubtless will hinge on how compelling readers find these examples.

You can leaf through the table of contents to get a better idea of the topics covered in these chapters.

In the final two chapters, on Poisson approximation and number theory, the applications of probability to other branches of mathematics come to the fore. These chapters are hardly in the mainstream of stochastic processes and are meant for independent reading as much as for classroom presentation.

All chapters come with exercises. (In this second printing, some additional exercises are included at the end of the book.) These are not graded by difficulty, but hints are provided for some of the more difficult ones. My own practice is to require one problem for each hour and a half of lecture. Students are allowed to choose among the problems within each chapter and are graded on the best of the solutions they present. This strategy provides an incentive for the students to attempt more than the minimum number of problems.

I would like to thank my former and current UCLA and University of Michigan students for their help in debugging this text. In retrospect, there were far more contributing students than I can possibly credit. At the risk of offending the many, let me single out Brian Dolan, Ruzong Fan, David Hunter, Wei-hsun Liao, Ben Redelings, Eric Schadt, Marc Suchard, Janet Sinsheimer, and Andy Ming-Ham Yip. I also thank John Kimmel of Springer–Verlag for his editorial assistance.

Finally, I dedicate this book to my mother, Alma Lange, on the occasion of her 80th birthday. Thanks, Mom, for your cheerfulness and generosity in raising me. You were, and always will be, an inspiration to the whole family.

Contents

1

Basic Notions of Probability Theory

1.1 Introduction

This initial chapter covers background material that every serious student of applied probability should master. In no sense is the chapter meant as a substitute for a previous course in applied probability or for a future course in measure-theoretic probability. Our comments are merely meant as reminders and as a bridge. Many mathematical facts will be stated without proof. This is unsatisfactory, but it is even more unsatisfactory to deny students the most powerful tools in the probabilist's toolkit. Quite apart from specific tools, the language and intellectual perspective of modern probability theory also furnish an intuitive setting for solving practical problems. Probability involves modes of thought that are unique within mathematics. As a brief illustration of the material reviewed, we derive properties of the multivariate normal distribution in the final section of this chapter. Later chapters will build on the facts and vocabulary mentioned here and provide more elaborate applications.

1.2 Probability and Expectation

The layman's definition of probability is the long-run frequency of success over a sequence of independent, identically constructed trials. Although this law of large numbers perspective is important, mathematicians have found it helpful to put probability theory on an axiomatic basis [24, 53, 60, 80,

K. Lange, *Applied Probability,*
DOI 10.1007/978-1-4419-7165-4_1, © Springer Science+Business Media, LLC 2010

166, 171, 208]. The modern theory begins with the notion of a sample space Ω and a collection $\mathcal{F}$ of subsets from Ω subject to the following conventions:

(1.2a) $\Omega \in \mathcal{F}$.

(1.2b) If $A \in \mathcal{F}$, then its complement $A^c \in \mathcal{F}$.

(1.2c) If $A_1, A_2, \ldots$ is a finite or countably infinite sequence of subsets from $\mathcal{F}$, then $\bigcup_i A_i \in \mathcal{F}$.

Any collection $\mathcal{F}$ satisfying these postulates is termed a σ-field or σ-algebra. Two immediate consequences of the definitions are that the empty set $\emptyset \in \mathcal{F}$ and that if $A_1, A_2, \ldots$ is a finite or countably infinite sequence of subsets from $\mathcal{F}$, then $\bigcap_i A_i = (\bigcup_i A_i^c)^c \in \mathcal{F}$. In probability theory, we usually substitute everyday language for set theory language. Table 1.1 provides a short dictionary for translating equivalent terms.

TABLE 1.1. A Brief Dictionary of Set Theory and Probability Terms

Set Theory	**Probability**	**Set Theory**	**Probability**
set	event	null set	impossible event
union	or	universal set	certain event
intersection	and	pairwise disjoint	mutually exclusive
complement	not	inclusion	implication

The axiomatic setting of probability theory is completed by introducing a probability measure or distribution Pr on the events in $\mathcal{F}$. This function should satisfy the properties:

(1.2d) $\Pr(\Omega) = 1$.

(1.2e) $\Pr(A) \geq 0$ for any $A \in \mathcal{F}$.

(1.2f) $\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$ for any countably infinite sequence of mutually exclusive events $A_1, A_2, \ldots$ from $\mathcal{F}$.

A triple $(\Omega, \mathcal{F}, \Pr)$ constitutes a probability space. An event $A \in \mathcal{F}$ is said to be null when $\Pr(A) = 0$ and almost sure when $\Pr(A) = 1$.

Example 1.2.1 *Discrete Uniform Distribution*

One particularly simple sample space is the set $\Omega = \{1, \ldots, n\}$. Here the natural choice of $\mathcal{F}$ is the collection of all subsets of Ω. The uniform distribution (or normalized counting measure) attributes probability $\Pr(A) = \frac{|A|}{n}$ to a set A, where $|A|$ denotes the number of elements of A. Most of the counting arguments of combinatorics presuppose the discrete uniform distribution. ■

Example 1.2.2 *Continuous Uniform Distribution*

A continuous analog of the discrete uniform distribution is furnished by Lebesgue measure on the unit interval $[0,1]$. In this case, the best one can do is define $\mathcal{F}$ as the smallest σ-algebra containing all closed subintervals $[a,b]$ of $\Omega = [0,1]$. The events in $\mathcal{F}$ are then said to be Borel sets. Henri Lebesgue was able to show how to extend the primitive identification $\Pr([a,b]) = b - a$ of the probability of an interval with its length to all Borel sets [171]. Invoking the axiom of choice from set theory, one can prove that it is impossible to attach a probability consistently to all subsets of $[0,1]$. The existence of nonmeasurable sets makes the whole enterprise of measure-theoretic probability more delicate than mathematicians anticipated. Fortunately, one can ignore such subtleties in most practical problems. ■

The next example is designed to give readers a hint of the complexities involved in defining probability spaces.

Example 1.2.3 *Density in Number Theory*

Consider the natural numbers $\Omega = \{1, 2, \ldots\}$ equipped with the density function

$$\mathrm{den}(A) \quad = \quad \lim_{n \to \infty} \frac{|A \cap \{1, 2, \ldots, n\}|}{n}.$$

Clearly, $0 \le \mathrm{den}(A) \le 1$ whenever $\mathrm{den}(A)$ is defined. Some typical densities include $\mathrm{den}(\Omega) = 1$, $\mathrm{den}(\{j\}) = 0$, and $\mathrm{den}(\{j, 2j, 3j, 4j, \ldots\}) = 1/j$. Any σ-algebra $\mathcal{F}$ containing each of the positive integers $\{j\}$ fails the test of countable additivity stated in postulate (1.2f) above. Indeed,

$$\mathrm{den}(\Omega) \quad \neq \quad 0 \quad = \quad \sum_{j=1}^{\infty} \mathrm{den}(\{j\}).$$

Note that $\mathrm{den}(A)$ does satisfy the test of finite additivity. Of course, it is possible to define many legitimate probability distributions on the positive integers. ■

In practice, most questions in probability theory revolve around random variables rather than sample spaces. Readers will doubtless recall that a random variable X is a function from a sample space Ω into the real line R. This is almost correct. To construct a consistent theory of integration, one must insist that a random variable be measurable. This technical condition requires that for every constant c, the set $\{\omega \in \Omega : X(\omega) \le c\}$ be an event in the σ-algebra $\mathcal{F}$ attached to Ω. Measurability can also be defined in terms of the Borel sets $\mathcal{B}$ of R, which comprise the smallest σ-algebra containing all intervals $[a,b]$ of R. With this definition in mind, X is measurable if and only if the inverse image $X^{-1}(B)$ of every Borel set B is an event in $\mathcal{F}$. This is analogous to but weaker than defining continuity by requiring that the inverse image of every open set be open. Almost every conceivable function $X : \Omega \mapsto \mathsf{R}$ qualifies as measurable. Formal verification of measurability usually invokes one or more of the many closure properties of measurable functions. For instance, measurability is preserved under the formation of finite sums, products, maxima, minima, and limits of measurable functions. For this reason, we seldom waste time checking measurability.

Measurable functions are candidates for integration. The simplest measurable function is the indicator 1_A of an event A. The integral or expectation $\mathrm{E}(1_A)$ of 1_A is just the corresponding probability $\Pr(A)$. Integration is first extended to simple functions $\sum_{i=1}^n c_i 1_{A_i}$ by the linearity device

$$\begin{aligned} \mathrm{E}\left(\sum_{i=1}^n c_i 1_{A_i}\right) &= \sum_{i=1}^n c_i \,\mathrm{E}(1_{A_i}) \\ &= \sum_{i=1}^n c_i \Pr(A_i) \end{aligned}$$

and from there to the larger class of integrable functions by appropriate limit arguments. Although the rigorous development of integration is one of the intellectual triumphs of modern mathematics, we record here only some of the basic facts. The two most important are linearity and nonnegativity:

(1.2g) $\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$.

(1.2h) $\mathrm{E}(X) \geq 0$ for any $X \geq 0$.

From these basic properties, a host of simple results flow. As one example, the inequality $|\,\mathrm{E}(X)| \leq \mathrm{E}(|X|)$ holds whenever $\mathrm{E}(|X|) < \infty$. As another example, taking expectations in the identity $1_{A\cup B} = 1_A + 1_B - 1_{A\cap B}$ produces the identity $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. Of course, one can prove this and similar equalities without introducing expectations, but the application of the expectation operator often streamlines proofs.

One of the most impressive achievements of Lebesgue's theory of integration is that it identifies sufficient conditions for the interchange of limits and integrals. Fatou's lemma states that

$$\mathrm{E}\left(\liminf_{n\to\infty} X_n\right) \leq \liminf_{n\to\infty} \mathrm{E}(X_n)$$

for any sequence $X_1, X_2, \ldots$ of nonnegative random variables. Recall that $\liminf_{n\to\infty} a_n = \lim_{n\to\infty} \inf\{a_k\}_{k\geq n}$ for any sequence a_n. In the present case, each sample point ω defines a different sequence $a_n = X_n(\omega)$.

If the sequence of random variables X_n is increasing as well as nonnegative, then the monotone convergence theorem

$$\lim_{n\to\infty} \mathrm{E}(X_n) = \mathrm{E}\left(\lim_{n\to\infty} X_n\right) \tag{1.1}$$

holds, with the possibility $\mathrm{E}(\lim_{n\to\infty} X_n) = \infty$ included. Again we need look no further than indicator functions to apply the monotone convergence theorem. Suppose $A_1 \subset A_2 \subset \cdots$ is an increasing sequence of events with limit $A_\infty = \cup_{n=1}^\infty A_n$. Then the continuity property

$$\lim_{n\to\infty} \Pr(A_n) = \Pr(A_\infty)$$

follows trivially from the monotone convergence theorem. Experts might rightfully object that this is circular reasoning because the continuity of probability is one of the ingredients that goes into constructing a rigorous theory of integration in the first place. However, this misses the psychological point that it is easier to remember and apply a general theorem than various special cases of it.

Example 1.2.4 *Borel-Cantelli Lemma*

Suppose a sequence of events $A_1, A_2, \ldots$ satisfies $\sum_{i=1}^\infty \Pr(A_i) < \infty$. The Borel-Cantelli lemma says only finitely many of the events occur. To prove this result, let 1_{A_i} be the indicator function of A_i, and let N be the infinite sum $\sum_{i=1}^\infty 1_{A_i}$. The monotone convergence theorem implies that

$$\mathrm{E}(N) = \sum_{i=1}^\infty \Pr(A_i).$$

If $\mathrm{E}(N) < \infty$ as assumed, then $N < \infty$ with probability 1. In other words, only finitely many of the A_i occur. ■

The dominated convergence theorem relaxes the assumptions that the sequence $X_1, X_2, \ldots$ is monotone and nonnegative but adds the requirement that all X_n satisfy $|X_n| \le Y$ for some dominating random variable Y with finite expectation. Assuming that $\lim_{n\to\infty} X_n$ exists, the interchange (1.1) is again permissible. If the dominating random variable Y is constant, then most probabilists refer to the dominated convergence theorem as the bounded convergence theorem. Our next example illustrates the power of the dominated convergence theorem.

Example 1.2.5 *Differentiation Under an Expectation Sign*

Let X_t denote a family of random variables indexed by a real parameter t such that (a) $\frac{d}{dt}X_t(\omega)$ exists for all sample points ω and (b) $|\frac{d}{dt}X_t| \le Y$ for some dominating random variable Y with finite expectation. We claim that $\frac{d}{dt}\mathrm{E}(X_t)$ exists and equals $\mathrm{E}(\frac{d}{dt}X_t)$. To prove this result, consider the difference quotient

$$\frac{\mathrm{E}(X_{t+\Delta t}) - \mathrm{E}(X_t)}{\Delta t} = \mathrm{E}\left(\frac{X_{t+\Delta t} - X_t}{\Delta t}\right).$$

For any sample point ω, the mean value theorem implies that

$$\begin{aligned}\left|\frac{X_{t+\Delta t}(\omega) - X_t(\omega)}{\Delta t}\right| &= \left|\frac{d}{ds}X_s(\omega)\right| \\ &\leq Y(\omega)\end{aligned}$$

for some s between t and $t + \Delta t$. Because the difference quotients converge to the derivative in a dominated fashion as Δt tends to 0, application of the dominated convergence theorem finishes the proof.

As a straightforward illustration, consider the problem of calculating the first moment of a random variable Z from its characteristic function $\mathrm{E}(e^{itZ})$. Assuming that $\mathrm{E}(|Z|)$ is finite, define the family of random variables $X_t = e^{itZ}$. It is then clear that the derivative $\frac{d}{dt}X_t(\omega) = iZ(\omega)e^{itZ(\omega)}$ exists for all sample points ω and that $Y = |iZe^{itZ}| = |Z|$ furnishes an appropriate dominating random variable. Hence, $\mathrm{E}(Z)$ equals the value of $-i\frac{d}{dt}\,\mathrm{E}(e^{itZ})$ at $t = 0$. ■

1.3 Conditional Probability

Constructing a rigorous theory of conditional probability and conditional expectation is as much a chore as constructing a rigorous theory of integration. Fortunately, most of the theoretic results can be motivated starting with the simple case of conditioning on an event of positive probability. In this case, we define the conditional probability

$$\Pr(B \mid A) = \frac{\Pr(B \cap A)}{\Pr(A)}$$

of any event B relative to A. Because the conditional probability $\Pr(B \mid A)$ is a legitimate probability measure, it is possible to define the conditional expectation $E(Z \mid A)$ of any integrable random variable Z. Fortunately, this boils down to nothing more than

$$\mathrm{E}(Z \mid A) = \frac{\mathrm{E}(Z1_A)}{\Pr(A)}. \tag{1.2}$$

Definition (1.2) has limited scope, and probabilists have generalized it by conditioning on a random variable rather than a single event. If X is a random variable taking only a finite number of values $x_1, \ldots, x_n$, then $\mathrm{E}(Z \mid X)$ is the random variable defined by $\mathrm{E}(Z \mid X = x_i)$ on the event $\{X = x_i\}$. Obviously, the conditional expectation operator inherits the properties of linearity and nonnegativity in Z from the ordinary expectation operator. In addition, there is the further connection

$$\begin{aligned}\mathrm{E}(Z) &= \sum_{i=1}^{n} \mathrm{E}(Z \mid X = x_i)\Pr(X = x_i) \\ &= \mathrm{E}[\mathrm{E}(Z \mid X)] \qquad (1.3)\end{aligned}$$

between ordinary and conditional expectations. The final property worth highlighting,

$$\mathrm{E}[f(X)Z] \quad = \quad \mathrm{E}[f(X)\,\mathrm{E}(Z \mid X)], \tag{1.4}$$

is a consequence of equation (1.3) and the obvious identity

$$\mathrm{E}[f(X)Z \mid X] \quad = \quad f(X)\,\mathrm{E}(Z \mid X).$$

Example 1.3.1 *The Hypergeometric Distribution*

Consider a finite sequence $X_1, \ldots, X_n$ of independent Bernoulli random variables with common success probability p. Here $\Pr(X_j = 1) = p$ and $\Pr(X_j = 0) = 1 - p$, and the sum $S_n = X_1 + \cdots + X_n$ follows a binomial distribution. The hypergeometric distribution can be recovered in this setting by conditioning. For $m < n$, define the shorter sum $S_m = X_1 + \cdots + X_m$ and calculate

$$\begin{aligned}
\Pr(S_m = j \mid S_n = k) \quad &= \quad \frac{\binom{m}{j} p^j (1-p)^{m-j} \binom{n-m}{k-j} p^{k-j} (1-p)^{n-m+j-k}}{\binom{n}{k} p^k (1-p)^{n-k}} \\
&= \quad \frac{\binom{m}{j}\binom{n-m}{k-j}}{\binom{n}{k}}.
\end{aligned}$$

The mean of this hypergeometric distribution is just the conditional expectation $\mathrm{E}(S_m \mid S_n = k)$. Using symmetry and the additivity of the conditional expectation operator, we find that

$$\begin{aligned}
\mathrm{E}(S_m \mid S_n = k) \quad &= \quad \sum_{i=1}^{m} \mathrm{E}(X_i \mid S_n = k) \\
&= \quad m\,\mathrm{E}(X_1 \mid S_n = k) \\
&= \quad \frac{m}{n}\,\mathrm{E}(S_n \mid S_n = k) \\
&= \quad \frac{mk}{n}.
\end{aligned}$$

It is noteworthy that the identity $\mathrm{E}(S_m \mid S_n) = \frac{m}{n} S_n$ does not require the X_j to be Bernoulli. ■

At the highest level of abstraction, we define conditional expectation $\mathrm{E}(Z \mid \mathcal{G})$ relative to a sub-σ-algebra $\mathcal{G}$ of the underlying σ-algebra $\mathcal{F}$. Here it is important to bear in mind that Z must be integrable and that in most cases $\mathcal{G}$ is the smallest σ-algebra making a random variable X or a random vector $(X_1, \ldots, X_n)$ measurable. The technical requirement that $\mathrm{E}(Z \mid \mathcal{G})$ be measurable with respect to $\mathcal{G}$ then means that $\mathrm{E}(Z \mid \mathcal{G})$ is a function of X or $(X_1, \ldots, X_n)$. Because $\mathcal{G}$ may have an infinity of events, we can no longer rely on defining $\mathrm{E}(Z \mid \mathcal{G})$ by naively conditioning on events of positive probability. The usual mathematical trick of turning a theorem into a definition, however, comes to the rescue. Thus, $\mathrm{E}(Z \mid \mathcal{G})$ is defined as the essentially unique integrable random variable that is measurable with respect to $\mathcal{G}$ and satisfies the analog

$$\mathrm{E}[1_C Z] \quad = \quad \mathrm{E}[1_C \, \mathrm{E}(Z \mid \mathcal{G})] \tag{1.5}$$

of equation (1.4) for every event C in $\mathcal{G}$. Hidden in this definition is an appeal to the powerful Radon-Nikodym theorem of measure theory. The upshot of these indirect arguments is that the conditional expectation operator is perfectly respectable and continues to enjoy the basic properties mentioned earlier.

In our study of martingales in Chapter 10, we will encounter increasing σ-algebras. We write $\mathcal{F} \subset \mathcal{G}$ if every event of $\mathcal{F}$ is also an event $\mathcal{G}$. In other words, $\mathcal{F}$ is less informative than $\mathcal{G}$. The "tower property"

$$\mathrm{E}[\mathrm{E}(Z \mid \mathcal{G}) \mid \mathcal{F}] \quad = \quad \mathrm{E}(Z \mid \mathcal{F}) \tag{1.6}$$

holds in this case because equation (1.5) implies

$$\begin{aligned} \mathrm{E}[1_C \, \mathrm{E}(Z \mid \mathcal{G})] &= \mathrm{E}[\mathrm{E}(1_C Z \mid \mathcal{G})] \\ &= \mathrm{E}(1_C Z) \\ &= \mathrm{E}[1_C \, \mathrm{E}(Z \mid \mathcal{F})] \end{aligned}$$

for every C in $\mathcal{F}$.

1.4 Independence

Two events A and B are independent if and only if

$$\Pr(A \cap B) \quad = \quad \Pr(A) \Pr(B).$$

This definition is equivalent to $\Pr(B \mid A) = \Pr(B)$ when $\Pr(A) > 0$. A finite or countable sequence $A_1, A_2, \ldots$ of events is independent provided

$$\Pr\left(\bigcap_{j=1}^{n} A_{i_j}\right) \quad = \quad \prod_{j=1}^{n} \Pr(A_{i_j})$$

for all finite subsequences $A_{i_1}, \ldots, A_{i_n}$. Pairwise independence is insufficient to imply independence. A sequence of random variables $X_1, X_2, \ldots$ is independent whenever the sequence of events $A_i = \{X_i \leq c_i\}$ is independent for all possible choices of the constants c_i. In practice, one usually establishes the independence of two random variables U and V by exhibiting them as measurable functions $U = f(X)$ and $V = g(Y)$ of known independent random variables X and Y.