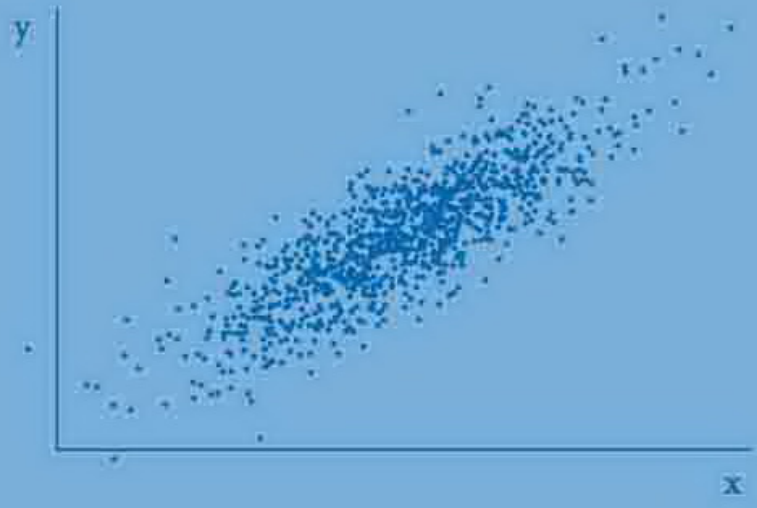
Springer Texts in Statistics

Anirban DasGupta

Fundamentals of Probability: A First Course



Springer

Springer Texts in Statistics

Series Editors:
G. Casella
S. Fienberg
I. Olkin

For other titles published in this series, go to
http://www.springer.com/series/417

Anirban DasGupta

Fundamentals of Probability: A First Course

Springer

Anirban DasGupta
Purdue University
Dept. Statistics & Mathematics
150 N. University Street
West Lafayette IN 47907
USA
dasgupta@stat.purdue.edu

ISSN 1431-875X
ISBN 978-1-4419-5779-5 e-ISBN 978-1-4419-5780-1
DOI 10.1007/978-1-4419-5780-1
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010924739

To the Memory of William Feller, whose books inspired my love of probability, and to Dev Basu, the greatest teacher I have known

Preface

Probability theory is one branch of mathematics that is simultaneously deep and immediately applicable in diverse areas of human endeavor. It is as fundamental as calculus. Calculus explains the external world, and probability theory helps predict a lot of it. In addition, problems in probability theory have an innate appeal, and the answers are often structured and strikingly beautiful. A solid background in probability theory and probability models will become increasingly more useful in the twenty-first century, as difficult new problems emerge, that will require more sophisticated models and analysis.

This is a text on the fundamentals of the theory of probability at an undergraduate or first-year graduate level for students in science, engineering, and economics. The only mathematical background required is knowledge of univariate and multivariate calculus and basic linear algebra. The book covers all of the standard topics in basic probability, such as combinatorial probability, discrete and continuous distributions, moment generating functions, fundamental probability inequalities, the central limit theorem, and joint and conditional distributions of discrete and continuous random variables. But it also has some unique features and a forward-looking feel. Some unique features of this book are its emphasis on conceptual discussions, a lively writing style, and on presenting a large variety of unusual and interesting examples; careful and more detailed treatment of normal and Poisson approximations (Chapters 6 and 10); better exposure to distribution theory, including developing superior skills in working with joint and conditional distributions and the bivariate normal distribution (Chapters 11, 12, and 13); a complete and readable account of finite Markov chains (Chapter 14); treatment of modern urn models and statistical genetics (Chapter 15); special efforts to make the book user-friendly, with unusually detailed chapter summaries, and a unified collection of formulas from the text, and from algebra, trigonometry, geometry, and calculus in the appendix of the book, for immediate and easy reference; and use of interesting *Use Your Computer* simulation projects as part of the chapter exercises to help students see a theoretical result evolve in their own computer work.

The exercise sets form a principal asset of this text. They contain a wide mix of problems at different degrees of difficulty. While many are straightforward, many others are challenging and require a student to think hard. These harder problems are always marked with an asterisk. The chapter ending exercises that are not marked

with an asterisk generally require only straightforward skills, and these are also essential for giving a student confidence in problem solving. The book also gives a set of supplementary exercises for additional homework and exam preparation. The supplementary problem set has 185 word problems and a very carefully designed set of 120 true/false problems. Instructors can use the true/false problems to encourage students to learn to think and also quite possibly for weekly homework. The total number of problems in the book is 810.

Students who take a course from this book should be extremely well-prepared to take more advanced probability courses and also courses in statistical theory at the level of Bickel and Doksum (2001), and Casella and Berger (2001). This book alone should give many students a solid working knowledge of basic probability theory, together with some experience with applications. The sections in the text that are marked with an asterisk are optional, and they are not essential for learning the most basic theory of probability. However, these sections have significant reference value, and instructors may choose to cover some of them at their discretion.

The book can be used for a few different types of one-semester courses; for example, a course that primarily teaches univariate probability, a course that caters to students who have already had some univariate probability, or a course that does a bit of both. The book can also be used to teach a course that does some theory and then some applications. A few such sample one-semester course outlines using this book are:

Sample course one: Univariate and some urn models Sections 1.1–1.5; 2.1; 3.1–3.4; 4.1–4.9, 4.10.1; 6.1–6.7; 7.1–7.5; 7.7.1; 8.1–8.4; 9.1–9.4; 10.1–10.5, 10.7; 15.4, 15.5

Sample course two: Mostly multivariate, with Markov chains and some urn models A four week review of univariate probability, followed by Sections 11.1–11.5; 12.1–12.6; 13.1–13.5; 8.6; 14.1–14.6; 15.1, 15.2, 15.4–15.6

Sample course three: Univariate, discrete multivariate, some Markov chains, and genetics Sections 1.1–1.5; 3.1–3.4; 4.1–4.6, 4.8, 4.9, 4.12; 6.3, 6.4, 6.6, 6.7, 6.9; 7.1, 7.3–7.5, 7.6.1; 8.1–8.6; 9.1–9.4; 10.1–10.4; 11.1–11.4; 14.1–14.3; 15.7–15.9.

A companion second volume of this book is planned for late 2010. The second volume will cater primarily to graduate students in mathematics, statistics, and machine learning and will cover advanced distribution theory, asymptotic theory and characteristic functions, random walks, Brownian motion and the empirical process, Poisson processes, extreme value theory and concent ration inequalities, a survey of models, including martingales, copulas, and exponential families, and an introduction to MCMC.

Peter Hall, Stewart Ethier, Burgess Davis, B.V. Rao, Wei-Liem Loh, Dimitris Politis, Yosi Rinott, Sara van de Geer, Jayaram Sethuraman, and Rabi Bhattacharya made scholarly comments on various drafts of this book. I am thankful to all of them. I am specifically deeply indebted to Peter Hall for the extraordinary nature of his counsel and support and for his enduring and selfless friendship and warmth.

I simply could not have written this book without Peter's help and mentoring. For this, and for being a unique counselor and friend to me, I am grateful to Peter.

I also want to express my deep appreciation for all the help that I received from Stewart Ethier as I was writing this book. Stewart was most gracious, patient, thoughtful, and kind. Burgess Davis affectionately read through several parts of the book, corrected some errors, and was a trusted counselor. Eight anonymous reviewers made superb comments and helped me make this a better book. Springer's series editors, Peter Bickel, George Casella, Steve Feinberg, and Ingram Olkin, helped me in every possible way at all times. I am thankful to them.

John Kimmel, as always, was a pleasure to work with. John's professionalism and his personal qualities make him a really dear person. My production editor Susan Westendorf graciously handled every production related issue and it was my pleasure to work with her. My copyeditor Hal Henglin did an unbelievably careful and thoughtful job. Indeed, if it was not for Hal, I could not have put this book out in a readable form. The technical staff at SPi Technologies, Pondicherry, India did a terrific and timely job of resetting the book in Springer's textbook template. Doug and Cheryl Crabill helped me with my computer questions and solved my problems with mysterious and magical powers. Shanti Gupta brought me to the United States and cared for me and was a guardian and a mentor for more than 15 years. I miss Shanti very much. Larry Brown, Persi Diaconis, Jon Wellner, Steve Lalley, Jim Pitman, C.R. Rao, and Jim Berger have given me support and sincere encouragement for many of my efforts. I appreciate all of them.

Human life is unreasonably fragile. It is important that our fondness for our friends not remain unspoken. I am thankful to numerous personal friends for their affection, warmth, and company over the years. It is not possible to name all of them. But I am especially grateful and fortunate for the magnificent and endearing support, camaraderie, and concern of some of my best friends, Jenifer Brown, Len Haff, Peter Hall, Rajeeva Karandikar, T. Krishnan, Wei-Liem Loh, B.V. Rao, Herman Rubin, Bill Strawderman, Larry Wasserman, and Dr. Julie Marshburn, MD. They have given me much more than I have cared to give in return. I appreciate them and their friendship more than I can express.

I had my core training in probability at the fundamental level in Dev Basu's classes at the ISI. I never met another teacher like Basu. I was simply fortunate to have him as my teacher and to have known him for the rare human being that he was. Basu told us that we must read Feller. I continue to believe that the two volumes of Feller are two all-time classics, and it's hard not to get inspired about the study of randomness once one has read Feller. I dedicate this book to William Feller and Dev Basu for bringing me the joy of probability theory.

But most of all, I am in love with my family for their own endless love for as long as I have lived. I hope they like this book.

West Lafayette, Indiana *Anirban DasGupta*

Contents

Chapter 1
Introducing Probability

Probability is a universally accepted tool for expressing degrees of confidence or doubt about some proposition in the presence of incomplete information or uncertainty. By convention, probabilities are calibrated on a scale of 0 to 1; assigning something a zero probability amounts to expressing the belief that we consider it impossible, while assigning a probability of one amounts to considering it a certainty. Most propositions fall somewhere in between. For example, if someone pulls out a coin and asks if the coin will show heads when tossed once, most of us will be inclined to say that the chances of the coin showing heads are 50%, or equivalently .5. On the other hand, if someone asks what the chances are that gravity will cease to exist tomorrow, we will be inclined to say that the chances of that are zero. In these two examples, we assign the chances .5 and 0 to the two propositions because in our life experience we have seen or heard that normal coins tend to produce heads and tails in roughly equal proportions and also that, in the past, gravity has never ceased to exist. Thus, our probability statements are based at some level on experience from the past, namely the propensity with which things, which we call *events*, tend to happen. But, as a third example, suppose we are asked what the chances are that civilized life similar to ours exists elsewhere in the known universe. Now the chances stated will undoubtedly differ from person to person. Now there is no past experience that we can count on to make a probabilistic statement, but many of us will still feel comfortable making rough probability statements on such a proposition. These are based purely on individual belief and understanding, and we think of them as *subjective probabilities*.

Whether our probability statements are based on past experience or subjective personal judgments, they obey a common set of rules that we can use to treat probabilities in a mathematical framework and use them for making decisions, predictions, understanding complex systems, as intellectual experiments, and for entertainment. Probability theory is one of the most beautiful branches of mathematics; the problems that it can address and the answers that it provides are often strikingly structured and beautiful. At the same time, probability theory is one of the most applicable branches of mathematics. It is used as the primary tool for analyzing statistical methodologies; it is used routinely in nearly every branch of science, such as biology, astronomy and physics, medicine, economics, chemistry,

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_1,

sociology, ecology, and finance, among others. A background in the theory, models, and applications of probability is almost a part of basic education. That is how important it is.

For classic and lively introductions to the subject of probability, we recommend Feller (1968). Later references with interesting examples include Ross (1984), Stirzaker (1994), and Pitman (1992).

1.1 Experiments and Sample Spaces

Treatment of probability theory starts with the consideration of a *sample space*. The sample space is the set of all possible outcomes in some physical experiment. For example, if a coin is tossed twice and after each toss the face that shows is recorded, then the possible outcomes of this particular coin-tossing experiment, say ξ, are HH, HT, TH, TT, with H denoting the occurrence of heads and T denoting the occurrence of tails. We call

$$\Omega = \{HH, HT, TH, TT\}$$

the sample space of the experiment ξ.

We instinctively understand what an experiment means. An experiment is a physical enterprise that can, in principle, be repeated infinitely many times independently. For example,

$\xi =$ choose a number between 1 and 10 and record the value of the chosen number,
$\xi =$ toss a coin three times and record the sequence of outcomes,
$\xi =$ arrange five people in a lineup for taking a picture,
$\xi =$ distribute 52 cards in a deck of cards to four players so that each player gets 13 cards,
$\xi =$ count the number of calls you receive on your cell phone on a given day, and
$\xi =$ measure someone's blood pressure

are all activities that can, in principle, be repeated and are experiments. Notice that, for each of these experiments, the ultimate outcome is uncertain until the experiment has actually been performed. For example, in the first experiment above, the number that ultimately gets chosen could be any of $1, 2, \ldots, 10$. The set of all these possible outcomes constitutes the sample space of the experiment. Individual possible outcomes are called the *sample points* of the experiment.

In general, a sample space is a general set Ω, finite or infinite. An easy example where the sample space Ω is infinite is to toss a coin until the first time heads shows up and record the number of the trial at which the first head showed up. In this case, the sample space Ω is the *countably infinite* set

$$\Omega = \{1, 2, 3, \ldots\}.$$

Sample spaces can also be *uncountably infinite*; for example, consider the experiment of choosing a number *at random* from the interval $[0, 1]$. Although we do not yet know what choosing a number at random from $[0, 1]$ means, we understand that the chosen number could be any number in $[0, 1]$, so the sample space of such an experiment should be $\Omega = [0, 1]$. In this case, Ω is an uncountably infinite set. In all cases, individual elements of a sample space will be denoted as ω. The first task is to define *events* and explain what is meant by the probability of an event.

Loosely speaking, events are collections of individual sample points. For example, in the experiment of tossing a coin twice, consider the collection of sample points $A = \{HT, TH\}$. This collection corresponds to an interesting statement or proposition, namely that when a coin is tossed twice, it will show one head and one tail. A particular collection of sample points may or may not turn out to be an interesting statement in every example. But it will nevertheless be an event. Here is the formal definition of an event.

Definition 1.1. Let Ω be the sample space of an experiment ξ. Then any subset A of Ω, including the empty set ϕ and the entire sample space Ω, is called an *event*. Events may contain even one single sample point ω, in which case the event is a *singleton set* $\{\omega\}$. We will want to assign probabilities to events. But we want to assign probabilities in such a way that they are logically consistent. In fact, this cannot be done in general if we insist on assigning probabilities to arbitrary collections of sample points, i.e., arbitrary subsets of the sample space Ω. We can only define probabilities for such subsets of Ω that are tied together like a family, the exact concept being that of a σ-field. In most applications, including those cases where the sample space Ω is infinite, events that we would want to normally think about will be members of such an appropriate σ-field. So we will not mention the need for consideration of σ-fields further and get along with thinking of events as subsets of the sample space Ω, including in particular the empty set ϕ and the entire sample space Ω itself.

1.2 Set Theory Notation and Axioms of Probability

Set theory notation will be essential in our treatment of events because events are sets of sample points. So, at this stage, it might be useful to recall the following common set theory notation:

Given two subsets A and B of a set Ω,

A^c = set of points of Ω not in A,
$A \cap B$ = set of points of Ω that are in both A and B,
$A \cup B$ = set of points of Ω that are in at least one of A and B,
$A \triangle B$ = set of points of Ω that are in exactly one of A and B,
$A - A \cap B$ = set of points of Ω that are in A but not in B.

If Ω is the sample space of some experiment and A and B are events in that experiment, then the probabilistic meaning of this notation would be as follows: Given two events A and B in some experiment,

$A^c = A$ does not happen,
$A \cap B =$ both A and B happen; the notation AB is also sometimes used to mean $A \cap B$,
$A \cup B =$ at least one of A and B happens,
$A \triangle B =$ exactly one of A and B happens,
$A - A \cap B = A$ happens, but B does not.

Example 1.1. This example is to help interpret events of various types using the symbols of set operation. This becomes useful for calculating probabilities by setting up the events in set theory notation and then using a suitable rule or formula. For example,

$$\text{at least one of } A, B, C = A \cup B \cup C;$$

$$\text{each of } A, B, C = A \cap B \cap C;$$

$$A, \text{ but not } B \text{ or } C = A \cap B^c \cap C^c;$$

$$A \text{ and exactly one of } B \text{ or } C = A \cap (B \triangle C) = (A \cap B \cap C^c) \cup (A \cap C \cap B^c);$$

$$\text{none of } A, B, C = A^c \cap B^c \cap C^c.$$

It is also useful to recall the following elementary facts about set operations.

Proposition.
(a) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$;
(b) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
(c) $(A \cup B)^c = A^c \cap B^c$;
(d) $(A \cap B)^c = A^c \cup B^c$.

Now, here is a definition of what counts as a legitimate probability of events.

Definition 1.2. Given a sample space Ω, a probability or a *probability measure* on Ω is a function P on subsets of Ω such that

(a) $P(A) \geq 0$ for any $A \subseteq \Omega$;
(b) $P(\Omega) = 1$;
(c) given disjoint subsets $A_1, A_2, \ldots$ of Ω, $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Property (c) is known as *countable additivity*. Note that it is not something that can be proved, but it is like an assumption or an *axiom*. In our experience, we have seen that operating as if the assumption is correct leads to useful and credible answers to many problems, so we accept it as a reasonable assumption. Not all probabilists agree that countable additivity is natural, but we will not get into that debate in this book. One important point is that finite additivity is subsumed in countable additivity i.e., if there is some finite number m of disjoint subsets $A_1, A_2, \ldots, A_m$ of Ω, then $P(\cup_{i=1}^{m} A_i) = \sum_{i=1}^{m} P(A_i)$. Also, it is useful to note that the last two conditions in the definition of a probability measure imply that $P(\phi)$, the probability of the empty set or the *null event*, is zero.

One notational convention is that, strictly speaking, for an event that is just a singleton set $\{\omega\}$, we should write $P(\{\omega\})$ to denote its probability. But, to reduce clutter, we will simply use the more convenient notation $P(\omega)$.

One pleasant consequence of the axiom of countable additivity is the following intuitively plausible result.

Theorem 1.1. *Let $A_1 \supset A_2 \supset A_3 \supset \cdots$ be an infinite family of subsets of a sample space Ω such that $A_n \downarrow A$. Then, $P(A_n) \to P(A)$ as $n \to \infty$.*

Proof. On taking the complements $B_i = A_i^c, i \geq 1, B = A^c$, the result is equivalent to showing that if $B_1 \subset B_2 \subset B_3 \cdots, B_n \uparrow B$, then $P(B_n) \to P(B)$.

Decompose B_n for a fixed n into disjoint sets as $B_n = \cup_{i=1}^n (B_i - B_{i-1})$, where $B_0 = \phi$ and the difference notation $B_i - B_{i-1}$ means $B_i \cap B_{i-1}^c$. Therefore,

$$P(B_n) = \sum_{i=1}^{n} P(B_i - B_{i-1})$$

$$\Rightarrow \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i - B_{i-1}) = \sum_{i=1}^{\infty} P(B_i - B_{i-1}) = P(B),$$

as $\cup_{i=1}^{\infty} (B_i - B_{i-1}) = B$.

Remark. Interestingly, if we assume the result of this theorem as an axiom and also assume finite additivity, then countable additivity of a probability measure follows.

1.3 How to Interpret a Probability

Many think that probabilities do not exist in real life. Nevertheless, a given or a computed value of the probability of some event A can be used in order to make conscious decisions. The entire subject of statistics depends on the use of probabilities. We depend on probabilities to make simple choices in our daily lives. For example, we carry an umbrella to work if the weather report gives a high probability of rain. Where do these probabilities come from? Two common interpretations are the following.

Long-run frequency interpretation. If the probability of an event A in some actual physical experiment ξ is p, then we believe that if ξ is repeated independently over and over again, then in the long run the event A will happen $100p\%$ of the time. We apply the long-run percentage to the one-time experiment that will actually be conducted. For better or worse, such probabilities that appear to come from an actual physical random process are called *frequentist probabilities*. Frequentist probabilities make sense in situations where we can obtain actual physical experience or data. For example, we can gather experience about a particular game in a casino and come to reasoned conclusions about the chances of winning.

Subjective Probabilities. At the heart of frequentist probabilities is the implicit assumption of repeatability of some genuine physical process. We gather experience from repeated experimentation and apply the past experience to make probabilistic statements. But we cannot gather actual experience if we want to assign a probability that the subsurface ocean in a moon of Saturn has microbial life or that the Big Bang actually happened. In such situations, we are forced to use probabilities based on beliefs or feelings based on personal or collective knowledge, the so-called *subjective probabilities*. For example, I should say that the probability that the Big Bang actually happened is .8 if I feel that it is just as certain as a red ball being drawn out from a box that has 80 red balls and 20 green balls. An obvious problem is that different people will assign different subjective probabilities in such a situation, and we cannot try to verify *whose belief is correct* by gathering experience, or data. Nevertheless, we are forced to use subjective probabilities in all kinds of situations because the alternative would be to do nothing. Regardless of which type of probability we may use, the manipulations and the operations will fortunately be the same. But, once a probability statement has been made in some specific problem, it is often a good idea to ask where this probability came from. The interested reader can learn from Basu (1975), Berger (1986), or Savage (1954) about the lively and yet contentious philosophical debates about the meaning of probability and for provocative and entertaining paradoxes and counterexamples.

Example 1.2. Consider our previous experiment ξ of tossing a coin twice and recording the outcome after each toss. A valid probability measure P on the sample space $\Omega = \{HH, HT, TH, TT\}$ of this experiment is one that assigns probability $\frac{1}{4}$ to *each* of the four sample points; i.e., $P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4}$. By the additivity property assumed in the definition, if we consider the event $A = \{HT, TH\}$ = the statement that exactly one head and exactly one tail will be obtained, then $P(A) = P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. If we believed in this probability, then a bet that offers to pay us ten dollars should the event A happen and require us to pay ten dollars if it does not happen would be considered a *fair bet*. Indeed, the original development of probability was motivated by betting and gambling scenarios involving coin, dice, or card games. Because of this, and also because they seem to provide an endless supply of interesting problems and questions, many of our examples will be based on suitable coin, dice, and card experiments.

Definition 1.3. Let Ω be a finite sample space consisting of N sample points. We say that the sample points are *equally likely* if $P(\omega) = \frac{1}{N}$ for each sample point ω.

An immediate consequence, due to the additivity axiom, is the following useful formula.

Proposition. Let Ω be a finite sample space consisting of N equally likely sample points. Let A be any event, and suppose A contains n distinct sample points. Then

$$P(A) = \frac{n}{N} = \frac{\text{number of sample points favorable to } A}{\text{total number of sample points}}.$$

Remark. In many experiments, when we assume that the sample points are equally likely, we do so expecting that the experiment has been conducted in an unbiased or fair way. For example, if we assign probability .5 (or 50%) to heads being obtained when a coin is tossed just once, we do so thinking that the coin in question is just a normal coin and has not been manipulated in any way. Indeed, we say that a coin is a *fair coin* if $P(H) = P(T) = .5$ when the coin is tossed once. Similarly, we say that a die is a fair die if $P(1) = P(2) = \cdots = P(6) = \frac{1}{6}$ when the die is rolled once. The assumption of equally likely sample points is immensely useful in complicated experiments with large sample spaces, where physically listing all the sample points would be difficult or even impossible. However, it is important to remember that the assumption of equally likely sample points *cannot* be made in every problem. In such cases, probabilities of events cannot be calculated by the convenient method of taking the ratio of favorable sample points to the total number of sample points, and they will have to be calculated by considering what the probabilities of the different sample points are.

Example 1.3 (A Computer Simulation). This example illustrates the long-run frequency interpretation of probabilities. We simulate the roll of a fair die on a computer. According to the definition of a fair die and the long-run frequency interpretation of probabilities, we should see that the percentage of times that any face appears should settle down near $\frac{100}{6}\% = 16.67\%$ after many rolls. The word *many* cannot be quantified in general. The main point is that we should expect heterogeneity and oscillations in the percentages initially, but as we increase the number of rolls, the percentages should all approach $\frac{100}{6}\% = 16.67\%$. Here is a report of a computer simulation.

Number of rolls	% of 1	% of 2	% of 3	% of 4	% of 5	% of 6
20	20	10	5	15	25	25
50	6	24	30	18	14	8
100	18	21	15	12	11	23
250	16.8	15.2	19.6	14.0	18.4	16
1000	17.6	17.3	16.9	15.8	15.3	17.1

Unmistakably, we see that the percentages appear to approach some limiting value when the number of rolls increases; indeed, they will all approach 16.67% when the number of rolls goes to infinity.

1.4 Calculating Probabilities

Probabilities are useful for making decisions or predictions, but only if we can calculate them. If we cannot calculate a probability, then obviously we cannot assess if it is large or small or something in between. In the simplest experiments, we will typically be able to calculate probabilities by examining the sample points. In more complex experiments, this would no longer be feasible. That is when formulas and

theorems that tell us how to calculate a probability under a given set of assumptions will be useful. We will see some simple experiments and then a number of basic formulas in this section.

1.4.1 Manual Counting

We first describe a collection of examples of simple experiments and the associated sample spaces where the assumption of equally likely sample points seems reasonable and then calculate probabilities of some interesting events. These experiments are simple enough that we can just list the sample points after a little thinking.

Example 1.4. Let ξ be the experiment of tossing a coin three times and recording the outcome after each toss. By inspection, we find that the sample space is $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$; indeed, since on each individual toss we have two possible outcomes, the number of sample points in the overall experiment is $2 \times 2 \times 2 = 8$, which is what we see in Ω. Suppose now that we take each of the eight sample points to be equally likely. This corresponds to an expression of our belief that the coin being tossed is a fair coin and that subsequent tosses are not affected by what may have been the outcomes in the tosses already completed; this latter concept is formally known as *independence* and will be treated formally later.

Under the equally likely assumption, then, $P(\text{At least one head is obtained}) = P\{HHH, HHT, HTH, HTT, THH, THT, TTH\} = \frac{7}{8}$. Alternatively, we could have calculated the probability that no heads are obtained at all, the probability of which is $P(TTT) = \frac{1}{8}$, and obtained $P(\text{At least one head is obtained})$ as

$$P(\text{At least one head is obtained}) = 1 - P(\text{No heads are obtained}) = 1 - \frac{1}{8} = \frac{7}{8}.$$

The event where no heads are obtained is the *complement* of the event where at least one head is obtained, and always $P(A) + P(A^c) = 1$, with A^c denoting the complement of A.

Likewise, $P(\text{At least one head and at least one tail are obtained}) = 1 - P(HHH) - P(TTT) = 1 - \frac{1}{8} - \frac{1}{8} = \frac{6}{8} = .75$.

The experiment of this example is simple enough that we can just list the sample points and calculate probabilities of events by counting favorable sample points.

Example 1.5 (Motivating Disjoint Events). Let ξ be the experiment of rolling a die twice and recording the outcome after each roll. Then there are $6 \times 6 = 36$ sample points, and the sample space is $\Omega = \{11, 12, 13, \ldots, 64, 65, 66\}$. Consider the following two events:

$$A = \text{the sum of the two numbers is odd;}$$
$$B = \text{the product of the two numbers is odd.}$$

Then, the favorable sample points for A are those for which one number is even and the other is odd; that is, sample points like 12 or 14, etc. By simple counting, there are 18 such favorable sample points so $P(A) = \frac{18}{36} = .5$. On the other hand, the favorable sample points for B are those for which both numbers are odd; that is, sample points like 11 or 13, etc. There are nine such favorable sample points, so $P(B) = \frac{9}{36} = .25$.

Interestingly, there are *no sample points that are favorable to both A and B*; in set theory notation, the *intersection* of A and B is empty (that is, $A \cap B = \phi$). Two such events A and B are called disjoint or mutually exclusive events, and then $P(A \cap B) = 0$.

Definition 1.4. Two events A and B are said to be disjoint or mutually exclusive if $A \cap B = \phi$, in which case $P(A \cap B) = 0$.

Example 1.6 (With and Without Replacement). Consider the experiment ξ where two numbers are chosen simultaneously *at random* from $\{0, 1, 2, \ldots, 9\}$. Since the numbers are chosen simultaneously, by implication they must be different; such sampling is called sampling *without replacement*. Probabilistically, sampling without replacement is also the same as drawing the two numbers one at a time with the restriction that the same number cannot be chosen twice. If the numbers are chosen one after the other and the second number could be equal to the first number, then the sampling is called sampling *with replacement*. In this example, we consider sampling without replacement. Consider the events

$$\begin{aligned} A &= \text{the first chosen number is even;} \\ B &= \text{the second chosen number is even;} \\ C &= \text{both numbers are even;} \\ D &= \text{at least one of the two numbers is even.} \end{aligned}$$

The sample space $\Omega = \{01, 02, 03, \ldots, 96, 97, 98\}$ has $10 \times 9 = 90$ sample points. Suppose that, due to the random or unbiased selection of the two numbers, we assign an equal probability, $\frac{1}{90}$, of selecting any of the 90 possible pairs. Event A is favored by the sample points $\{01, 02, \ldots, 88, 89\}$; thus, A is favored by $5 \times 9 = 45$ sample points, so $P(A) = 45/90 = .5$. Similarly, $P(B)$ is also .5. Event C is favored by those sample points that are in both A and B; i.e., in set theory notation, $C = A \cap B$. By direct listing, $A \cap B = \{02, 04, \ldots, 86, 88\}$; there are $5 \times 4 = 20$ such sample points, so $P(C) = P(A \cap B) = 20/90 = 2/9$. On the other hand, event D is favored by those sample points that favor A or B, or perhaps both; i.e., D is favored by sample points that favor at least one of A, B. In set theory notation, $D = A \cup B$, and by direct listing, it is verified that $P(D) = P(A \cup B) = 70/90 = 7/9$. We note that the collection of sample points that favor at least one of A, B can be found by writing the sample points in A, then writing the sample points in B, and eventually taking out those sample points that were written twice; i.e., the sample

points in $A \cap B$. So, we should have $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/2 - 2/9 = 7/9$, which is what we found by direct listing. Indeed, this is a general rule.

Addition Rule. For any two events A, B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

1.4.2 General Counting Methods

In more complicated experiments, it might be difficult or even impossible to manually list all the sample points. For example, if you toss a coin 20 times, the total number of sample points would be $2^{20} = 1{,}048{,}576$, which is larger than a million. Obviously we do not want to calculate probabilities for such an example by manual listing and manual counting.

Some facts about counting and basic combinatorics will be repeatedly useful in complex experiments, so it is useful to summarize them before we start using them.

Proposition.

(a) The number of ways of linearly arranging n distinct objects when the order of arrangement matters = $n!$.

(b) The number of ways of choosing r distinct objects from n distinct objects when the order of selection is important = $n(n-1)\cdots(n-r+1)$.

(c) The number of ways of choosing r distinct objects from n distinct objects when the order of selection is not important = $\binom{n}{r} = \frac{n!}{r!(n-r)!}$.

(d) The number of ways of choosing r objects from n distinct objects if the same object could be chosen repeatedly $= n^r$.

(e) The number of ways of distributing n distinct objects into k distinct categories when the order in which the distributions are made is not important and n_i objects are to be allocated to the ith category $= \binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-n_2-\cdots n_{k-1}}{n_k}$ $= \frac{n!}{n_1!n_2!\cdots n_k!}$.

Example 1.7. Tim has eight pairs of trousers, 15 shirts, six ties, and four jackets, of which two pairs of trousers, five shirts, two ties, and two jackets are green. Suppose that one morning Tim selects his outfit completely at random. Then, the total number of possible ways that he could select his outfit is $8 \times 15 \times 6 \times 4 = 2880$. These are the sample points of Tim's experiment. Since selection is completely at random, we assume that the sample points are equally likely. There are $2 \times 5 \times 2 \times 2 = 40$ possible ways that he could choose a completely green outfit, so $P(\text{Tim is dressed completely in green on a particular day}) = 40/2880 = .014$. Notice that in this example there were far too many sample points to actually list them. Nevertheless, we could calculate the required probability by simple *counting*. Counting methods are thus extremely useful in calculating probabilities when sample points are equally likely, and we will see more sophisticated examples later.

Example 1.8. A carton of eggs has 12 eggs, of which three happen to be bad, although we do not know that there are some bad eggs. We want to make a three-egg omelet. There are $\binom{12}{3} = \frac{12!}{3!9!} = 220$ ways to select three eggs from the carton of 12 eggs. It seems reasonable to assume that the three eggs are selected without any bias; i.e., at random. Then, each sample point is equally likely. The omelet will not contain any bad eggs if our three eggs are all chosen from the nine that are good eggs. This can be done in $\binom{9}{3} = 84$ ways. Therefore, $P(\text{The three-egg omelet contains no bad eggs}) = 84/220 = .38$.

Example 1.9. Suppose six distinguishable cookies are distributed completely at random to six children, with it being possible that the same child could get more than one cookie. Thus, there are $6^6 = 46,656$ sample points; i.e., there are 46,656 ways to distribute the six cookies among the six children.

The exactly equitable case is when each child gets exactly one cookie, although who gets which cookie is flexible. The number of ways to distribute six cookies to six children in any arbitrary way is $6! = 720$, so the probability that this will happen is $720/46656 = .015$. The complement is that at least one child gets no cookies at all, which therefore has the probability $1 - .015 = .985$.

Example 1.10 (The Shoe Problem). Suppose there are five pairs of shoes in a closet and four shoes are taken out at random. What is the probability that among the four that are taken out, there is at least one complete pair?

The total number of sample points is $\binom{10}{4} = 210$. Since selection was done completely at random, we assume that all sample points are equally likely. At least one complete pair would mean two complete pairs, or exactly one complete pair and two other nonconforming shoes. Two complete pairs can be chosen in $\binom{5}{2} = 10$ ways. Exactly one complete pair can be chosen in $\binom{5}{1}\binom{4}{2} \times 2 \times 2 = 120$ ways. The $\binom{5}{1}$ term is for choosing the pair that is complete; the $\binom{4}{2}$ term is for choosing two incomplete pairs, and then from each incomplete pair one chooses the left or the right shoe. Thus, the probability that there will be at least one complete pair among the four shoes chosen is $(10 + 120)/210 = 13/21 = .62$.

Example 1.11 (Avoiding Tedious Listing). Suppose three balls are distributed completely at random into three urns. What is the probability that exactly one urn remains empty?

There are $3^3 = 27$ sample points, which we assume to be equally likely. If exactly one urn is to remain empty, then the two other urns receive all the three balls, one getting two balls and the other getting one. This can be done in $\binom{3}{1}\left(\binom{3}{2}\binom{1}{1} + \binom{3}{1}\binom{2}{2}\right) = 18$ ways. Hence, the probability that exactly one urn will remain empty is $18/27 = .667$, which can also be verified by listing the 27 sample points.

Example 1.12 (Bridge). Bridge is a card game in which 52 cards are distributed to four players, say North, South, East, and West, each receiving 13 cards. It is assumed that distribution is done at random. Consider the events

A = North has no aces;

B = neither North nor South has any aces;

C = North has all the aces;

D = North and South together have all the aces.

For $P(A)$, if North has no aces, his 13 cards must come from the other 48 cards, so $P(A) = \binom{48}{13}/\binom{52}{13} = .304$. Similarly, $P(B) = \binom{48}{13}\binom{35}{13}/\left(\binom{52}{13}\binom{39}{13}\right) = 46/833 = .055$. Note that D is probabilistically equivalent to the statement that neither East nor West has any aces, and therefore $P(D) = P(B) = .055$. Finally, for $P(C)$, if North has all the aces, then his other nine cards come from the 48 non-ace cards, so $P(C) = \binom{4}{4}\binom{48}{9}/\binom{52}{13} = 11/4165 = .0026$.

Example 1.13 (Five-Card Poker). In five-card poker, a player is given five cards from a full deck of 52 cards at random. Various named hands of varying degrees of rarity exist. In particular, we want to calculate the probabilities of A = *two pairs* and B = a *flush*. Two pairs is a hand with two cards each of two different denominations and the fifth card of some other denomination; a flush is a hand with five cards of the same suit, but the cards cannot be of denominations in a sequence. Then, $P(A) = \binom{13}{2}[\binom{4}{2}]^2\binom{44}{1}/\binom{52}{5} = .04754$.

To find $P(B)$, note that there are ten ways to select five cards from a suit such that the cards are in a sequence, namely $\{A, 2, 3, 4, 5\}, \{2, 3, 4, 5, 6\}, \cdots, \{10, J, Q, K, A\}$, so $P(B) = \binom{4}{1}\left(\binom{13}{5} - 10\right)/\binom{52}{5} = .00197$.

Example 1.14 (Clever Counting). Suppose n integers are chosen with replacement (that is, the same integer could be chosen repeatedly) at random from $\{1, 2, \cdots, N\}$. We want to calculate the probability that the chosen numbers arise according to some nondecreasing sequence. This is an example of clever counting.

Take a nondecreasing sequence of n numbers and combine it with the full set of numbers $\{1, 2, \cdots, N\}$ to form a set of $n + N$ numbers. Now rearrange these numbers in a nondecreasing order. Put a bar between consecutive distinct numbers in this set and a dot between consecutive equal numbers in this set. The number to the right of each dot is an element of the original n-number sequence. There are n dots in this picture, and they can be positioned at n places out of $N + n - 1$ places. Therefore, the probability that the original n-member sequence is nondecreasing is $\binom{N+n-1}{n}/N^n$.

1.5 Inclusion-Exclusion Formula

The inclusion-exclusion formula is a formula for the probability that at least one of n general events $A_1, A_2, \ldots, A_n$ will happen. The formula has many applications and is also useful for providing upper and lower bounds for the probability that at least one of $A_1, A_2, \ldots, A_n$ will happen.

Theorem 1.2. *Let $A_1, A_2, \ldots, A_n$ be n general events. Then,*

$$P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i) - \sum_{1\le i<j\le n} P(A_i \cap A_j) + \sum_{1\le i<j<k\le n} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Proof. The theorem is proved by induction. Suppose it is known to be true for $n-1$ general events $A_1, A_2, \ldots, A_{n-1}$. Define $A = \cup_{i=1}^{n-1} A_i$ and $B = A_n$. Then, by the addition rule for two events,

$$P(\cup_{i=1}^n A_i) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = \sum_{i=1}^{n-1} P(A_i) - \sum_{1\le i<j\le n-1} P(A_i \cap A_j) + \cdots + (-1)^n P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) + P(A_n) - P\left(U_{i=1}^{n-1}(A_i \cap A_n)\right).$$

Applying the inclusion-exclusion formula to the $(n-1)$ events $A_i \cap A_n, i = 1, 2, \cdots, n-1$, and rearranging terms, the expression in the theorem follows.

Here are some examples of applications of the inclusion-exclusion formula.

Example 1.15 (Missing Faces in Dice Rolls). Suppose a fair die is rolled n times. We want to calculate the probability that at least one of the six sides of the die never shows up in these n rolls.

To do this, define A_i = side i never shows up in the n rolls, $1 \le i \le 6$. Then, assuming that all 6^n sample points are equally likely,

$$P(A_i) = 5^n/6^n; P(A_i \cap A_j) = 4^n/6^n; P(A_i \cap A_j \cap A_k) = 3^n/6^n,$$

etc., and these hold for any $i, i < j, i < j < k$, etc. Plugging this into the inclusion-exclusion formula,

$$\begin{aligned} p_n &= P(\text{At least one of the six sides never shows up}) \\ &= \binom{6}{1}(5^n/6^n) - \binom{6}{2}(4^n/6^n) + \binom{6}{3}(3^n/6^n) - \binom{6}{4}(2^n/6^n) + \binom{6}{5}(1/6^n) \\ &= 6\ (5^n/6^n) - 15\ (4^n/6^n) + 20\ (3^n/6^n) - 15\ (2^n/6^n) + 6/6^n. \end{aligned}$$

By computing, we find that $p_{10} = .73, p_{12} = .56, p_{13} = .49, p_{15} = .36, p_{20} = .15, p_{25} = .06, p_{35} = .01$. Note that $1 - p_n$ is the probability that each of the six faces will show up within n rolls. In particular, it takes 13 rolls of a fair die to have a better than 50% chance that each of the six sides would show up. A plot of the probability that each of the six faces of a fair die will show up within n rolls of the die is given to read off p_n for a given n (see Figure 1.1).

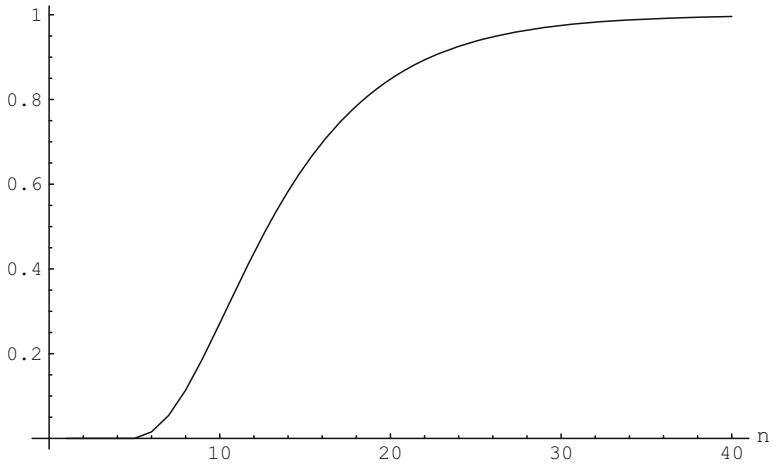


Fig. 1.1 Probability that each of the six races of a fair die will show up within n rolls

Example 1.16 (Missing Suits in a Bridge Hand). Consider a specific player, say North, in a bridge game. We want to calculate the probability that North's hand is void in at least one suit. To do this, denote the suits as 1, 2, 3, 4 and let A_i = North's hand is void in suit i.

Then, by the inclusion-exclusion formula,

$$P(\text{North's hand is void in at least one suit}) = P(A_1 \cup A_2 \cup A_3 \cup A_4)$$

$= 4\binom{39}{13}/\binom{52}{13} - 6\binom{26}{13}/\binom{52}{13} + 4\binom{13}{13}/\binom{52}{13} = .051$, which is small, but not very small.

Example 1.17. Here is an easier example of an application of the inclusion-exclusion formula. Suppose a cabinet contains two white cups, two red cups, two white plates, and two red plates. Suppose the four cups are placed at random on the four plates. We want to find the probability that no cup is on a plate with a matching color.

Define

A = at least one white cup is on a white plate;
B = at least one red cup is on a red plate;
C = the first white cup is on a white plate;
D = the second white cup is on a white plate.

Then we want to find $P(A \cup B)^c$. But, in this example, a moment's thinking shows that $A = B$. So, $A \cup B = A$, and we want $P(A^c)$. On the other hand, $A = C \cup D$, and by the inclusion-exclusion formula, $P(C \cup D) = P(C) + P(D) - P(C \cap D) = 1/2 + 1/2 - 4/24 = 5/6$. Therefore, our required probability is $P(A^c) = 1 - P(A) = 1 - P(C \cup D) = 1 - 5/6 = 1/6$.

1.6 ∗ Bounds on the Probability of a Union

If we denote

$$S_1 = \sum_{i=1}^{n} P(A_i), S_2 = \sum_{1 \le i < j \le n} P(A_i \cap A_j), S_3 = \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k),$$

etc., then the inclusion-exclusion formula says that $P(\cup_{i=1}^{n} A_i) = S_1 - S_2 + S_3 - \cdots$. The inclusion-exclusion formula can be hard to apply exactly because the quantities S_j for large indices j can be difficult to calculate. However, fortunately, the inclusion-exclusion formula leads to bounds in both directions for the probability of the union of n general events. We have the following series of bounds.

Theorem 1.3 (Bonferroni Bounds). *Given n events $A_1, A_2, \cdots, A_n$, let $p_n = P(U_{i=1}^{n} A_i)$. Then,*

$$p_n \le S_1; p_n \ge S_1 - S_2; p_n \le S_1 - S_2 + S_3; \cdots .$$

In addition,

$$P(\cap_{i=1}^{n} A_i) \ge 1 - \sum_{i=1}^{n} P(A_i^c).$$

Example 1.18. Suppose each of ten events $A_1, A_2, \cdots A_{10}$ has probability .95 or more; thus, $P(A_i^c) \le .05$ for each i. From the last Bonferroni bound given above, $P(\cap_{i=1}^{n} A_i) \ge 1 - 10 \times .05 = .5$. Each A_i by itself has a 95% probability or more of occurring. But that does not mean that with a high probability all ten events will occur. What kinds of probability assurances can we provide that indeed all ten events will occur? The bound we just derived says that we can be at least 50% sure that all ten events will occur. This is typically rather crude, but these bounds are sometimes used by statisticians to make overall accuracy statements of their inferences when they have made a number of inferences simultaneously.

Here are two better bounds on p_n.

Theorem 1.4.
(a) **(Galambos-Simonelli Bound).**

$$S_1 - S_2 + \frac{2}{n-1} S_3 \le p_n \le S_1 - \frac{2}{n} S_2;$$

(b) **(Chung-Erdös Bound)**

$$p_n \ge \frac{S_1^2}{2S_2 + S_1}.$$

See Galambos and Simonelli (1996) and Rao (1973) for all of these bounds.

1.7 Synopsis

Some key facts and formulas in this chapter are now restated here.

(a) Given any two events, A, B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
(b) More generally, given n events $A_1, \ldots, A_n$,

$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n).$$

(c) Union corresponds to at least one; intersection corresponds to both if there are two events, and it corresponds to all the events occurring if there are more than two events.
(d) In simple problems, you can list all the sample points and find probabilities by manual counting. However, in more complex examples, you must use suitable counting formulas to find probabilities.
(e) If there are N equally likely sample points in some example and n of them are favorable to some event A, then $P(A) = \frac{n}{N}$.

1.8 Exercises

Exercise 1.1. Define a sample space and count the number of sample points for each of the following experiments:

(a) Select at random a catcher, pitcher, first baseman, and second baseman from a group of 15 baseball players.
(b) Select at random a chairman and a deputy chairman from a department of 40 professors.
(c) Select at random a best-tasting, a second-best-tasting, and a worst-tasting wine from a selection of 20 different types of wine.
(d) The possible scores on a certain test are H, P, and F, for High Pass, Pass, and Fail, respectively. Give this test to two students and record their scores on the test.
(e) Toss a coin until a head appears, and record the number of tosses required.

Exercise 1.2. Suppose Mark pulls a coin from his pocket and tosses it four times and always obtains a tail. But the coin looks normal. Would you start to suspect that Mark pulled a biased coin?

Exercise 1.3. A telephone number consists of ten digits, of which the first digit is one of $1, 2, \ldots, 9$ and the others can be $0, 1, 2, \ldots, 9$. What is the probability that 0 appears at most once in a telephone number, if all the digits are chosen completely at random?

Exercise 1.4. Mark goes out to dinner twice a week. If he chooses the days to go out at random, what is the probability that he goes out on exactly one weeknight?

Exercise 1.5. The population of Danville is 20,000. Can it be said with certainty that there must be two or more people in Danville with exactly the same three initials?

Exercise 1.6 (Skills Exercise). Events A, B, and C are defined in a sample space Ω. Find expressions for the following probabilities in terms of $P(A), P(B), P(C), P(AB), P(AC), P(BC)$, and $P(ABC)$; here AB means $A \cap B$, etc.:

(a) the probability that exactly two of A, B, C occur;
(b) the probability that exactly one of these events occurs;
(c) the probability that none of these events occur.

Exercise 1.7 (Skills Exercise). Let E, F, and G be three events. Find expressions for the following events:

(a) only E occurs;
(b) both E and G occur, but not F;
(c) all three occur;
(d) at least one of the events occurs;
(e) at most two of them occur.

Exercise 1.8. Mrs. Jones predicts that if it rains tomorrow it is bound to rain the day after tomorrow. She also thinks that the chance of rain tomorrow is 1/2 and that the chance of rain the day after tomorrow is 1/3. Are these subjective probabilities consistent with the axioms and theorems of probability?

Exercise 1.9. In which of the following are events A and B mutually exclusive?

(a) Roll two dice. A is the event of a sum of 9; B is the event of a double (i.e., the same value on both dice).
(b) Draw 13 cards from a deck of 52 cards. A is the event of drawing at least one club; B is the event of drawing no aces.
(c) Toss a coin twice. A is the event of a head on the first toss; B is the event of a head on the second toss.

Exercise 1.10. Consider the experiment of tossing a coin three times. Give verbal descriptions for the following events:

(a) $\{HHH, HHT, HTH, HTT\}$.
(b) $\{HHH, HHT, HTH, THH\}$.

(c) $\{HHT, HHH, TTH, TTT\}$.
(d) $\{HTH, HTT, TTT, TTH\}$.

Exercise 1.11 (Odd Man Out). Each of three people toss a coin. What is the probability of *someone* being the "odd man out"? This means that two of them obtain an identical outcome, while the odd man gets a different one.

Exercise 1.12 (Elementary Number Theory). One number is chosen at random from 0 to 9999. What is the probability that it is divisible by 5? That it is divisible by both 2 and 5? That it is divisible by 2 but not by 5?

Exercise 1.13. The letters in the word FULL are rearranged at random. What is the probability that it still spells FULL?

Exercise 1.14. A seafood omelet will be made with two eggs, a piece of salmon, and a slice of cheese by choosing the items at random from four eggs, two pieces of salmon, and three slices of cheese. One egg, one piece of salmon, and one slice of cheese have gone bad. What is the probability that the omelet will contain at least one bad item?

Exercise 1.15 (Coincidence). Three families, each with three members, are lined up at random for a picture. What is the probability that members of each family happen to be together (that is, not separated by someone from another family) in the picture?

Exercise 1.16. Twenty cookies are to be distributed to ten children. In how many ways can the cookies be distributed if (a) any cookie can be given to any child; (b) If two cookies are to be given to each child?

Exercise 1.17 (Bridge Deals). What is the total possible number of deals in a bridge game? A deal is a distribution of 52 cards to four players, each receiving 13 cards.

Exercise 1.18. In a building with six floors, an elevator starts with four people at the ground floor. What is the probability that the four people get off at exactly two floors?

Exercise 1.19. An urn contains five red, five black, and five white balls. If three balls are chosen without replacement at random, what is the probability that they are of exactly two different colors?

Exercise 1.20 (A Problem of Tom Sellke). Suppose that cards are picked at random from a full deck of 52 cards.

(a) What is the probability that exactly one jack, one queen, and one king have been picked from the deck when the first ace turns up?
(b) What is the probability that exactly two jacks, two queens, and two kings have been picked from the deck when the second ace turns up?

Exercise 1.21 (Coincidence, Again). Four men throw their watches into the sea, and the sea brings back to each man at random one watch. What is the probability that exactly one man gets his own watch back?

Exercise 1.22. The refrigerators in seven households need repair, and each owner calls the town handyman on a randomly chosen day of the week. What is the probability that the handyman gets at least one day of the week off?

Exercise 1.23. A fair die is rolled thrice. What is the probability that the sum is at least nine but at most 15?

Exercise 1.24. What is the probability that North receives an odd number of spades in his hand in a bridge game? Count zero as even.

Exercise 1.25 (Bad Luck). Jeff and Donna have three children. Two are chosen at random on each day of the week to help with the dishes. What is the probability that at least one child gets chosen every day of the week?

Exercise 1.26 (Check Your Intuition). An urn contains three red and three blue balls. Half the balls are removed at random and then one ball is selected from the rest. What are the chances that this ball will be red?

Exercise 1.27. * A wooden cube with painted faces is sawed up into 1000 little cubes, all of the same size. The little cubes are then mixed up, and one is chosen at random. What is the probability of its having just two painted faces?

Exercise 1.28. One of the numbers 2, 4, 6, 7, 8, 11, 12, and 13 is chosen at random as the numerator of a fraction, and then one of the remaining numbers is chosen at random as the denominator of the fraction. What is the probability of the fraction being in lowest terms?

Exercise 1.29. * **(Logic).** Suppose six customers stand in line at a box office, three with five-dollar bills and three with ten-dollar bills. Suppose each ticket costs 5 dollars, and the box office has no money initially. What is the probability that none of the customers have to wait for change?

Exercise 1.30. * **(Logic).** Eight pawns are placed on eight random squares on a chess board. What is the probability that no two pawns are in the same row or the same column?

Exercise 1.31. A fair coin is tossed n times. What is the probability of at least one head run of length 2 or more if $n = 3, 4, 5, 6$?

Exercise 1.32. Show that it is impossible for the total number of distinct events in an experiment to be 28.

Exercise 1.33. Which is more likely:

(a) obtaining at least one six in six rolls of a fair die
    or
(b) obtaining at least one double six in six rolls of a pair of fair dice?

Remark. This question was posed to Isaac Newton and has some history associated with it.

Exercise 1.34. * **(Coincidence).** Suppose each of n sticks are broken into two pieces, one longer than the other. Then the $2n$ pieces are paired up to make n new sticks. Find the probabilities that the pieces are all paired up in their original order and that each long piece is paired up with some short piece. (There are two events in this problem.)

Exercise 1.35. In a completely dark room with ten chairs, six people come and occupy six chairs at random. What is the probability that at least one of three specific chairs gets occupied?

Exercise 1.36. A group of six men and 12 women are partitioned into six committees of three people each. What is the probability that each committee contains a male member?

Exercise 1.37. * **(The General Shoes Problem).** There are n pairs of shoes of n distinct colors in a closet and $2m$ are pulled out at random from the $2n$ shoes. What is the probability that there is at least one complete pair among the shoes pulled?

Exercise 1.38. Find the probability that, after n rolls of a fair die, the sum of the rolls will be no less than $6n - 1$.

Exercise 1.39. * **(Clever Counting).** n balls from a total of N balls, labeled as $1, 2, \cdots, N$, are taken out from an urn and the label of each written down. Find the probability that the labels form an increasing sequence.

Hint: See the text for a similar example.

Exercise 1.40. * n people are lined up at random for a photograph. What is the probability that a specified set of r people happen to be next to each other?

Exercise 1.41. Mark and his wife, with n other people, are lined up at random for a photograph. What is the probability that they are separated by exactly k people in the photograph?

Exercise 1.42 (Poker). In a five-card poker game, a straight is five cards in a sequence, but not all of the same suit. A straight flush is five cards in a sequence, all of the same suit.

Find the probabilities of obtaining a straight hand or a straight flush.

Exercise 1.43. Calculate the following probabilities in bridge:

(a) Neither North nor South has any spades.
(b) Each of North and South has only clubs and spades.

Exercise 1.44. Calculate the probability that, in bridge, the hand of at least one player is void in a particular suit.

Exercise 1.45. Give a probabilistic proof that the product of any n consecutive positive integers is divisible by $n!$.

Exercise 1.46. * **(The Rumor Problem).** In a town with n residents, someone starts a rumor by saying it to one of the other $n-1$ residents. Thereafter, each recipient passes the rumor on to one of the other residents, chosen at random. What is the probability that by the kth time that the rumor has been told it has not come back to someone who has already heard it?

Exercise 1.47 (Not Equally Likely Sample Points). A couple wants to have children untill they have their first daughter, but they will not have more than three children. Find the probability that they will have more boys than girls.

Exercise 1.48. Prove or disprove:

(a) $A \cup (B \triangle C) = (A \cup B) \triangle (A \cup C)$.
(b) $A \triangle (B \cup C) = (A \triangle B) \cup (A \triangle C)$.
(c) $A \triangle (B \cap C) = (A \triangle B) \cap (A \triangle C)$.

Exercise 1.49 (Use Your Computer). Using a computer, simulate the experiment of tossing a fair coin $n = 50$ times. Perform your simulation 500 times, and count how many times you saw a head run of length three or more, of length four or more, and of length five or more. Before you start your simulation, write down your guesses for what your simulation will actually show. Then, compare the guesses with the actual simulation.

Exercise 1.50 (Use Your Computer). Using a computer, simulate a five-card poker game in which a player gets five of the 52 cards at random. Perform your simulation 500 times, and count how many times you got a straight, and how many times you got two pairs. Compare the theoretical values of their probabilities against your simualtion. (The theoretical value of two pairs is an example in the text; the straight is an exercise above.)

Exercise 1.51 (Use Your Computer). Using a computer, simulate the experiment of rolling a fair die $n = 20$ times. Perform your simulation 500 times, and count how many times you saw each of the six faces show up. Compare the theoretical value of the probability that each face will show up against your simualtion. (The theoretical value is an example in the text.)

References

Basu, D. (1975). Statistical information and likelihood, Sankhyá Ser. A, 37, 1–71.
Berger, J. (1986). *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.
Feller, W. (1968). *Introduction to Probability Theory and Its Applications*, Vol. I, Wiley, New York.
Galambos, J. and Simonelli, I. (1996). *Bonferroni-Type Inequalities with Applications*, Springer-Verlag, New York.
Pitman, J. (1992). *Probability*, Springer-Verlag, New York.
Rao, C.R. (1973). *Linear Statistical Inference and Applications*, Wiley, New York.
Ross, S. (1984). *A First Course in Probability*, Macmillan, New York.
Savage, L. (1954). *The Foundations of Statistics*, Wiley, New York.
Stirzaker, D. (1994). *Elementary Probability*, Cambridge University Press, London.

Chapter 2
The Birthday and Matching Problems

In this chapter, we offer a glimpse into some problems that have earned the status of being classics in counting and combinatorial probability. They have an entertainment value, and they also present some surprises in their solutions and the final answers. The problems we present are generally known as the *birthday problem* and the *matching problem*. For greater exposure to the material in this chapter, we recommend Feller, W. (1968), Diaconis and Holmes (2002), Blom et al. (1994), DasGupta (2005), Mckinney (1966), Abramson and Moser (1970), Diaconis and Mosteller (1989), Barbour and Hall(1984), Barbour et al. (1992), Gani (2004), Ivchenko and Medvedev (1997), Johnson and Kotz (1977), and Karlin and McGregor (1965).

2.1 The Birthday Problem

The birthday problem has earned the status of being a classic in introductory probability. The canonical birthday problem asks the following question

Example 2.1. Suppose n unrelated people are gathered together and that each person has an equal probability of being born on any day of the calendar year. Assuming that a calendar year has 365 days, what is the probability that we will find two or more people in the gathering with the same birthday?

An exact formula is easily found. There are $(365)^n$ possible choices of birthdays for the set of n people. (The assumption of *unrelatedness* is supposed to rule out a priori knowledge of identical birthdays, etc.) Now consider the event where we will *not* be able to find two or more people in the gathering with the same birthday. This is the same as saying that the n people have n distinct birthdays. There are $365 \times 364 \times 363 \times \cdots \times (366-n) = \binom{365}{n} n!$ ways that n people can have n distinct birthdays. Under the assumption of equally likely sample points, we then have, for $n \le 365$,

$$\begin{aligned} p_n &= P(\text{There exist two or more people in the gathering with the same birthday}) \\ &= 1 - \binom{365}{n} n!/(365)^n \end{aligned}$$

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_2,

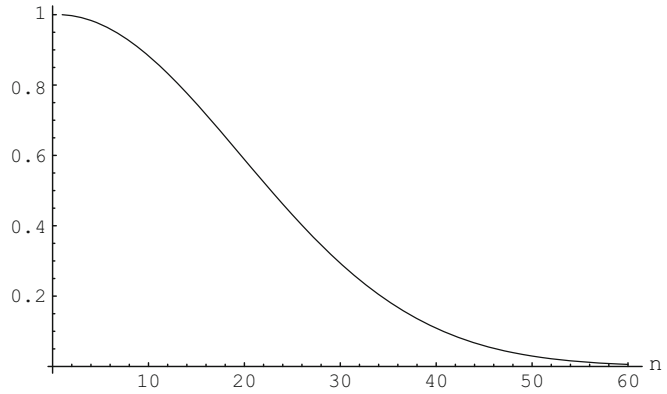


Fig. 2.1 Probability that n unrelated people have n different birthdays

This can be computed exactly for a given value of n, and calculation gives the following values:

n	2	4	5	10	15	20	22	23	30	40	50	60
p_n	.0027	.0163	.0271	.1169	.2529	.4114	.4757	.5073	.7063	.8912	.9703	.9941

We see that, under the assumptions that we have made, it takes *only* 23 people to be more sure than not that there will be people with common birthdays in the gathering. This comes to some as quite a surprise. A plot of $1 - p(n) = P(n$ unrelated people have n different birthdays) is given in Figure 2.1.

*2.1.1 * Stirling's Approximation*

Although p_n can be easily computed (*Mathematica* gave the value .999999 when $n = 98$), a sharp analytical approximation to p_n can be useful for making further deductions. For ease of explanation, we will first derive an approximation nonrigorously. However, even the nonrigorous method will produce a correct approximation!

To do this, writing m for 365,

$$\begin{aligned}
1 - p_n &= \frac{m(m-1)(m-2)\cdots(m-(n-1))}{m^n} \\
&= \frac{m}{m}\frac{m-1}{m}\frac{m-2}{m}\cdots\frac{m-(n-1)}{m} \\
&= 1\left(1-\frac{1}{m}\right)\left(1-\frac{2}{m}\right)\cdots\left(1-\frac{n-1}{m}\right) \\
&= \left(1-\frac{1}{m}\right)^{m\times\frac{1}{m}}\left(1-\frac{2}{m}\right)^{m\times\frac{1}{m}}\cdots\left(1-\frac{n-1}{m}\right)^{m\times\frac{1}{m}} \\
&\approx e^{-\frac{1}{m}}e^{-\frac{2}{m}}\cdots e^{-\frac{n-1}{m}} \\
&= e^{-\frac{1+2+\cdots+n-1}{m}} = e^{-\frac{n(n-1)}{2m}} = e^{-\frac{\binom{n}{2}}{m}}.
\end{aligned}$$

Let us now derive this approximation by using Stirling's approximation to factorials, which says

$$k! \approx e^{-k} k^{k+1/2} \sqrt{2\pi},$$

where the $\approx$ notation means that the ratio converges to one as $k \to \infty$. Applying this separately to 365! and $(365-n)!$, we have

$$1 - p_n = 365!/((365-n)! \times 365^n) \sim e^{-n} (365/(365-n))^{365-n+1/2}.$$

Taking the logarithm and using the approximation $\log(1+x) \approx x - x^2/2$ for small x, we get with a bit of algebra

$$\log\,(1-p_n) \approx (365-n+1/2)[n/(365-n) - n^2/(2(365-n)^2)] - n$$

$$= n/(2(365-n)) - n^2/(2(365-n)) = -\binom{n}{2}/(365-n).$$

Exponentiating, an approximation to p_n is

$$p_n \approx 1 - e^{-\binom{n}{2}/365}$$

if n is moderately large but small compared to 365, so we can write 365 in place of $365-n$ in $\binom{n}{2}/(365-n)$.

If we use this approximation with $n = 23$, we get $p_n \approx .5000$, while, as we saw above, the true value is $p_{23} = .5073$; clearly, the approximation is accurate even for n in the range of $n = 20$.

2.2 The Matching Problem

Example 2.2. Yet another problem in probability theory with celebrity status is the *matching problem*. Some popular variants of it are:

(a) n people throw their hats in the air and the wind brings each of them one hat at random. What is the probability that at least one person gets his own hat back?
(b) At a party of n couples, the men are paired up at random with the women for a dance. What is the probability that at least one man dances with his wife?
(c) One holds two decks of well-shuffled cards, one deck in the right hand and the other deck in the left hand. The cards are picked one by one from the top from each deck. What is the probability that on at least one draw the same card will be picked from both decks?

The mathematical formulation is that n numbers, say $1, 2, \ldots, n$, are arranged in a random manner on a line. If $\pi(i)$ is the number occupying the ith location, what is the probability that for at least one $i, \pi(i) = i$? The problem can be solved by applying the inclusion-exclusion formula.

Define event $A_i = \{\pi(i) = i\}, 1 \le i \le n$. Then, we want to find $P(\cup_{i=1}^n A_i)$. Clearly, $P(A_i) = 1/n\ \forall i$; $P(A_i \cap A_j) = (n-2)!/n!,\ \forall i < j$; $P(A_i \cap A_j \cap A_k) = (n-3)!/n!,\ \forall i < j < k$, etc.

Thus, by the inclusion-exclusion formula,

$$\begin{aligned} P(\cup_{i=1}^n A_i) &= \binom{n}{1}1/n - \binom{n}{2}(n-2)!/n! + \binom{n}{3}(n-3)!/n! - \cdots + (-1)^{n+1}/n! \\ &= 1 - 1/2! + 1/3! - \cdots + (-1)^{n+1}1/n! \\ &= 1 - [1/2! - 1/3! + \cdots + (-1)^n/n!]. \end{aligned}$$

This is an exact formula for all $n \ge 2$. However, there is a simple and elegant approximation that is extremely accurate even for very small n. Recall that the exponential function has the power series expansion

$$e^x = 1 + x + x^2/2! + x^3/3! + \cdots\ \forall x \in \mathcal{R}.$$

Therefore, by using $x = -1$ for large n,

$$e^{-1} \approx 1/2! - 1/3! + \cdots + (-1)^n/n!,$$

and hence $p_n = P(\cup_{i=1}^n A_i) \approx 1 - e^{-1} = .6321$ for large n. For $n = 5$, the exact formula gives $p_n = .6333$, while for $n = 7$, the exact formula already gives $p_n = .6321$, which agrees with the limiting value of the sequence p_n up to four decimal places.

In fact, an elegant generalization of this specific result is available. This is known as the *Poisson approximation to the number of matches*, and the link will be clear when we later define a Poisson distribution. For now, we state this result. The theorem follows from an application of the general Poisson approximation result given in Theorem 6.10 in this text.

Theorem 2.1. *Let $N = N_n$ = number of locations i such that $\pi(i) = i$. Then, for any fixed $k \ge 0$,*

$$\lim_{n\to\infty} P(N = k) = e^{-1}/k!.$$

For example, if 25 couples attend a party and the husbands are paired up at random with the women, then the probability that at least two husbands will dance with their respective wives is approximately $\sum_{k=2}^{25} e^{-1}/k! = .2642$.

2.3 Synopsis

(a) In the *birthday problem*, the probability that n people in a gathering have n distinct birthdays is $\binom{365}{n}n!/(365)^n$. This is $< .5$ if n is 23 or more. This surprises many people.

(b) In the *matching problem*, the probability of at least one match, if there are a total number of n locations, is given by the exact formula

$$1 - [1/2! - 1/3! + \cdots + (-1)^n/n!].$$

Even for n as small as 7, this expression is nearly equal to $1 - e^{-1}$.

(c) In both the birthday problem and the matching problem, useful approximations using more sophisticated techniques are available.

2.4 Exercises

Exercise 2.1. Suppose n unrelated people are gathered together. What is the smallest n for which chances are $>50\%$ that there will be two or more people born in the same calendar month?

Exercise 2.2. * Suppose n unrelated families, defined as the husband, the wife, and one child, are gathered together. What is the smallest n for which chances are $> 50\%$ that there will be two or more *families* completely matched in birthdays (i.e., the two husbands have the same birthday, so do the two wives, and so do the two children)?

Exercise 2.3 (Use Your Computer). Simulate the birthday problem with $n = 30, 60, 100$ people. Perform 500 simulations, and count how many times you got: (a) at least one pair of people with a common birthday; (b) at least two different pairs of people with two distinct common birthdays (e.g., two people born on July 1 and two others born on December 6); (c) at least one set of three people with the same birthday. Which of (a), (b), and (c) was the least frequent?

References

Abramson, M. and Moser, W. (1970). More birthday surprises, Am. Math. Mon., 77, 856–858.

Barbour, A. and Hall, P. (1984). On the rate of Poisson convergence, Math. Proc. Cambridge Philos. Soc., 95, 473–480.

Barbour, A., Holst, L., and Janson, S. (1992). *Poisson Approximation*, Clarendon Press, New York.

Blom, G., Holst, L., and Sandell, D. (1994). *Problems and Snapshots from the World of Probability*, Springer, New York.

DasGupta, A. (2005). The matching, birthday, and the strong birthday problems: A contemporary review, J. Statist. Planning Inference, 130, 377–389.

Diaconis, P. and Holmes, S. (2002). A Bayesian peek at Feller Volume I, Sankhya, Special Issue in Memory of Dev Basu, A. DasGupta, ed., 64(3), Part 2, 820–841.

Diaconis, P. and Mosteller, F. (1989). Methods for studying coincidences, J. Am. Statist. Assoc. 84(408), 853–861.

Feller, W. (1968). *Introduction to Probability Theory and Its Applications*, Wiley, New York.

Gani, J. (2004). Random allocation and urn models, J. Appl. Prob., 41A, 313–320.
Ivchenko, G. and Medvedev, Yu. I. (1997). The contribution of the Russian mathematicians to the study of urn models, in *Probabilistic Methods in Discrete Mathematics*, V. Kolchin, V. Kozlov, Y. Pavlov, and Y. Prokhorov eds., 1–9, VSP, Utrecht.
Johnson, N. and Kotz, S. (1977). *Urn Models and Their Application*, Wiley, New York.
Karlin, S. and McGregor, J. (1965). Ehrenfest urn models, J. Appl. Prob., 2, 352–376.
Mckinney, E. (1966). Classroom notes; generalized birthday problem, Am. Math. Mon., 73(4), 385–387.

Chapter 3
Conditional Probability and Independence

Both conditional probability and independence are fundamental concepts for probabilists and statisticians alike. Conditional probabilities correspond to updating one's beliefs when new information becomes available, a natural human instinct. Independence corresponds to irrelevance of a piece of new information, even when it is made available. Additionally, the assumption of independence can and does significantly simplify development, mathematical analysis, and justification of tools and procedures. Indeed, nearly every key result in probability and statistics was first derived under suitable independence assumptions and then extended to selected cases where independence may be lacking. These two topics together also provide the reader with a supply of fascinating problems and often very pretty solutions.

3.1 Basic Formulas and First Examples

We start with an elementary motivating example.

Example 3.1. Consider the experiment ξ of rolling a fair die twice, and consider the events

$$A = \text{the first roll is 6;}$$
$$B = \text{the sum of the two rolls is 12.}$$

Under the assumption of equally likely sample points, $P(B) = 1/36$. However, if someone were to tell us that the first roll was definitely a six, then intuitively we would feel that in view of this new information, a sum of 12 now seems more likely; indeed, most would say that since the second roll can be any of $1, 2, \cdots 6$, but we have no other information about the second roll, if we knew that the first roll was a six, then the chance that the sum of the two rolls is 12 should be updated to 1/6. We call this updated probability the *conditional probability* of B given A and write it as $P(B|A)$. The conditional probability tells us *among the times that A has already happened how often B also happens*. This motivates the definition of $P(B|A)$ as given below.

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_3,

Definition 3.1. Let A and B be general events with respect to some sample space Ω, and suppose $P(A) > 0$. The conditional probability of B given A is defined as

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

Some immediate consequences of the definition of a conditional probability are the following.

Theorem 3.1.

(a) **(Multiplicative Formula).** *For any two events A and B such that $P(A) > 0$, $P(A \cap B) = P(A)P(B|A)$.*

(b) *For any two events A and B such that $0 < P(A) < 1$, $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$.*

(c) **(Total Probability Formula).** *If $A_1, A_2, \ldots, A_k$ form a* partition *of the sample space Ω (i.e., $A_i \cap A_j = \phi$ for all $i \neq j$, and $\cup_{i=1}^k A_i = \Omega$) and if $0 < P(A_i) < 1$ for all i, then*

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i).$$

Proof. Part (a) is just a restatement of the definition of $P(B|A)$. Part (b) follows from part (a) and the fact that $B = (B \cap A) \cup (B \cap A^c)$, which gives $P(B) = P(B \cap A) + P(B \cap A^c) = P(B|A)P(A) + P(B|A^c)P(A^c)$. For part (c), simply observe that $B = \cup_{i=1}^k (B \cap A_i)$ and therefore $P(B) = \sum_{i=1}^k P(B \cap A_i) = \sum_{i=1}^k P(B|A_i)P(A_i)$.

Theorem 3.2 (Hierarchical Multiplicative Formula). *Let $A_1, A_2, \cdots, A_k$ be k general events in a sample space Ω. Then,*

$$P(A_1 \cap A_2 \cap \ldots ... \cap A_k) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_k|A_1 \cap A_2 \cap \ldots ... \cap A_{k-1}).$$

The proof is a simple exercise and is omitted.

We will now see a spectrum of elementary examples.

Example 3.2. One of the cards from a deck of 52 cards is missing from the deck, but we do not know which one. One card is chosen at random from the remaining 51 cards. We want to find the probability that it is a spade. Define events $A =$ the missing card is a spade; $B =$ the card chosen from the imputed deck is a spade.

Then, by the total probability formula,

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = 12/51 \times 1/4 + 13/51 \times 3/4 = 1/4.$$

Thus, we have the interesting conclusion that although there is nothing like 12.5 spade cards in an imputed deck of 51 cards, $P(B)$ is still 1/4, as it was for the case of a complete deck.

Example 3.3. A certain item is produced in a factory on one of three machines, A, B, or C. The percentages of items produced on machines A, B, and C, are 50%, 30%, and 20% respectively, and 4%, 2%, and 4%, respectively, of their products are defective. We want to know what percentage of all copies of this item are defective.

By defining A_1, A_2, A_3 as the event that a randomly selected item was produced by machine A, B, C, respectively, and defining D as the event that it is a defective item, by the total probability formula, $P(D) = \sum_{i=1}^{3} P(D|A_i)P(A_i) = .04 \times .5 + .02 \times .3 + .04 \times .2 = .034$; i.e., 3.4% of all copies of the item produced in the factory are defective.

Example 3.4. One of two urns has a red and b black balls, and the other has c red and d black balls. One ball is chosen at random from each urn, and then one of these two balls is chosen at random. What is the probability that this ball is red?

If each ball selected from the two urns is red, then the final ball is definitely red. If one of those two balls is red, then the final ball is red with probability 1/2. If none of those two balls is red, then the final ball cannot be red.

Thus, $P(\text{The final ball is red}) = a/(a+b) \times c/(c+d) + 1/2 \times (a/(a+b) \times d/(c+d) + b/(a+b) \times c/(c+d)) = \frac{2ac+ad+bc}{2(a+b)(c+d)}$.

As an example, suppose $a = 99, b = 1, c = 1, d = 1$. Then $\frac{2ac+ad+bc}{2(a+b)(c+d)} = .745$. Although the total percentage of red balls in the two urns is more than 98%, the chance that the final ball selected would be red is just about 75%.

3.2 More Advanced Examples

We will now see a spectrum of examples that have something interesting about them and are at a somewhat more advanced level.

Example 3.5 (An Example Open to Interpretation). This example was given to us in a lecture when this author was a student of Dev Basu.

Mrs. Smith has two children. On a visit to the Smith household, you request a glass of water, and a boy brings it over. What is the probability that Mrs. Smith's other child is a boy?

Some people give the answer that this probability is 1/2. Others argue that with obvious notation, the sample space of the experiment is $\Omega = \{BB, BG, GB, GG\}$, and the required probability is $\frac{P(BB)}{P\{BB,BG,GB\}} = 1/3$.

Actually, the question does not have a unique answer because the experiment to choose the child to carry the glass of water has not been specified. For instance, if Mrs. Smith will always send a girl child with the water if she has a girl, then the correct answer to the question is neither 1/2 nor 1/3 but 1!

Suppose Mrs. Smith chooses one of the two children at random to carry the glass of water if both children are of the same sex and chooses the male child with probability p if the children are of different sex. Then,

$$P(\text{The other child is a boy}|\text{The chosen child is a boy}) = \frac{1/4}{1/4 + p/2} = 1/(2p+1).$$

If $p = .5$, then this is $1/2$. Otherwise, the answer depends on Mrs. Smith's state of mind.

Example 3.6 (A Simple Weather Forecasting Model). Suppose the weather on any day is dry or wet, and with probability p it is the same as whatever the weather was on the day before and with probability $1 - p$ it is different from what it was the day before.

Suppose it is dry today, and let $p_n = P(\text{It will be dry } n \text{ days from today})$, with the convention that $p_0 = 1$. Then, by the total probability formula,

$$p_n = p \times p_{n-1} + (1-p) \times (1 - p_{n-1}) = (2p-1)p_{n-1} + (1-p), n \geq 1.$$

The interesting question is what can we say about the state of the weather a long time into the future; i.e., as $n \to \infty$? It turns out that a sequence satisfying the *recursion relation* given above must have a limit. If we call this limit θ, then it follows that

$$\theta = (2p-1)\theta + (1-p) \Rightarrow \theta = 1/2;$$

i.e., with this simple model for weather forecasting, we can only do as well as a coin toss for predicting into the distant future. The reader should prove that the limit θ of the sequence p_n indeed exists by proving that the sequence is a *Cauchy sequence*.

Example 3.7 (Does the First Chooser Have an Advantage?). Suppose that in a lottery there are n tickets, of which a prespecified set of m tickets will win a prize. There are n players, and they will choose one ticket at random successively from the available tickets. We want to calculate the probability that the ith player (that is, the player to choose his ticket after $i-1$ players have already chosen their tickets) will win a prize for a general i.

First, obviously, the probability that the player to buy the first ticket wins a prize is m/n. The probability that the player to buy next would buy a winning ticket depends on whether the first player had bought a winning ticket or not. Thus, by the total probability formula,

$$P(\text{The second player wins a prize}) = (m-1)/(n-1) \times m/n + m/(n-1) \times (1 - m/n) = m/n$$

with a little algebra. Next, the probability that the third player wins a prize can be found by formally defining $A_i = \text{player } i \text{ wins a prize}$ and on using

$$P(A_3) = P(A_3A_2A_1) + P(A_3A_2A_1^c) + P(A_3A_2^cA_1) + P(A_3A_2^cA_1^c)$$

$= (m/n) \times (m-1)/(n-1) \times (m-2)/(n-2) + (1-m/n) \times m/(n-1) \times$ (m-1)/ $(n-2) + m/n \times (1-(m-1)/(n-1)) \times (m-1)/(n-2) + (1-m/n) \times (1-m/(n-1)) \times m/(n-2) = m/n$, again on doing the algebra.

In general, by conditioning on how many of the first $i-1$ players won a prize and applying the total probability formula, it does follow that for the ith player, for a general i, the probability of selecting a winning ticket is indeed m/n; the probability does not depend on i. That is, there is no advantage in selecting one's ticket early.

Example 3.8 (A Clever Conditioning Argument). Coin A gives heads with probability s and coin B gives heads with probability t. They are tossed alternately, starting off with coin A. We want to find the probability that the first head is obtained on coin A. If we find this probability to be $>.5$, we will have to agree that in this example (contrary to the preceding one) starting first has an advantage!

We find this probability by conditioning on the outcomes of the first two tosses; more precisely, define

$$A_1 = \{H\} = \text{the first toss gives } H;\ A_2 = \{TH\};\ A_3 = \{TT\}.$$

Also, let $A =$ the first head is obtained on coin A.

One of the three events A_1, A_2, A_3 must happen, and they are also mutually exclusive. Therefore, by the total probability formula,

$$\begin{aligned} P(A) &= \sum_{i=1}^{3} P(A_i)P(A|A_i) = s \times 1 + (1-s)t \times 0 + (1-s)(1-t)P(A) \\ &\Rightarrow P(A) = s/[1-(1-s)(1-t)] = s/(s+t-st). \end{aligned}$$

As an example, let $s = .4, t = .5$. Note that coin A is biased against heads. Even then, $s/(s+t-st) = .57 > .5$. We see that, contrary to our previous example, now *there is an advantage in starting first*.

3.3 Independent Events

Independence of events corresponds to lack of probabilistic information in one event A about some other event B; i.e., even if knowledge that some event A has occurred was available, it would not cause us to modify the chances of the event B. Here is a simple example.

Example 3.9. Consider the experiment ξ of rolling a fair die twice, and let

$$\begin{aligned} A &= \text{the first roll is an even number;} \\ B &= \text{the sum of the two rolls is an even number.} \end{aligned}$$

Assuming that each sample point has an equal probability $1/36$, $P(B) = P(A) = 1/2$, and $P(A \cap B) = 1/4$, by direct counting. Therefore, $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/4}{1/2} = 1/2$, implying that $P(B|A) = P(B)$, i.e., the knowledge that A has occurred

did not cause us to alter the chance of occurrence of B. In such a case, we say A and B are independent events.

Definition 3.2. Two events A and B are called independent if $P(B|A) = P(B) \Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(A \cap B) = P(A)P(B)$.

Caution. Disjointness of two events should not be confused with their independence. In fact, if A and B are disjoint events with nonzero probabilities, then they *cannot* be independent. After all, if they are disjoint, then as soon as A occurs, B becomes impossible! Thus, there is a lot of information in A about B and they cannot be independent.

In applications, we often have to deal with more than two events simultaneously. But we may still want to know if they are independent. Fortunately, the concept of independence extends in a natural way to any number of events. The idea is that no subcollection of the events should give any probabilistic information about any other nonoverlapping subcollection of events. A mathematically equivalent way to state that is the following.

Definition 3.3. A collection of events $A_1, A_2, \cdots, A_n$ are said to be *mutually independent* (or just independent) if for each $k, 1 \leq k \leq n$, and for any k of the events, $A_{i_1}, \cdots, A_{i_k}, P(A_{i_1} \cap \cdots A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$. They are called *pairwise independent* if this property holds for $k = 2$.

We will now look at a number of illustrative examples.

Example 3.10. Here is an example where we can sense intuitively that the events under consideration are independent. Suppose a card is chosen at random from a deck of 52 cards, and let A and B respectively be the events that the card is an ace and that the card is a spade. Then, $P(A) = 1/13, P(B) = 1/4$, *and* $P(A \cap B) = 1/52$ because there are four aces and 13 spades in a full deck. Clearly, then, $P(A \cap B) = P(A)P(B)$, and so A and B are independent events.

Example 3.11. Suppose a fair die is rolled twice, and let A and B be the events that the sum of the two rolls is 7 and that the first roll is j, where j is any given number $1, 2, \cdots,$ or 6. Then $P(A) = 6/36 = 1/6, P(B) = 1/6$, and $P(A \cap B) = 1/36$. So A and B are independent events.

Now, change the event A to the event that the sum of the two rolls is 8. Then, A and B are not necessarily independent events. Why? For instance, with $j = 1$, $P(A|B) = 0$, but the unconditional probability $P(A)$ is not zero. Therefore, A and B cannot be independent events.

Example 3.12. In some applications, we assume that certain physical systems or processes behave independently and transfer that physical assumption into independence of appropriate events. Suppose Jack and his wife, Sarah, have three cars, and on any given winter morning, the cars run, *independently* of each other, with probabilities .9, .95, and .99, respectively.

Because we assume that the cars run or not independently of each other, we can easily calculate probabilities such as

1. $P(\text{At least one of the three cars runs}) = 1 - P(\text{None of the three cars runs})$ $=1-.1 \times .05 \times .01 = .99995$ or
2. $P(\text{All three cars run}) = .9 \times .95 \times .99 = .84645$.

Example 3.13 (System Reliability). Here is an example of independence applied to system reliability. Suppose a long-haul airplane has four engines and needs three or more engines to work in order to fly. Another airplane has two engines and needs one engine to fly. We assume that the engines are independent and suppose each has a constant probability p of staying functional during a flight.

Then,

$$P(\text{The first airplane can fly}) = p^4 + 4p^3(1-p) = 4p^3 - 3p^4$$

and

$$P(\text{The second airplane can fly}) = p^2 + 2p(1-p).$$

We want to know which plane is safer. The second plane is safer if and only if $4p^3 - 3p^4 < p^2 + 2p(1-p)$, and this happens to be true for all $p, 0 < p < 1$. Thus, the second plane is always safer. Note that we could not have reached this conclusion if the engines on the airplanes were not assumed to be independent.

Example 3.14 (Lotteries). Although many people buy lottery tickets out of an expectation of good luck, probabilistically speaking, buying lottery tickets is usually a waste of money. Here is an example. Suppose that in a weekly state lottery five of the numbers $00, 01, \ldots, 49$ are selected without replacement at random and someone who holds exactly those numbers wins the lottery. Then, the probability that someone holding one ticket will be the winner in a given week is $\frac{1}{\binom{50}{5}} = 4.72 \times 10^{-7}$. Suppose this person buys a ticket every week for 40 years. Then, the probability that he will win the lottery in at least one week is $1 - (1 - 4.72 \times 10^{-7})^{52 \times 40} = .00098 < .001$, still a very small probability. We assumed in this calculation that the weekly lotteries are all mutually independent, a reasonable assumption. The calculation would fall apart if we did not make this independence assumption.

Example 3.15. Peter and Karen take turns, starting with Karen, rolling a fair die. The first to obtain a six wins. Then,

$$P(\text{Karen wins}) = 1/6 + 5/6 \times 5/6 \times 1/6 + 5/6 \times 5/6 \times 5/6 \times 5/6 \times 1/6 + \cdots$$

$= 1/6(1 + \sum_{i=1}^{\infty}(5/6)^{2i}) = 1/6 \times (1 + \frac{25/36}{1-25/36}) = 6/11 > .5$. Thus, Karen does have an advantage due to starting first. We have assumed in this calculation that the successive rolls of the die are mutually independent.

Example 3.16 (An Interesting Example due to Emanuel Parzen). Consider two dice, with the side probabilities being $p_j, 1 \leq j \leq 6$ for the first die and $q_j, 1 \leq j \leq 6$ for the second die. That is, we are just assuming that these are two arbitrarily loaded dice. The question is whether we can choose p_j, q_j in any way whatsoever such that the sum of the numbers obtained on tossing the dice once each has an equal probability of being any of $2, 3, \cdots 12$. The answer, interestingly, is that we cannot choose p_j, q_j in any way at all to make this happen.

To sketch the proof, suppose we could. Then, since the sums of 2 and 12 will have equal probabilities, we must have $p_1 q_1 = p_6 q_6 \Rightarrow q_1 = p_6 q_6 / p_1$. It follows, after some algebra, on using this that $(p_1 - p_6)/(q_1 - q_6) \leq 0 \Rightarrow p_1 q_1 + p_6 q_6 \leq p_1 q_6 + p_6 q_1$. But this means that

$$P(\text{The sum is 7}) \geq p_1 q_6 + p_6 q_1 \geq p_1 q_1 + p_6 q_6 = P(\text{The sum is 2 or 12}),$$

a contradiction, because by assumption $P(\text{The sum is 2 or 12})$ is supposed to be twice the probability that the sum is 7. Hence, we cannot construct two dice in any way to make the sum have an equal probability of taking the values $2, 3, \ldots 12$.

3.4 Bayes' Theorem

It is not uncommon to see the conditional probabilities $P(A|B)$ and $P(B|A)$ be confused with each other. Suppose that in some group of lung cancer patients we see a large percentage of smokers. If we define B to be the event that a person is a smoker and A to be the event that a person has lung cancer, then all we can conclude is that in our group of people $P(B|A)$ is large. But we cannot conclude from just this information that smoking increases the chance of lung cancer; i.e., that $P(A|B)$ is large. In order to calculate a conditional probability $P(A|B)$ when we know the *other* conditional probability $P(B|A)$, a simple formula known as *Bayes' theorem* is useful. The formula is named after the Reverend Thomas Bayes, who (essentially) obtained this formula in the eighteenth century. Here is a statement of a general version of Bayes' theorem.

Theorem 3.3. *Let $\{A_1, A_2, \cdots, A_m\}$ be a partition of a sample space Ω. Let B be some fixed event. Then*

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}.$$

Proof. By the definition of conditional probability, $P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}$ by the multiplicative formula and the total probability formula.

We will now show a number of examples illustrating the use of Bayes' theorem.

Example 3.17 (Inaccurate Blood Tests). A certain blood test for a disease gives a positive result 90% of the time among patients having the disease. But it also gives a positive result 25% of the time among people who do not have the disease. It is believed that 30% of the population has this disease. What is the probability that a person with a positive test result indeed has the disease?

In medical terminology, the 90% value is called the *sensitivity* of the test, and $100 - 25 = 75\%$ is called the *specificity* of the test. Often, the sensitivity and the specificity would be somewhat higher than what they are in this example.

Define

$$A = \text{the person has the disease;}$$

$$B = \text{the blood test gives a positive result for the person.}$$

Then, by Bayes' theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{.9 \times .3}{.9 \times .3 + .25 \times (1 - .3)} = .607.$$

Before the test was given, the physician had the *a priori* probability of 30% that the person has the disease. After a blood test came out positive, the physician has the *posterior* probability of 60.7% that the person has the disease. If the physician wants to be more sure that the person has the disease, then she will give some other test or repeat the blood test for another independent confirmation of this positive result.

Example 3.18 (Multiple-Choice-Exams). Suppose that the questions in a multiple-choice exam have five alternatives each, of which a student picks one as the correct alternative. A student either knows the truly correct alternative with probability .7 or he randomly picks one of the five alternatives as his choice. Suppose a particular problem was answered correctly. We want to know the probability that the student really knew the correct answer. Define

$$A = \text{the student knew the correct answer;}$$

$$B = \text{the student answered the question correctly.}$$

We want to compute $P(A|B)$. By Bayes' theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{1 \times .7}{1 \times .7 + .2 \times .3} = .921.$$

Before the student answered the question, our probability that he would know the correct answer to the question was .7, but once he answered it correctly, the posterior probability that he knew the correct answer increased to .921. This is exactly what Bayes' theorem does: it updates our *prior* belief to the *posterior* belief when evidence becomes available.

Example 3.19 (Clever Conditioning Argument). Lafayette is connected to Gary by two different roads, and Gary is connected to Chicago by two different roads. Also, Lafayette is connected to Chicago by Amtrak. Each road and the Amtrak service are operational on a winter day with probability .9, mutually independently. If Mark was able to travel from Lafayette to Chicago on a given winter day, what is the probability that Amtrak was operational? Define

A_i = ith road connection between Lafayette and Chicago is operational, $i = 1, 2, 3, 4$;
A_5 = Amtrak is closed;
B = Mark was able to travel from Lafayette to Chicago.

Then,

$$P(A_5|B) = \frac{P(A_5 \cap B)}{P(B)} = \frac{P(A_5)P(A_1 \cup A_2)P(A_3 \cup A_4)}{P(A_5^c) + P(A_5)P(A_1 \cup A_2)P(A_3 \cup A_4)}$$
$$= \frac{.1 \times (1 - .1^2) \times (1 - .1^2)}{.9 + .1 \times (1 - .1^2) \times (1 - .1^2)} = .098.$$

Therefore, the probability that Amtrak was operational given that Mark was able to travel is $1 - .098 = .902$.

Example 3.20 (A Counterintuitive Result). Suppose Jeff hits the bull's eye 60% of the times he shoots at it, and Jen hits the bull's eye 90% of the times she shoots at it. Suppose both shoot simultaneously and only one hits. What is the probability that it was Jen?

Let A = Jen hit the bull's eye; B = exactly one of Jeff and Jen hit the bull's eye. Then,

$$P(A|B) = .9 \times .4/(.9 \times .4 + .1 \times .6) = .536,$$

which is *lower* than $.9/(.9 + .6) = .6$. Often, people give the .6 value as the answer to the question. It would be helpful for the reader to think about why the intuitive result is not correct.

Example 3.21. 75% of Democrats and 25% of Republicans are pro-choice. In the population, 48% are Democrats and 52% are Republicans. If Cathy is pro-life, what is the probability that she is a Republican?

Let A = Cathy is a Republican; B = Cathy is pro-life. Then, by Bayes' theorem,

$$P(A|B) = .75 \times .52/(.75 \times .52 + .25 \times .48) = .765.$$

Note that it is slightly larger than .75. This is because there are slightly more Republicans than Democrats in this population.

3.5 Synopsis

(a) The conditional probability of B given A is defined as

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

assuming that $P(A) > 0$.

(b) The multiplicative formulas say that

$$P(A \cap B) = P(A)P(B|A).$$

More generally,

$$P(\cap_{i=1}^{k} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_k|A_1 \cap A_2 \cap \cdots \ldots \cap A_{k-1}).$$

(c) The total probability formulas say that

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

More generally, $P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$ if $A_1, \ldots, A_k$ form a partition of the sample space Ω.

(d) Bayes' theorem tells you how to find $P(A|B)$ if you know $P(B|A)$, $P(B|A^c)$, and $P(A)$. Bayes' theorem says

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

More generally, $P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}$ if $A_1, \ldots, A_m$ form a partition of the sample space Ω.

3.6 Exercises

Exercise 3.1. A fair die is rolled twice. What is the probability that at least one of the two rolls is a six if the sum of the two rolls is at least 10?

Exercise 3.2. Jen will call Cathy on Saturday with a 60% probability. She will call Cathy on Sunday with an 80% probability. The probability that she will call on neither of the two days is 10%. What is the probability that she will call on Sunday if she calls on Saturday?

Exercise 3.3. On a table there is a double-headed coin and there is a fair coin. One of them was randomly chosen and tossed, and the outcome was a head. What is the probability that the lower side of this coin is a tail?

Exercise 3.4. Suppose $P(A) = P(B) = .9$. Give a useful lower bound on $P(B|A)$.

Exercise 3.5. Two distinct cards are drawn one at a time from a deck of 52 cards. The first card chosen is the ace of spades. What is the probability that the second card is neither an ace nor a spade?

Exercise 3.6. On 10% of the days, Sam's computer does not work. If the computer does not work, then he turns on the television, and independently of the computer, on 30% of the days he finds nothing interesting on television. If he finds nothing interesting on television, he calls his wife at work. Independently of television programming and the computer, on 75% of all occasions she does not pick up the phone. In that case, Sam goes to bed at 8:00 PM. On what percentage of days does Sam go to bed at 8:00 PM?

Exercise 3.7. What is the probability that four cards drawn at random from a deck of cards are of different denominations if they are of different suits?

Exercise 3.8. In an urn, there are three black balls, four green balls, and five red balls. Two balls are randomly drawn from the urn and are found to be of the same color. What is the probability that they are both red?

Exercise 3.9. North and South have all the aces between them in a bridge game. What is the probability that one has one ace and the other has three?

Exercise 3.10. Three independent proofreaders will read a manuscript. The probabilities that a particular error will be detected by them are respectively $.92, .85$, and $.95$. What is the probability that the error will go undetected?

Exercise 3.11. * Two numbers are chosen without replacement from $1, 2, \ldots 100$. What is the probability that the smaller number is greater than 20 if the larger number is smaller than 60?

Exercise 3.12. Sam, Fred, and Vishy in that order take turns shooting at a bull's eye. Their success rates are 20%, 30%, and 40%. What are the probabilities of each making the first hit at the bull's eye?

Exercise 3.13 (Communication Channel). A signal passes through three filtering networks. The signal can be a $+$ signal or a $-$ signal. Each network has a probability p of transmitting the signal as it was received and probability $1 - p$ of transmitting it as the wrong signal. Assume that the networks function independently. If a signal is received as a $+$, what is the probability that it was sent as a $+$? You may assume that the prior probability is 50% that it was sent as a $+$.

Exercise 3.14. * In a country, 60% of parents have one child, 30% have two children, and 10% have three children. A randomly chosen child turns out to be a boy. What is the probability that he has no older brother?

Exercise 3.15. * A fair die is tossed repeatedly until every face has occurred at least once. Give a clever proof and a direct proof that the probability that 1 is the last face to show up is $\frac{1}{6}$.

Exercise 3.16. * The probability that a coin will show all heads or all tails when tossed four times is .25. What is the probability that it will show two heads and two tails?

Exercise 3.17. Sam tosses two fair dice and Fred tosses one. Sam wins if his sum is at least twice as much as Fred's score. What is the probability that Sam wins?

Exercise 3.18. A statistics department has five assistant professors, two associate professors, and six full professors. Two of these 13 faculty members are chosen at random. What is the probability that the junior member is an assistant professor if the senior member is a full professor? It is assumed that we know that the two chosen are always of different ranks.

Exercise 3.19. * **(Lottery).** In a state lottery, five numbers from $00, 01, \ldots, 49$ are selected at random. Someone holding exactly those five numbers wins. For how many weeks must a couple buy tickets to have a 5% probability that at least one of them wins at least once?

Exercise 3.20. A true-false question will be posed to a couple on a game show. The husband and the wife each pick the correct answer with probability p. Should the couple decide to let one of them answer the question or decide that they will give the common answer if they agree and give one of the two answers at random if they disagree?

Exercise 3.21. Of the paintings in a certain gallery, 25% are not original. A certain collector makes an error in judging an item's authenticity 15% of the time, whether the painting is an original or not. If she purchases an item believing that it is an original, what is the probability that it is a fake?

Exercise 3.22. You have requested that a friend water your plant while you are on vacation for a week. Without water, it will die with probability x, and with water it will die with probability y. On returning, you find your plant dead. What is the probability that your friend forgot to water it?

Exercise 3.23. * **(Conditional Independence).** Events A and B are called *conditionally independent* given C if $P(A \cap B|C) = P(A|C)P(B|C)$.

(a) Give an example of events A, B, C such that A and B are not independent but are conditionally independent given C.
(b) Give an example of events A, B, C such that A and B are independent but are not conditionally independent given C.

Exercise 3.24 (Polygraphs). Polygraphs are routinely administered to job applicants for sensitive government positions. Suppose someone actually lying fails the polygraph 90% of the time but someone telling the truth also fails the polygraph 15% of the time. If a polygraph indicates that an applicant is lying, what is the probability that he is in fact telling the truth? Assume a general *prior* probability p that the person is telling the truth.

Exercise 3.25. A gambler has in his pocket a fair coin and a two-headed coin. He selects one of the coins at random; when he flips it, it shows heads.

(a) What is the probability that it is the fair coin?
(b) Suppose that he flips the same coin a second time and again it shows heads. Now what is the probability that it is the fair coin?
(c) Suppose that he flips the same coin a third time and it shows tails. Now what is the probability that it is the fair coin?

Exercise 3.26 (Casino Slot Machines). A typical slot machine in a casino has three wheels, each marked with 20 symbols spaced equally around the wheel. The machine is constructed so that on each play the three wheels spin independently, and each wheel is equally likely to show any one of its 20 symbols when it stops spinning. On the central wheel, nine out of the 20 symbols are bells, while there is only one bell on the left wheel and one bell on the right wheel. The machine pays out the jackpot only if the wheels come to rest with each wheel showing a bell.

(a) Find the probability of hitting the jackpot.
(b) Find the probability of getting two bells but not the jackpot.

Exercise 3.27. Three independent events occur with probabilities .5, .75, and .9, respectively. Given that two of the three occurred, what is the probability that the second event did not occur?

Exercise 3.28. * **(Empty cells).** Three balls will be distributed independently into one of three cells. For each ball, the probabilities that the ball will be dropped into the three cells are x, y, and $1-x-y$, respectively. If we know that exactly one cell remains empty, what is the probability that it is the first cell?

Exercise 3.29. * **(Random Matrix).** The diagonal elements a, c of a 2×2 symmetric matrix are chosen independently at random from $1, 2, \ldots, 5$, and the off-diagonal element is chosen at random from $1, \ldots \min(a, c)$. Find the probability that the matrix is nonsingular.

Exercise 3.30. A, B, C, D are independent events, each with probability .5. What is the probability that at least one of A, B, C, D happens?

Exercise 3.31. * **(Craps).** You are to roll a pair of fair dice. If you get a sum of 7 or 11, you win; if you get a sum of 2, 3, or 12, you lose. If the sum k is none of these, you keep rolling the pair of dice until you again get a sum of k or a sum of 7, whichever happens first. If it is 7 that occurs first, you lose; if it is k that occurs first, then you win. Show that the probability that you will win in this game is .493.

Exercise 3.32 (Yahtzee). Five fair dice are to be rolled. Find the probability of getting a full house, which is three rolls of one number and two rolls of some other number.

Exercise 3.33 (The Parking Problem). At a parking lot, there are 12 spaces arranged in a row. A man observed that there were eight cars parked and that the four empty spaces were adjacent to each other. Given that there are four empty spaces, is this arrangement surprising?

Exercise 3.34 (Another Parking Problem). A car is parked among N cars in a row, not at either end. On his return, its owner finds that exactly r of the N spaces are still occupied. What is the probability that both of his neighboring spaces are empty?

Exercise 3.35 (Use Your Computer). Simulate the lottery problem in which three numbers are picked from $00, 01, 02, \ldots, 24$, and a player with exactly those three numbers is the winner. Perform the simulation 500 times, and count (a) how many times a fixed player, say one holding the numbers 5, 10, and 15, won, and (b) how many times the fixed player matched two of the three winning numbers.

Exercise 3.36 (Use Your Computer). Simulate the airplane reliability problem, taking it to be a three-out-of-four system and taking the reliability of each individual engine to be .99. Perform the simulation 1000 times, and count how many times, if any, the plane would be forced to make an emergency landing.

Chapter 4
Integer-Valued and Discrete Random Variables

In this chapter, we introduce the concept of random variables and their distributions. In some sense, the entire subject of probability and statistics is about distributions of random variables. Random variables, as the very name suggests, are quantities that vary over time or from individual to individual, and the reason for the variability is some underlying random process. We try to understand the behavior of a random variable by analyzing the probability structure of that underlying random mechanism or process. Random variables, like probabilities, originated in gambling. Therefore, the random variables that come to us *more naturally* are integer-valued random variables; e.g., the sum of the two rolls when a die is rolled twice. Integer-valued random variables are special cases of what are known as discrete random variables. We study integer-valued and discrete random variables and their basic properties in this chapter. Random variables with an intrinsically more abstract and complex structure will be studied after we introduce *probability density functions*.

4.1 Mass Function

We start with a mathematical formulation for a random variable. An example will help motivate the definition.

Example 4.1. Consider the experiment ξ of rolling a fair die twice. The sample space Ω of this experiment has the 36 sample points $\Omega = \{11, 12, 13, \ldots, 64, 65, 66\}$. Consider now the sum of the two rolls, and call it X. We realize immediately that the value of X depends on which sample point prevails when the experiment ξ is conducted; e.g., if $\omega = 11$ prevails, then $X = 2$, but if $\omega = 65$ prevails, then $X = 11$. That is, X is a *function of* ω. Indeed, this is how random variables are defined. Although in general X can take values in esoteric spaces, we will confine ourselves to the case where X is real-valued and then quickly specialize even further to the case where X is integer-valued, or more generally *discrete*.

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_4,

Here is the definition of a real-valued random variable.

Definition 4.1. Let Ω be a sample space corresponding to some experiment ξ and let $X : \Omega \to \mathcal{R}$ be a function from the sample space to the real line. Then X is called a *random variable*.

Let us see some simple illustrative examples of random variables.

Example 4.2. Let ξ be the experiment of tossing a coin twice and let X be the number of times in the two tosses that a head is obtained. Denote the four sample points as $HH = \omega_1, HT = \omega_2, TH = \omega_3, TT = \omega_4$. Then, $X(\omega_1) = 2, X(\omega_2) = X(\omega_3) = 1, X(\omega_4) = 0$. This is the function that formally defines the words "number of times in the two tosses that a head is obtained."

Example 4.3. Let ξ be the experiment of rolling a die twice, and let X be the sum of the two rolls. Denoting the sample points as $11 = \omega_1, 12 = \omega_2, 13 = \omega_3, \ldots, 66 = \omega_{36}$, we have $X(\omega_1) = 2, X(\omega_2) = 3, X(\omega_3) = 4, \ldots, X(\omega_{36}) = 12$. This then is the function that formally defines the words "sum of the two rolls."

From the point of view of understanding the behavior of a random variable, the important thing is to know the probabilities with which X takes its different possible values. If we could ascertain these probabilities, then we could exploit them to our advantage. For example, if we have to decide whether to bet fair money that the sum of two rolls of a fair die will be 8 or more, we will surely decide against it if we know that the chances of it happening are less than 50%. This motivates the definition of the distribution of a random variable.

Definition 4.2. Let $X : \Omega \to \mathcal{R}$ be a discrete random variable taking a finite or countably infinite number of values $x_1, x_2, x_3, \ldots$. The probability distribution or the *probability mass function* (pmf) of X is the function $p(x) = P(X = x), x = x_1, x_2, x_3, \ldots$, and $p(x) = 0$ otherwise.

It is common not to explicitly mention the phrase "$p(x) = 0$ otherwise," and we will generally follow this convention. Some authors use the phrase *mass function* instead of *probability mass function*.

Remark. For any pmf, one must have $p(x) \geq 0$ for any x, and $\sum_i p(x_i) = 1$. The second property articulates the fact that something has to happen; i.e., X *must* take one of the values $x = x_1, x_2, x_3, \ldots$. Thus, any function satisfying these two properties for some set of numbers $x_1, x_2, x_3, \ldots$ is a valid pmf. They may not correspond to interesting random variables, but they are all valid pmfs.

Caution. It is important to know that not all random variables that we study in applications are discrete. There are plenty of random variables that arise very naturally in applications and take *uncountably many* possible values. A simple example is the value of a *random fractional number*. Such a variable, by its very definition, can take all values in the unit interval $[0, 1]$, so it is not discrete. Describing the behavior of such random variables probabilistically requires tools that are different from pmfs. We do not discuss them in this chapter. But it may be helpful to just mention the basic idea of how we deal with such random variables, which we call *continuous random variables*. Let us see the random fraction example.

Example 4.4. Consider the sample space $\Omega = [0, 1]$, and define a random variable X by the function $X : \Omega \to \mathcal{R}$ as $X(\omega) = \omega, \omega \in \Omega$. If we equip Ω with a *uniform probability measure* P (i.e., $P(a \le \omega \le b) = b - a$ for $0 \le a \le b \le 1$), then, because $X(\omega) = \omega$ itself, we also have $P(a \le X \le b) = b - a, 0 \le a \le b \le 1$. In particular, if we define a function $f(x)$ on [0,1] as $f(x) \equiv 1$, then we have

$$P(a \le X \le b) = b - a = \int_a^b f(x)dx.$$

The function $f(x)$ is called the *probability density function* (pdf) of X. It plays the role of the pmf for discrete random variables, but $f(x)$ *does not* mean $P(X = x)$. Indeed, by choosing $a = b = x$ in the above, we see that for these continuous random variables, $P(X = x)$ is always zero! There are so many possibilities for X that any one specific possible value is infinitely unlikely! We will discuss continuous random variables and pdfs in great detail in later chapters. For now, it is important to know that they are there and that we analyze them probabilistically by using pdfs rather than pmfs.

4.2 CDF and Median of a Random Variable

A second important definition is that of a *cumulative distribution function* (CDF). The CDF gives the probability that a random variable X is less than or equal to any given number x. In other words, the CDF measures the probability that has been accumulated up to and including a given number x, hence the name *cumulative* distribution function. It is important to understand that the notion of a CDF is universal to all random variables; it is not limited to only the discrete ones or only the continuous ones. However, operationally, pmfs are simpler to work with when we have discrete random variables; the examples below will show that. So, although the notion of a CDF provides a common and unified background for treatment of all types of random variables, we tend to depend more on the pmf for discrete random variables.

Definition 4.3. The *cumulative distribution function* (CDF) of a random variable X is the function $F(x) = P(X \le x), x \in \mathcal{R}$.

Let us see some examples.

Example 4.5. Let ξ be the experiment of tossing a fair die twice and let X be the number of heads obtained. Then X takes the possible values $x_1 = 0, x_2 = 1, x_3 = 2$. Also, $P(X = 0) = P(TT) = 1/4, P(X = 1) = P(\{HT, TH\}) = 1/2$, and $P(X = 2) = P(HH) = 1/4$. We then have the following pmfs for X:

x	0	1	2
$p(x)$	.25	.5	.25

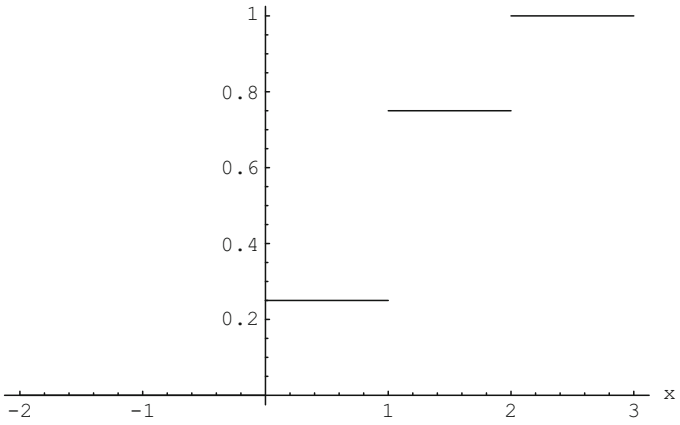


Fig. 4.1 CDF of the number of heads in two tosses of a fair coin

As regards the CDF of X, since X does not take any negative values, the CDF $F(x)$ is zero for any negative x. However, X takes the value $x = 0$ with a positive probability, namely .25. Thus, as soon as x reaches the zero value, the CDF $F(x)$ takes a jump and becomes equal to .25. Then, between $x = 0$ and $x = 1$, X does not take any other values, and no new probability is accumulated. So the CDF stays stuck at the .25 value until x reaches the value $x = 1$, and now it takes another jump of size .5, which is the probability that $X = 1$. The next jump is at $x = 2$, when the CDF takes another jump of size .25, and thereafter the CDF takes no further jumps. In symbols, the CDF $F(x)$ in this example is the jump function

$$\begin{aligned} F(x) &= 0 \text{ if } x < 0 \\ &= .25 \text{ if } 0 \le x < 1 \\ &= .75 \text{ if } 1 \le x < 2 \\ &= 1 \text{ if } x \ge 2. \end{aligned}$$

A plot of the CDF is helpful for understanding and is given in Figure 4.1.

It is clear that because this CDF increases by jumps, it does not attain all values between 0 and 1. For example, there is no x at which $F(x) = .5$. If there was, that value could stake a claim to splitting the distribution into two halves, 50% of the probability below and 50% above. However, there do exist values x such that $P(X \le x) \ge .5$, and at the same time $P(X \ge x)$ is also $\ge .5$. Such a number x is called a *median* of the distribution. We define it formally.

Definition 4.4. Let X have the CDF $F(x)$. Any number m such that $P(X \le m) \ge .5$ and also $P(X \ge m) \ge .5$ is called a median of F, or equivalently a median of X.

Remark. The median of a random variable *need not be* unique. A simple way to characterize all the medians of a distribution is available.

Proposition. Let X be a random variable with the CDF $F(x)$. Let m_0 be the first x such that $F(x) \geq .5$, and let m_1 be the last x such that $P(X \geq x) \geq .5$. Then, a number m is a median of X if and only if $m \in [m_0, m_1]$.

A proof of this uses property (c) below of an arbitrary CDF, known as the *right continuity* of a CDF. We will omit the proof.

The CDF of any random variable satisfies a set of properties. Conversely, any function satisfying these properties is a valid CDF; i.e., it will be the CDF of some appropriately chosen random variable. These properties are given in the next result.

Theorem 4.1. *A function $F(x)$ is the CDF of some real-valued random variable X if and only if it satisfies all of the following properties:*

(a) $0 \leq F(x) \leq 1 \ \forall x \in \mathcal{R}$.
(b) $F(x) \to 0$ *as* $x \to -\infty$, *and* $F(x) \to 1$ *as* $x \to \infty$.
(c) *Given any real number* a, $F(x) \downarrow F(a)$ *as* $x \downarrow a$.
(d) *Given any two real numbers* $x, y, x < y$, $F(x) \leq F(y)$.

Remark. The only part of this theorem that is not obvious is part (c). The best proof of part (c) just uses the fact that if a_n is a sequence decreasing to a real number a, then the intervals $(-\infty, a_n] \downarrow (-\infty, a]$, so by the property of continuity of probabilities, $P((-\infty, a_n]) = F(a_n) \downarrow P((-\infty, a]) = F(a)$.

Property (c) is called *continuity from the right*, or simply right continuity. It is clear that a CDF need not be continuous from the left; indeed, in the example of tossing a fair coin twice, the CDF of the number of heads has jumps at the points $0, 1, 2$, and at those points the value of the CDF is *not* equal to the left limit of the function. Thus, at the jump points, the CDF is not left continuous. More precisely, one has the following result.

Proposition. Let $F(x)$ be the CDF of some random variable X. Then, for any x,

(a) $P(X = x) = F(x) - \lim_{y \uparrow x} F(y) = F(x) - F(x-)$, including those points x for which $P(X = x) = 0$, and
(b) $P(X \geq x) = P(X > x) + P(X = x) = (1 - F(x)) + (F(x) - F(x-)) = 1 - F(x-)$.

Example 4.6 (Dice Sum). Consider the experiment of rolling a fair die twice, and let X be the sum of the two rolls. Then X takes the values $x_1 = 2, x_2 = 3, \ldots, x_{11} = 12$. For example, $P(X = 7) = P(\{16, 25, 34, 43, 52, 61\}) = 6/36 = 1/6$. We can easily find the probabilities of all the possible values of X by direct counting, and we get the following pmf of X.

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

We see from the table that a median of the distribution is 7.

Example 4.7 (Indicator Variables). Consider again the experiment of rolling a fair die twice, and now define a random variable Y as follows:

$$Y = 1 \text{ if the sum of the two rolls } X \text{ is an even number,}$$
$$Y = 0 \text{ if the sum of the two rolls } X \text{ is an odd number.}$$

If we let A be the event that X is an even number, then $Y = 1$ if A happens and $Y = 0$ if A does not happen. Such random variables are called *indicator random variables* and are immensely useful in mathematical calculations in many complex situations.

Definition 4.5. Let A be any event in a sample space Ω. The *indicator random variable* for A is defined as

$$I_A = 1 \text{ if } A \text{ happens,}$$
$$I_A = 0 \text{ if } A \text{ does not happen.}$$

Thus, the distribution of an indicator variable is simply $P(I_A{=}1){=}P(A)$; $P(I_A{=}0) {=}1 - P(A)$.

An indicator variable is also called a *Bernoulli variable* with parameter p, where p is just $P(A)$.

Example 4.8 (Bridge). Consider the random variable

$$X = \text{ number of aces in North's hand in a bridge game.}$$

Clearly, X can take any of the values $x = 0, 1, 2, 3, 4$. If $X = x$, then the other $13 - x$ cards in North's hand must be non-ace cards. Thus, the pmf of X is

$$P(X = x) = \frac{\binom{4}{x}\binom{48}{13-x}}{\binom{52}{13}}, x = 0, 1, 2, 3, 4.$$

In decimals, the pmf of X is

x	0	1	2	3	4
$p(x)$	.304	.439	.213	.041	.003

Once again, the CDF of X is a jump function, taking jumps at the values $0, 1, 2, 3, 4$, namely the possible values of X. The CDF is

$$\begin{aligned} F(x) &= 0 \text{ if } x < 0, \\ &= .304 \text{ if } 0 \le x < 1, \\ &= .743 \text{ if } 1 \le x < 2, \\ &= .956 \text{ if } 2 \le x < 3, \\ &= .997 \text{ if } 3 \le x < 4, \\ &= 1 \text{ if } x \ge 4. \end{aligned}$$

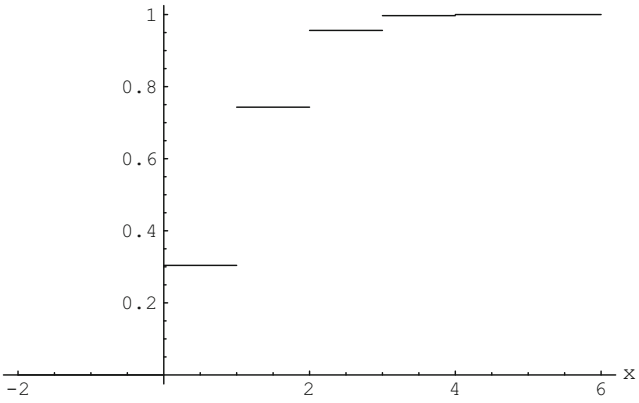


Fig. 4.2 CDF of the number of aces in the hand of one player in bridge

Once again, a plot of the CDF is given here for illustration (see Figure 4.2).

This is an example of the *hypergeometric distribution*, in which among N total objects some D are of type I and the other $N - D$ of type II, and n are selected without replacement from the N objects. Then, the pmf of the number of type I objects among the n selected is

$$P(X = x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}},$$

where $n + D - N \leq x \leq D$.

Example 4.9 (Long Head Runs and Pál Révész's Classroom Experiment). Suppose a fair coin is tossed $n = 5$ times. A *head run* is an uninterrupted sequence of heads; e.g., if the outcomes are $HTHHT$, then the first H is a run of length one, and the two successive heads HH in the third and fourth tosses form a head run of length two. For this particular sample point $HTHHT$, the length of the longest head run is two. For other sample points, it can be different. Define X = the length of the longest head run in $n = 5$ tosses of a fair coin. We want to find the distribution; i.e., the pmf of X.

The sample space of the experiment is

$$\begin{aligned}\Omega = \{&HHHHH, HHHHT, HHHTH, HHHTT, HHTHH, HHTHT,\\ &HHTTH, HHTTT, HTHHH, HTHHT, HTHTH, HTHTT,\\ &HTTHH, HTTHT, HTTTH, HTTTT, THHHH, THHHT,\\ &THHTH, THTHH, THTHT, THTTH, THTTT, TTHHH, TTHHT,\\ &TTHTH, THHTT, TTHTT, TTTHH, TTTHT, TTTTH, TTTTT\}.\end{aligned}$$

X takes the values $5, 4, 3, 3, 2, \ldots, 0$ corresponding to the 32 different sample points. Since each sample point has probability $\frac{1}{32}$, by direct counting, the pmf of X is as follows:

x	0	1	2	3	4	5
$p(x)$	$\frac{1}{32}$	$\frac{12}{32}$	$\frac{11}{32}$	$\frac{5}{32}$	$\frac{2}{32}$	$\frac{1}{32}$
$F(x)$	$\frac{1}{32}$	$\frac{13}{32}$	$\frac{24}{32}$	$\frac{29}{32}$	$\frac{31}{32}$	1

We notice that the probability of getting a head run of length three or more when the coin is tossed only five times is $P(X \geq 3) = 1 - P(X \leq 2) = 1 - F(2) = 1 - \frac{24}{32} = .25$. Most people feel surprised that such long head runs are fairly likely in such a small number of tosses.

The distribution of the length of the longest head run for a general number of tosses n is a hard and well-studied problem; one should *expect* that the longest head run would be of length $\frac{\log n}{\log 2}$. This is a classic result in discrete probability; see Erdös and Rényi (1970) and Erdös and Révész (1975). Thus, with $n =$ just 60 tosses, one should not be surprised to see six consecutive heads because $\frac{\log 60}{\log 2} \approx 6$. Pál Révész has conducted classroom experiments in which he asks one group of students to actually toss fair coins and report the length of the longest head run and have another group of students write imagined coin toss outcomes and report the length of the longest head run; with high accuracy, it is possible to identify from the reported values of the length of the longest head run whether a student actually tossed the coin or did a mental coin toss experiment!

Example 4.10 (How Common Is Bad Luck?). Suppose seven cookies are distributed independently and completely at random to seven children. Let X be the number of children who end up getting no cookies at all. We want to find the pmf of X.

This is a special case of the problem of *empty cells*. Precisely, if n balls (cookies) are distributed independently at random into m cells (children) and $\mu_0(m, n)$ denotes the number of empty cells (the number of children who receive no cookies at all), then it can be shown that, for $k \geq 1$,

$$P(\mu_0(m, n) \geq k) = \sum_{i=0}^{m-k} (-1)^i \binom{i + k - 1}{k - 1} \binom{m}{i + k} (m - i - k)^n / m^n;$$

from here, $P(\mu_0(m, n) = k)$ can be found by subtraction:

$$P(\mu_0(m, n) = k) = P(\mu_0(m, n) \geq k) - P(\mu_0(m, n) \geq k + 1).$$

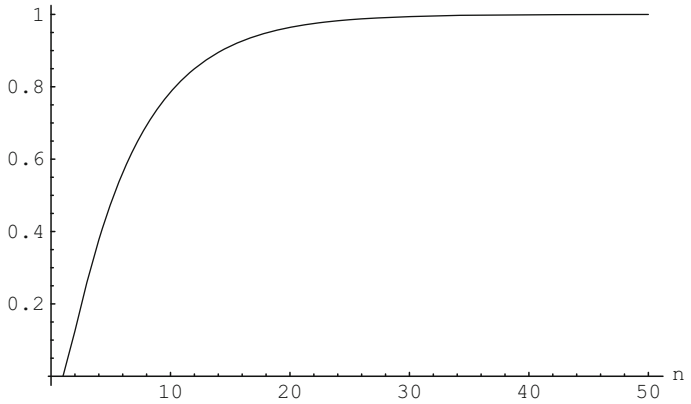


Fig. 4.3 Probability that at least one cell is empty if $2n$ balls are dropped in n cells

Using the formula above with $m = n = 7$ produces the following pmf for X, the number of children who do not receive any cookies at all:

x	0	1	2	3	4	5	6
$p(x)$	.006	.129	.428	.357	.077	.003	.000
$F(x)$	.006	.135	.563	.920	.997	1	1

What is the probability that as many as three children receive no cookies at all? This is $P(X \geq 3) = 1 - P(X \leq 2) = 1 - F(2) = 1 - .563 = .437$. There is almost a 44% probability that as many as three children receive no cookies at all.

From the general formula above, one can compute the probability that at least one cell will remain empty. When both the number of cells and the number of balls are large, this probability tends to be high unless the number of balls is a factor of magnitude bigger than the number of cells. A plot of this probability is given when the number of balls is twice the number of cells (see Figure 4.3). We can see that if 50 balls are randomly dropped into 25 cells, then with a very high probability some cells will remain empty.

4.2.1 Functions of a Random Variable

In applications, we are sometimes interested in the distribution of a function, say $g(X)$, of a basic random variable X; e.g., $g(X) = X^2$ or $g(X) = e^X$. A function of a random variable is also a random variable, so it has a distribution. If the function is not one-to-one, then a particular value y of $g(X)$ could be inherited from more than one value of X. One has to add the probabilities of all those values of X to find the probability that $g(X)$ takes the value y. Here is a formal statement for this argument.

Proposition (Function of a Random Variable). Let X be a discrete random variable and $Y = g(X)$ a real-valued function of X. Then, $P(Y=y) = \sum_{x:g(x)=y} p(x)$.

Example 4.11. Suppose X has the pmf $p(x) = \frac{c}{1+x^2}, x = 0, \pm 1, \pm 2, \pm 3$. Suppose we want to find the distribution of two functions of X:

$$Y = g(X) = X^3; Z = h(X) = \sin\left(\frac{\pi}{2}X\right).$$

First, the constant c must be explicitly evaluated. By directly summing the values,

$$\sum_x p(x) = \frac{13c}{5} \Rightarrow c = \frac{5}{13}.$$

Note that $g(X)$ is a one-to-one function of X, but $h(X)$ is not one-to-one. The values of Y are $0, \pm 1, \pm 8, \pm 27$. For example, $P(Y = 0) = P(X = 0) = c = 5/13; P(Y = 1) = P(X = 1) = c/2 = 5/26$, etc. In general, for $y = 0, \pm 1, \pm 8, \pm 27, P(Y = y) = P(X = y^{1/3}) = \frac{c}{1+y^{2/3}}$, with $c = 5/13$.

However, $Z = h(X)$ is not a one-to-one function of X. The possible values of Z are as follows:

x	$h(x)$
-3	1
-2	0
-1	-1
0	0
1	1
2	0
3	-1

So, for example, $P(Z = 0) = P(X = -2) + P(X = 0) + P(X = 2) = \frac{7}{5}c = 7/13$. The pmf of $Z = h(X)$ is

z	-1	0	1
$P(Z = z)$	3/13	7/13	3/13

Notice that Z has a *symmetric distribution*; i.e., Z and $-Z$ have the same pmf. This is not a coincidence. This is because X itself has a symmetric distribution and $Z = h(X)$ is an *odd function* of X. This is generally true.

Proposition. Suppose X has a distribution symmetric about zero; i.e., $P(X = x) = P(X = -x)$ for any x. Let $h(x)$ be an odd function; i.e., $h(-x) = -h(x)$ for any x. Then $Z = h(X)$ also has a distribution symmetric about zero.

4.2.2 Independence of Random Variables

Although we will not study probabilistic behavior of more than one random variable simultaneously in this chapter, it is useful to know the concept of *independent random variables* right now. The definition manifests the idea that no subset of a collection of random variables provides any probabilistic information about another nonoverlapping subset of that collection of variables.

Definition 4.6. Let $X_1, X_2, \ldots, X_k$ be $k \geq 2$ discrete random variables defined on the same sample space Ω. We say that $X_1, X_2, \ldots, X_k$ are *independent* if $P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_m = x_m)$, $\forall\, m \leq k$, and all $x_1, x_2, \ldots, x_m$.

It follows from the definition of independence of random variables that if X_1 and X_2 are independent, then any function of X_1 and any function of X_2 are also independent. In fact, we have a more general result.

Theorem 4.2. *Let $X_1, X_2, \ldots, X_k$ be $k \geq 2$ discrete random variables, and suppose they are independent. Let $U = f(X_1, X_2, \ldots, X_i)$ be some function of $X_1, X_2, \ldots, X_i$, and $V = g(X_{i+1}, \ldots, X_k)$ be some function of $X_{i+1}, \ldots, X_k$. Then, U and V are independent.*

This result is true of *any types of random variables* $X_1, X_2, \ldots, X_k$, not just discrete ones. We will omit a formal proof of this theorem, but it is clear why the theorem must be true. An event such as $\{f(X_1, X_2, \ldots, X_i) = c\}$ and another event such as $\{g(X_{i+1}, \ldots, X_k) = b\}$ are events involving nonoverlapping subsets of a set of independent random variables and therefore must be independent events.

A common notation widely used in probability and statistics is now introduced.

If $X_1, X_2, \ldots, X_k$ are independent, and moreover have the same CDF, say F, then we say that $X_1, X_2, \ldots, X_k$ are iid (or IID) and write $X_1, X_2, \ldots, X_k \overset{iid}{\sim} F$.

Example 4.12 (Two Simple Illustrations). Consider the experiment of tossing a fair coin (or any coin) four times. Suppose X_1 is the number of heads in the first two tosses and X_2 is the number of heads in the last two tosses. Then, it is intuitively clear that X_1 and X_2 are independent because the first two tosses have no information regarding the last two tosses. The independence can be easily verified mathematically by using the definition of independence.

Next, consider the experiment drawing 13 cards at random from a deck of 52 cards. Suppose X_1 is the number of aces and X_2 is the number of clubs among the 13 cards. Then, X_1, and X_2 are not independent. For example, $P(X_1 = 4, X_2 = 0) = 0$, but $P(X_1 = 4)$ and $P(X_2 = 0)$ are both >0, so $P(X_1 = 4)P(X_2 = 0) > 0$. So, X_1 and X_2 cannot be independent.

4.3 Expected Value of a Discrete Random Variable

By definition, a random variable takes different values on different occasions. It is natural to want to know what value it takes on average. Averaging is a very primitive concept. A simple average of just the possible values of the random variable will be misleading because some values may have so little probability that they are relatively inconsequential. The average or the mean value, also called the expected value of a random variable, is a weighted average of the different values of X, weighted according to how important the value is. Here is the definition.

Definition 4.7. Let X be a discrete random variable. We say that the *expected value* of X exists if $\sum_i |x_i|\, p(x_i) < \infty$, in which case the expected value is defined as

$$\mu = E(X) = \sum_i x_i\, p(x_i).$$

For notational convenience, we simply write $\sum_x xp(x)$ instead of $\sum_i x_i\, p(x_i)$. The expected value is also known as *the expectation or the mean* of X.

Remark. If the set of possible values of X is infinite, then the infinite sum $\sum_x xp(x)$ can take different values on rearranging the terms of the infinite series unless $\sum_x |x| p(x) < \infty$. So, as a matter of definition, we have to include the qualification that $\sum_x |x| p(x) < \infty$.

If the sample space Ω of the underlying experiment is finite or countably infinite, then we could also calculate the expectation by averaging directly over the sample space.

Proposition (Change of Variable Formula). Suppose the sample space Ω is finite or countably infinite and X is a discrete random variable with expectation μ. Then,

$$\mu = \sum_x xp(x) = \sum_\omega X(\omega) P(\omega),$$

where $P(\omega)$ is the probability of the sample point ω.

Proof. $\sum_\omega X(\omega)P(\omega) = \sum_x \sum_{\omega: X(\omega)=x} X(\omega)P(\omega) = \sum_x x \sum_{\omega: X(\omega)=x} P(\omega)$ $= \sum_x xp(x)$.

Important Point. Although it is not the focus in this chapter, in applications we are often interested in more than one variable at the same time. To be specific, consider two discrete random variables X and Y defined on a common sample space Ω. Then, we could construct new random variables out of X and Y; for example, XY, $X + Y$, $X^2 + Y^2$, etc. We can then talk of their expectations as well because after all these are random variables, too. Here is a quick example. Roll a fair die four times, and let X and Y be the number of ones and number of sixes obtained in the four

rolls, respectively. Then, we can ask what the expectation of $X + Y$, or $|X - Y|$, is which would correspond to the number of rolls in which either a one or a six was obtained the absolute difference between the number of ones and the number of sixes. The point is that we are often interested in studying more than one random variable with respect to a common underlying experiment ξ, and we may then want to compute the expectation of some function of these random variables. Here is a general definition of expectation of a function of more than one random variable.

Definition 4.8 (Function of Several Random Variables). Let $X_1, X_2, \ldots, X_n$ be n discrete random variables, all defined on a common sample space Ω, with a finite or a countably infinite number of sample points. We say that the expectation of a function $g(X_1, X_2, \ldots, X_n)$ exists if $\sum_\omega |g(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))| P(\omega) < \infty$, in which case the expected value of $g(X_1, X_2, \ldots, X_n)$ is defined as

$$E[g(X_1, X_2, \ldots, X_n)] = \sum_\omega g(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) P(\omega).$$

Example 4.13. Here is a simple example. Consider two rolls of a fair die. Let X be the number of ones and Y the number of sixes obtained, and then consider the function $g(X, Y) = XY$. Now, the sample space of this experiment is

$$\Omega = \{11, 12, 13, \ldots, 64, 65, 66\}.$$

Each sample point ω has the probability $P(\omega) = \frac{1}{36}$, and the respective values of XY for these 36 sample points are $X(\omega_1)Y(\omega_1) = 2 \times 0 = 0; X(\omega_2)Y(\omega_2) = 1 \times 0 = 0$, etc. By considering all 36 sample points, we obtain as the expected value of XY

$$E(XY) = 0 \times \frac{1}{36} + 0 \times \frac{1}{36} + \cdots + 0 \times \frac{1}{36} = \frac{2}{36} = \frac{1}{18}.$$

4.4 Basic Properties of Expectations

The next few results summarize the most fundamental properties of expectations.

Proposition.
(a) If there exists a finite constant c such that $P(X = c) = 1$, then $E(X) = c$.
(b) If X and Y are random variables defined on the same sample space Ω with finite expectations, and if $P(X \leq Y) = 1$, then $E(X) \leq E(Y)$.
(c) If X has a finite expectation, and if $P(X \geq c) = 1$, then $E(X) \geq c$. If $P(X \leq c) = 1$, then $E(X) \leq c$.

Proof. We assume all the random variables are discrete, although the results hold for arbitrary random variables. For part (a), $E(X) = \sum_\omega X(\omega) P(\omega) = \sum_\omega c P(\omega) = c \sum_\omega P(\omega) = c$.

Likewise, for part (b), by hypothesis, for each ω, $Y(\omega) \geq X(\omega) \Rightarrow Y(\omega)P(\omega) \geq X(\omega)P(\omega) \Rightarrow \sum_\omega Y(\omega)P(\omega) \geq \sum_\omega X(\omega)P(\omega) \Rightarrow E(Y) \geq E(X)$.

Part (c) follows from part (b).

Proposition (Linearity of Expectations). Let $X_1, X_2, \ldots, X_k$ be random variables defined on the same sample space Ω and $c_1, c_2, \ldots, c_k$ any real-valued constants. Then, provided $E(X_i)$ exists for every X_i,

$$E\left(\sum_{i=1}^{k} c_i X_i\right) = \sum_{i=1}^{k} c_i E(X_i);$$

in particular, $E(cX) = cE(X)$ and $E(X_1 + X_2) = E(X_1) + E(X_2)$ whenever the expectations exist.

Proof. We assume that the sample space Ω is finite or countably infinite. Then, by the change-of-variable formula,

$$\begin{aligned} E\left(\sum_{i=1}^{k} c_i X_i\right) &= \sum_{\omega} \left[\sum_{i=1}^{k} c_i X_i(\omega)\right] P(\omega) \\ &= \sum_{\omega} \left[\sum_{i=1}^{k} c_i X_i(\omega) P(\omega)\right] = \sum_{i=1}^{k} \sum_{\omega} [c_i X_i(\omega) P(\omega)] \\ &= \sum_{i=1}^{k} c_i \sum_{\omega} [X_i(\omega) P(\omega)] = \sum_{i=1}^{k} c_i E(X_i). \end{aligned}$$

The following fact also follows easily from the definition of the pmf of a function of a random variable. The result says that the expectation of a function of a random variable X can be calculated directly using the pmf of X itself without having to calculate the pmf of the function.

Proposition (Expectation of a Function). Let X be a discrete random variable on a sample space Ω with a finite or countable number of sample points and $Y = g(X)$ a function of X. Then,

$$E(Y) = \sum_{\omega} Y(\omega) P(w) = \sum_{x} g(x) p(x),$$

provided $E(Y)$ exists.

A very important property of independent random variables is the following factorization result on expectations.

Theorem 4.3. *Suppose $X_1, X_2, \ldots X_k$ are independent random variables. Then, provided each expectation exists,*

$$E(X_1 X_2 \cdots X_k) = E(X_1) E(X_2) \cdots E(X_k).$$

Proof. We prove this for $k = 2$; the general case then follows by induction. To do this,

$$E(X_1X_2) = \sum_{x_1,x_2} x_1x_2P(X_1 = x_1, X_2 = x_2) = \sum_{x_1,x_2} x_1x_2P(X_1 = x_1)P(X_2 = x_2)$$

$$= \sum_{x_1} x_1P(X_1 = x_1) \times \sum_{x_2} x_2P(X_2 = x_2) = E(X_1)E(X_2).$$

4.5 Illustrative Examples

Let us now see some more illustrative examples.

Example 4.14. Let X be the number of heads obtained in two tosses of a fair coin. We have worked out the pmf of X as $p(0) = p(2) = 1/4$, $p(1) = 1/2$. Therefore, $E(X) = 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1$. Since the coin is fair, we expect to see it show heads 50% of the number of times it is tossed, which is 50% of 2; i.e., 1.

Example 4.15 (Dice Sum)). Let X be the sum of the two rolls when a fair die is rolled twice. The pmf of X was worked out to be $p(2) = p(12) = 1/36$; $p(3) = p(11) = 2/36$; $p(4) = p(10) = 3/36$; $p(5) = p(9) = 4/36$; $p(6) = p(8) = 5/36$; $p(7) = 6/36$. Therefore, $E(X) = 2 \times 1/36 + 3 \times 2/36 + 4 \times 3/36 + \cdots + 12 \times 1/36 = 7$. This can also be seen by letting $X_1 =$ the face obtained on the first roll and $X_2 =$ the face obtained on the second roll, and by using $E(X) = E(X_1 + X_2) = E(X_1) + E(X_2) = 3.5 + 3.5 = 7$.

Let us now make this problem harder. Suppose that a fair die is rolled ten times and X is the sum of all ten rolls. The pmf of X is no longer so simple; it will be cumbersome to write it down. But, if we let $X_i =$ the face obtained on the ith roll, it is still true by the linearity of expectations that $E(X) = E(X_1+X_2+\cdots+X_{10}) = E(X_1) + E(X_2) + \cdots + E(X_{10}) = 3.5 \times 10 = 35$. We can easily compute the expectation, although the pmf would be difficult to write down.

Example 4.16 (Bridge). Let X be the number of aces in the hand of North in a bridge game. Then, from the pmf of X previously worked out, $E(X) = 0 \times .304 + 1 \times .439 + 2 \times .213 + 3 \times .041 + 4 \times .003 = 1$. This again makes common sense because there are four aces in the deck, and we should expect that one ace will go to each of the four players.

Example 4.17. Suppose seven cookies are distributed independently and completely at random to seven children. We previously worked out the pmf of X, the number of children who receive no cookies at all. Therefore, the expected number of children who receive no cookies is $E(X) = 0 \times .006 + 1 \times .129 + 2 \times .428 + 3 \times .357 + 4 \times .077 + 5 \times .003 + 6 \times .000 = 2.38$. Note that it is not possible to intuitively guess that 2.38 will turn out to be the expected value of X in this example. Also, note that 2.38 is not one of the possible values of X.

Example 4.18 (A Random Variable without a Finite Expectation). Let X take the positive integers $1, 2, 3, \ldots$ as its values with the pmf $p(x) = P(X = x) = \frac{1}{x(x+1)}, x = 1, 2, 3, \ldots$. This is a valid pmf because obviously $\frac{1}{x(x+1)} > 0$ for any $x = 1, 2, 3, \ldots$, and also the infinite series $\sum_{x=1}^{\infty} \frac{1}{x(x+1)}$ sums to 1, a fact from calculus. Now, $E(X) = \sum_{x=1}^{\infty} xp(x) = \sum_{x=1}^{\infty} x\frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \frac{1}{x+1} = \sum_{x=2}^{\infty} \frac{1}{x} = \infty$, also a fact from calculus.

This example shows that not all random variables have finite expectations. Here, the reason for the infiniteness of $E(X)$ is that X takes large integer values x with probabilities $p(x)$ that are not adequately small. The large values are realized sufficiently often that on average X becomes larger than any given finite number.

Example 4.19 (An Interesting Coin-Tossing Expectation). Suppose a fair coin is tossed n times. Let X be the number of heads obtained; then, $n - X$ is the number of tails obtained, and therefore $W = g(X) = \max\{X, n - X\}$ is the larger of the number of heads and the number of tails. We do not have enough tools right now to find the expected value of W for a general n. However, we can compute it directly for small n. As an example, suppose $n = 4$. Then there are $2^4 = 16$ sample points in the experiment, and it is easy to verify by simple counting that the pmf of X is

x	0	1	2	3	4
$p(x)$	1/16	4/16	6/16	4/16	1/16

Therefore, by the formula for the expectation of a function of a random variable X,

$$E(W) = E[\max\{X, 4 - X\}] = \sum_{x=0}^{4} g(x)p(x) = 4 \times 1/16 + 3 \times 4/16$$
$$+2 \times 6/16 + 3 \times 4/16 + 4 \times 1/16 = 44/16 = 2.75.$$

Thus, even if X and Y have the same expectation, say μ, the larger of X and Y does not have that same expectation μ.

4.6 Using Indicator Variables to Calculate Expectations

The zero-one nature of indicator random variables is extremely useful for calculating expectations of certain integer-valued random variables whose distributions are sometimes so complicated that it would be difficult to find their expectations directly from the definition. We describe the technique and some illustrations of it below.

Proposition. Let X be an integer-valued random variable such that it can be represented as $X = \sum_{i=1}^{m} c_i I_{A_i}$ for some m, constants $c_1, c_2, \ldots, c_m$, and suitable events $A_1, A_2, \ldots, A_m$. Then, $E(X) = \sum_{i=1}^{m} c_i P(A_i)$.

Proof. $E(X) = E[\sum_{i=1}^{m} c_i I_{A_i}] = \sum_{i=1}^{m} c_i E[I_{A_i}] = \sum_{i=1}^{m} c_i P(A_i)$.

Here are some illustrative examples.

Example 4.20 (Coin Tosses). Suppose a coin that has probability p of showing heads in any single toss is tossed n times, and let X denote the number of times in the n tosses that a head is obtained. Then, $X = \sum_{i=1}^{n} I_{A_i}$, where A_i is the event that a head is obtained in the ith toss. Therefore, $E(X) = \sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{n} p = np$.

A direct calculation of the expectation would involve finding the pmf of X and obtaining the sum $\sum_{x=0}^{n} xP(X = x)$; it can also be done that way, but that is a much longer calculation.

The random variable X of this example is a *binomial random variable* with parameters n and p. Its pmf is given by the formula $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, 1, 2, \ldots, n$. We will study binomial random variables in greater detail later.

Example 4.21 (The Matching Problem). Recall the matching problem in which n entities, say $1, 2, \ldots, n$, are linearly arranged at random in the locations marked as $1, 2, \ldots, n$. Suppose that after rearrangement the number at location i is $\pi(i)$. We want to study the number of matches defined as $X =$ number of locations i such that $\pi(i) = i$.

We use the indicator variable method to find the expected number of matches. To do this, define $A_i =$ there is a match at location i. Then $X = \sum_{i=1}^{n} I_{A_i}$. Now, for any i, $P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$, and therefore we have the quite elegant result that whatever n is, $E(X) = \sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{n} \frac{1}{n} = n \times \frac{1}{n} = 1$. Again, direct verification of this would require us to calculate the pmf of the number of matches for any given n. If we do, then we will find from algebra that $\sum_{k=1}^{n} kP(X = k) = 1$ for any n.

Example 4.22 (Missing Faces in Die Rolls). Suppose a fair die is rolled n times, and let X be the number of faces that never show up in these n rolls. Define $A_i =$ the ith face is missing. Then, $X = \sum_{i=1}^{6} I_{A_i}$. For any i, $P(A_i) = (\frac{5}{6})^n$. Therefore, $E(X) = \sum_{i=1}^{6} P(A_i) = 6 \times (\frac{5}{6})^n$. For example, if a fair die is rolled ten times, then the expected number of missing faces is $6 \times (\frac{5}{6})^{10} = .97 \approx 1$. Again, direct calculation of this expectation would be a much more complicated exercise.

Example 4.23 (Consecutive Heads in Coin Tosses). Suppose a coin with probability p for heads in a single toss is tossed n times. How many times can we expect to see a head followed by at least one more head? For example, if $n = 5$ and we see the outcomes *HTHHH*, then we see a head followed by at least one more head twice.

Define $A_i =$ the ith and the $(i + 1)$th toss both result in heads. Then $X =$ number of times a head is followed by at least one more head $= \sum_{i=1}^{n-1} I_{A_i}$, so $E(X) = \sum_{i=1}^{n-1} P(A_i) = \sum_{i=1}^{n-1} p^2 = (n - 1)p^2$. For example, if a fair coin is tossed 20 times, we can expect to see a head followed by another head about five times $(19 \times .5^2 = 4.75)$.

4.7 The Tail Sum Method for Calculating Expectations

Another useful technique for calculating expectations of nonnegative integer-valued random variables is based on the CDF of the random variable rather than directly on the pmf. This method is useful when calculating probabilities of the form $P(X > x)$ is logically more straightforward than calculating $P(X = x)$ directly. Here is the expectation formula based on the tail CDF.

Theorem 4.4. *Let X take values $0, 1, 2, \ldots$. Then,*

$$E(X) = \sum_{n=0}^{\infty} P(X > n).$$

Proof. An informal proof starting from the right side of the formula is to note that

$$\begin{aligned}\sum_{n=0}^{\infty} P(X > n) &= [p(1) + p(2) + p(3) + \cdots] + [p(2) + p(3) + \cdots] \\ &\quad + [p(3) + p(4) + \cdots] + \cdots \\ &= p(1) + 2p(2) + 3p(3) + \cdots = E(X).\end{aligned}$$

Here are some examples of applications of the method.

Example 4.24 (Waiting Time to See the First Head). Suppose a coin with probability p for heads is tossed until a head is obtained for the first time. How many tosses will it take on average to see the first head? Let X denote the number of heads necessary to obtain the very first head. Then $X > n$ simply means that the first n tosses have all produced tails. Therefore,

$$E(X) = \sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} (1-p)^n = \frac{1}{1-(1-p)} = \frac{1}{p}.$$

X is called a *geometric random variable with parameter p*. If the coin is fair, this says that on average the first head will be seen at the second toss.

Example 4.25 (Family Planning). Suppose a couple will have children until they have at least one child of each sex. How many children can they expect to have? Let X denote the childbirth at which they have a child of each sex for the first time. Suppose the probability that any particular childbirth will be a boy is p and that all births are independent. Then,

$$P(X > n) = P(\text{The first } n \text{ children are all boys or all girls}) = p^n + (1-p)^n.$$

Therefore, $E(X) = 2 + \sum_{n=2}^{\infty} [p^n + (1-p)^n] = 2 + p^2/(1-p) + (1-p)^2/p = \frac{1}{p(1-p)} - 1$. If boys and girls are equally likely on any childbirth, then this says that a couple waiting to have a child of each sex can expect to have three children.

Example 4.26 (Elevator Stops in a High-Rise Building). Suppose a building has f floors above the basement, where it starts with N passengers. Assume that the passengers get off, independently of each other, at one of the f floors with an equal probability. Let X be the first floor after the basement at which the elevator stops to let a passenger off.

Then $X > n$ simply means that no passengers got off at floors $1, 2, \ldots, n$, so $P(X > n) = \frac{(f-n)^N}{f^N} = (1 - \frac{n}{f})^N$ for $n \leq f - 1$. Thus,

$$E(X) = 1 + \sum_{n=1}^{f-1} \left(1 - \frac{n}{f}\right)^N .$$

For example, if $f = 10$ and $N = 8$, then by computing with the formula above, $E(X) = 1.68$. If $f = 50$ and $N = 15$, then $E(X) = 3.65$. It would be interesting to take some real data on this in a high-rise hotel.

4.8 Variance, Moments, and Basic Inequalities

The expected value is calculated with the intention of understanding what a typical value of a random variable is. But two very different distributions can have exactly the same expected value. A common example is that of a return on an investment in a stock. Two stocks may have the same average return, but one may be much riskier than the other in the sense that the variability in the return is much higher for that stock. In that case, most risk-averse individuals would prefer to invest in the stock with less variability. Here is another example. Suppose a fair coin is to be tossed, and if it shows heads, I pay you one dollar, while if it shows tails, you pay me one dollar. This is a fair game. But suppose that, instead of one dollar, we raise the stakes: heads up, I pay you a hundred dollars, and tails up, you pay me a hundred dollars. This is still a fair game. But surely we will consider the second game to be riskier. Measures of risk or variability are of course not unique. Some natural measures that come to mind are $E(|X - \mu|)$, known as the *mean absolute deviation*, or $P(|X - \mu| > k)$, for some suitable k. However, neither of these two is the most common measure of variability. The most common measure is the *standard deviation* of a random variable. We will discuss later why it has become the most common measure of variability. Here is the definition.

Definition 4.9. Let a random variable X have a finite mean μ. The *variance* of X is defined as

$$\sigma^2 = E[(X - \mu)^2],$$

and the *standard deviation* of X is defined as $\sigma = \sqrt{\sigma^2}$.

Remark. Thus, variance measures average squared deviation from a very special value, namely the expected value. If X has a unit of measurement, then the standard deviation σ has the same unit. It is easy to prove that $\sigma^2 < \infty$ if and only

if $E(X^2)$, the *second moment* of X, is finite. It is not uncommon to mistake the standard deviation for the mean absolute deviation, but they are not the same.

Caution. The standard deviation of a random variable X and the mean absolute deviation are *not* the same. In fact, an inequality holds.

Proposition. $\sigma \geq E(|X - \mu|)$, and σ is strictly greater unless X is a constant random variable, namely $P(X = \mu) = 1$.

We list some basic properties of the variance of a random variable.

Proposition.
(a) $\text{Var}(cX) = c^2\text{Var}(X)$ for any real c.
(b) $\text{Var}(X + k) = \text{Var}(X)$ for any real k.
(c) $\text{Var}(X) \geq 0$ for any random variable X and equals zero only if $P(X = c) = 1$ for some real constant c.
(d) $\text{Var}(X) = E(X^2) - \mu^2$.

Proof. We prove part (a) and part (d). For part (a),

$$\text{Var}(cX) = E[(cX - E(cX))^2] = E[(cX - c\mu)^2] = E[c^2(X - \mu)^2] = c^2\text{Var}(X).$$

For part (d),

$$\begin{aligned}\text{Var}(X) &= E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] = E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2.\end{aligned}$$

The quantity $E(X^2)$ is called the *second moment* of X. The definition of a general moment is as follows.

Definition 4.10. Let X be a random variable and $k \geq 1$ a positive integer. Then $E(X^k)$ is called the *kth moment of X* and $E(X^{-k})$ is called the *kth inverse moment of X*, provided they exist.

We therefore have the following relationships involving moments and the variance:

variance = second moment − (first moment)^2;
second moment = variance + (first moment)^2.

In general, third or higher moments cannot be calculated if only the variance and the first moment are known.

Statisticians often use the third moment around the mean as a measure of lack of symmetry in the distribution of a random variable. The point is that if a random variable X has a symmetric distribution and has a finite mean μ, then all odd moments around the mean, namely $E[(X - \mu)^{2k+1}]$, will be zero if the moment exists. In particular, $E[(X - \mu)^3]$ will be zero. Likewise, statisticians also use the fourth moment around the mean as a measure of how spiky the distribution is around the mean. To

make these indices independent of the choice of unit of measurement (e.g., inches or centimeters), they use certain scaled measures of asymmetry and peakedness. Here are the definitions.

Definition 4.11.

(a) Let X be a random variable with $E[|X|^3] < \infty$. The *skewness* of X is defined as

$$\beta = \frac{E[(X-\mu)^3]}{\sigma^3}.$$

(b) Suppose X is a random variable with $E[X^4] < \infty$. The *kurtosis* of X is defined as

$$\gamma = \frac{E[(X-\mu)^4]}{\sigma^4} - 3.$$

The skewness β is zero for symmetric distributions, but the converse need not be true. The kurtosis γ is necessarily ≥ -2 but can be arbitrarily large, with spikier distributions generally having a larger kurtosis. But a very good interpretation of γ is not really available. We will later see that $\gamma = 0$ for all *normal distributions*, hence the motivation for subtracting 3 in the definition of γ.

4.9 Illustrative Examples

We now go back to the variance and calculate it in some examples. A general discussion about interpretation of the variance and the standard deviation will be made after these examples.

Example 4.27 (Variance of Number of Heads). Consider the experiment of two tosses of a fair coin, and let X be the number of heads obtained. Then, we have seen that $p(0) = p(2) = 1/4$ and $p(1) = 1/2$. Thus, $E(X^2) = 0 \times 1/4 + 1 \times 1/2 + 4 \times 1/4 = 3/2$ and $E(X) = 1$. Therefore, $\mathrm{Var}(X) = E(X^2) - \mu^2 = 3/2 - 1 = \frac{1}{2}$, and the standard deviation is $\sigma = \sqrt{.5} = .707$.

Example 4.28 (Variance of Dice Sum). Let X be the sum of two independent rolls of a fair die. Then, from the pmf of X previously derived, $E(X) = 7$ and $E(X^2) = 2^2 \times 1/36 + 3^2 \times 2/36 + 4^2 \times 3/36 + \cdots + 12^2 \times 1/36 = 329/6$, and therefore $\mathrm{Var}(X) = E(X^2) - \mu^2 = 329/6 - 49 = \frac{35}{6} = 5.83$ and the standard deviation is $\sigma = \sqrt{5.83} = 2.415$.

Example 4.29 (Variance in the Matching Problem). Again let X be the number of locations at which a match occurs when n numbers, say, $1, 2, \ldots, n$, are rearranged in a random order. We have previously seen by using the indicator variable method that $E(X) = 1$, whatever n is. We now use the indicator variable method to also calculate the variance.

To do this, define again $A_i =$ there is a match at location i. Then, $X = \sum_{i=1}^{n} I_{A_i}$. We first find the second moment of X.

Now, $X^2 = (\sum_{i=1}^n I_{A_i})^2 = (\sum_{i=1}^n [I_{A_i}]^2 + 2\sum_{1\le i<j\le n} I_{A_i} I_{A_j}) = (\sum_{i=1}^n I_{A_i} + 2\sum_{1\le i<j\le n} I_{A_i} I_{A_j})$.

Therefore,

$$E(X^2) = \sum_{i=1}^n P(A_i) + 2 \sum_{1\le i<j\le n} P(A_i \cap A_j).$$

For any i, $P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$, and for all $i, j, i < j$, $P(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$.

Therefore,

$$E(X^2) = n \times \frac{1}{n} + 2\binom{n}{2}\frac{1}{n(n-1)} = 1 + 1 = 2.$$

Therefore, $\text{Var}(X) = E(X^2) - [E(X)]^2 = 2 - 1 = 1$, regardless of the value of n.

To summarize, in the matching problem, regardless of the value of n, the mean and the variance of the number of matches are both 1. We will later see that this property of equality of the mean and variance is true for a well-known distribution called the *Poisson distribution*, and in fact *a Poisson distribution does provide an extremely accurate approximation* for the exact pmf of the number of matches.

Example 4.30 (A Random Variable with an Infinite Variance). If a random variable has a finite variance, then it can be shown that it must have a finite mean. This example shows that the converse need not be true.

Let X be a discrete random variable with the pmf $P(X = x) = \frac{c}{x(x+1)(x+2)}$, $x = 1, 2, 3, \ldots$, where the normalizing constant $c = 4$. The expected value of X is

$$E(X) = \sum_{x=1}^{\infty} x \times \frac{4}{x(x+1)(x+2)} = 4\sum_{x=1}^{\infty} \frac{1}{(x+1)(x+2)} = 4 \times 1/2 = 2.$$

Therefore, by direct verification, X has a finite expectation. Let us now examine the second moment of X.

$$E(X^2) = \sum_{x=1}^{\infty} x^2 \times \frac{4}{x(x+1)(x+2)} = 4\sum_{x=1}^{\infty} x \times \frac{1}{(x+1)(x+2)} = \infty$$

because the series $\sum_{x=1}^{\infty} x \times \frac{1}{(x+1)(x+2)}$ is not finitely summable, a fact from calculus. Since $E(X^2)$ is infinite but $E(X)$ is finite, $\sigma^2 = E(X^2) - [E(X)]^2$ must also be infinite.

Example 4.31 (Variance Can Mislead). It is possible for a random variable to be essentially very concentrated around some number and only rarely take a large value, thereby causing a large variance. The large variance gives the impression that there is a lot of uncertainty about which value X will take, but actually there is almost no uncertainty at all.

Let n be a fixed (large) positive integer, and suppose X has the pmf $P(X=0) = 1-\frac{1}{n}$, $P(X=n)=\frac{1}{n}$. Then, $E(X)=1$ and $E(X^2)=n$. Therefore, $\mathrm{Var}(X) = n-1$, which is large if n is large, although $P(X=0)\approx 1$.

4.9.1 Variance of a Sum of Independent Random Variables

If a collection of random variables are independent, then just like the expectation, the variance also adds up. Precisely, one has the following very useful fact.

Theorem 4.5. *Let $X_1, X_2, \ldots, X_n$ be n independent random variables. Then,*

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n).$$

Proof. It is enough to prove this result for $n = 2$ because we can then prove the general case by induction. By definition, writing $E(X_i) = \mu_i$,

$$\begin{aligned}\mathrm{Var}(X_1 + X_2) &= E[(X_1 + X_2) - E(X_1 + X_2)]^2 = E[(X_1 + X_2) - (\mu_1 + \mu_2)]^2 \\ &= E[(X_1 - \mu_1) + (X_2 - \mu_2)]^2 = E[(X_1 - \mu_1)^2] \\ &\quad + E[(X_2 - \mu_2)^2] + 2E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 0\end{aligned}$$

because $E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[(X_1 - \mu_1)]E[(X_2 - \mu_2)] = 0 \times 0 = 0$ by virtue of the independence of X_1 and X_2. This proves the result.

An important corollary of this result is the following formula for the mean, $\bar{X}$, of n independent and identically distributed (iid) random variables.

Corollary 4.1. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with a common variance $\sigma^2 < \infty$. Let $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$. Then* $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

Proof.

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

4.10 Utility of μ and σ as Summaries

The mean and the variance together have earned the status of being the two most common summaries of a distribution. Are there any justifications for it? One reason for using the standard deviation σ rather than, say, the mean absolute deviation $E(|X-\mu|)$ is that calculating and operating with σ is typically much easier. However, this is a matter of convenience. A better question is whether μ and σ are useful

summaries of the distribution of a random variable. The answer is a qualified yes. The inequalities in this section suggest that knowing just the values of μ and σ, it is in fact possible to say something useful about the full distribution.

4.10.1 Chebyshev's Inequality and the Weak Law of Large Numbers

Theorem 4.6.

(a) **(Chebyshev's Inequality).** *Suppose* $E(X)=\mu$ *and* $\mathrm{Var}(X)=\sigma^2$ *are assumed to be finite. Let* k *be any positive number. Then*

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

(b) **(Markov's Inequality).** *Suppose* X *takes only nonnegative values and that* $E(X)=\mu$ *is assumed to be finite. Let* c *be any positive number. Then,*

$$P(X \geq c) \leq \frac{\mu}{c}.$$

Proof. We prove part (a) assuming part (b), as part (b) is simpler and can be proved easily. Define $Y = (X-\mu)^2$. Then $E(Y) = E[(X-\mu)^2] = \sigma^2$. Therefore, by part (b),

$$P(|X-\mu| \geq k\sigma) = P((X-\mu)^2 \geq k^2\sigma^2) = P(Y \geq k^2\sigma^2) \leq \frac{E(Y)}{k^2\sigma^2}$$
$$= \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Remark. The virtue of these two inequalities is that they make no restrictive assumptions on the random variable X. Whenever μ and σ are finite, Chebyshev's inequality is applicable, and whenever μ is finite, Markov's inequality applies, provided the random variable is nonnegative. They are universal inequalities. Furthermore, the inequalities *cannot* be improved without additional restrictions; they are in fact attained by suitable distributions. However, the universal nature also makes them typically quite conservative. As an illustration, let us see the following example.

Example 4.32 (Conservativeness of Chebyshev's Inequality). Suppose X is the sum of two rolls of a fair die. We have computed μ and σ previously as $\mu = 7$, $\sigma = 2.415$. Choosing $k = 2$ in Chebyshev's inequality, the inequality tells us

$$P(|X-7| \geq 4.830) = P(X \geq 11.83) + P(X \leq 2.17) = P(X=2) + P(X=12)$$
$$= \frac{2}{36} = .056.$$

But all we can say from Chebyshev's inequality is that

$$P(|X - 7| \geq 4.830) = P(|X - \mu| \geq 2\sigma) \leq \frac{1}{4} = .25.$$

Clearly the bound obtained from Chebyshev's inequality is weak.

Although Chebyshev's inequality usually gives conservative estimates for tail probabilities, it does imply a major result in probability theory in a special case. Consider a random variable X with a finite mean μ and a finite variance σ^2. Suppose $X_1, X_2, \ldots, X_n$ are independent samples on this variable X. If the number of sample values n is large, then the mean of the sample values ought to be close to the mean in the entire population; i.e., $\bar{X}$ should be close to μ. The closeness of $\bar{X}$ to μ for large n can be given different meanings. The simplest interpretation is that, with a high probability, the mean of a large sample should be numerically close to the mean in the entire population. Chebyshev's inequality, although elementary, can give us this result when X has a finite variance. However, the finite-variance condition is not needed, although if the variance is not finite, then a much harder proof is necessary. Theorem 4.7 gives the property we are referring to.

Theorem 4.7 (Weak Law of Large Numbers). *Let $X_1, X_2, \ldots$ be iid random variables, with $E(X_i) = \mu, \text{Var}(X_i) = \sigma^2 < \infty$. Then, for any $\epsilon > 0$, $P(|\bar{X} - \mu| > \epsilon) \to 0$ as $n \to \infty$.*

Proof. By Chebyshev's inequality and our previously observed fact that $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$,

$$P(|\bar{X} - \mu| > \epsilon) = P\left(|\bar{X} - \mu| > \frac{\sqrt{n}\epsilon}{\sigma}\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{\left(\frac{\sqrt{n}\epsilon}{\sigma}\right)^2} = \frac{\sigma^2}{n\epsilon^2}$$

$$\to 0 \text{ as } n \to \infty.$$

The interpretation of this result is that if we take a large sample from a population, then most of the time our sample mean will be numerically close to the population mean. Occasionally, just by bad luck, even though we have taken a large sample, the sample mean will not be close enough to the population mean. But such bad luck will occur only occasionally; i.e., with a small probability.

There is a stronger version of the weak law of large numbers, which says that in fact, with certainty, $\bar{X}$ will converge to μ as $n \to \infty$. The precise mathematical statement is that

$$P\left(\lim_{n\to\infty} \bar{X} = \mu\right) = 1.$$

The only condition needed is that $E(|X_i|)$ should be finite. This is called the *strong law of large numbers*. It is impossible to prove it without using much more

sophisticated concepts and techniques than we are using here. The strong law, practically speaking, gives us an assurance that computational techniques such as *Monte Carlo* are certain to give us accurate results in problems such as numerical approximation of (complicated) definite integrals. We will not make further use of the strong law in this text.

*4.10.2 * Better Inequalities*

Now we return to the issue of the utility of μ and σ as summaries of a distribution and to describing inequalities that are somewhat stronger than Chebyshev's inequality.

Inequalities better than Chebyshev's or Markov's inequalities are available under additional restrictions on the distribution of the underlying random variable X. For now, we state three other inequalities that can sometimes give bounds better than what Chebyshev's or Markov's can give. It is also important to note that the first two inequalities below can handle one-sided deviations from the mean μ, although the only thing one can assert in general about one-sided deviations from using Chebyshev's inequality is still the $\frac{1}{k^2}$ bound.

Theorem 4.8.

(a) **(Cantelli's Inequality).** *Suppose* $E(X) = \mu$ *and* $\text{Var}(X) = \sigma^2$ *are assumed to be finite. Then,*

$$P(X - \mu \geq k\sigma) \leq \frac{1}{k^2 + 1},$$

$$P(X - \mu \leq -k\sigma) \leq \frac{1}{k^2 + 1}.$$

(b) **(Paley-Zygmund Inequality).** *Suppose* X *takes only nonnegative values, with* $E(X) = \mu$ *and* $\text{Var}(X) = \sigma^2$*, assumed to be finite. Then, for* $0 < c < 1$*,*

$$P(X > c\mu) \geq (1 - c)^2 \frac{\mu^2}{\mu^2 + \sigma^2}.$$

(c) **(Alon-Spencer Inequality).** *Suppose* X *takes only nonnegative integer values, with* $E(X) = \mu$ *and* $\text{Var}(X) = \sigma^2$ *assumed to be finite. Then,*

$$P(X = 0) \leq \frac{\sigma^2}{\mu^2 + \sigma^2}.$$

These inequalities may be seen in Rao (1973), Paley and Zygmund (1932), and Alon and Spencer (2000), respectively.

Example 4.33 (Matching Problem Revisited). Consider again the matching problem with X being the number of locations among n locations where a match occurs. Recall that, for any n, $E(X) = \text{Var}(X) = 1$. We can therefore assert from the Alon-Spencer inequality that, for any n, $P(X = 0) \le \frac{1}{1+1} = \frac{1}{2}$; i.e., there is always at least a 50% chance of some matches. Note that this particular bound cannot be improved because, for $n = 2$, $P(X = 0)$ is exactly $\frac{1}{2}$.

Next, if we apply the first of the two Cantelli inequalities with $k = 4$, then we get that, for any n, $P(X \ge 5) = P(X - 1 \ge 4) \le \frac{1}{4^2+1} < .06$. Thus, the chances of as many as five matches are always less than 6%, regardless of the value of n.

4.11 ∗ Other Fundamental Moment Inequalities

We saw several important inequalities, primarily based on the mean and variance, in the previous section. The area of probability inequalities is an extremely rich and diverse area. The reason is that inequalities are tremendously useful in giving approximate answers when the exact answer to a problem or a calculation is very hard or perhaps even impossible to obtain. We will periodically present and illustrate inequalities over the rest of the book. Some really basic inequalities based on moments are presented in this section.

Theorem 4.9.

(a) **(Cauchy-Schwarz Inequality).** *Let X and Y be two random variables such that $E(X^2)$ and $E(Y^2)$ are finite. Then,*

$$E(|XY|) \le \sqrt{E(X^2)}\sqrt{E(Y^2)}.$$

(b) **(Holder's Inequality).** *Let X and Y be two random variables and $1 < p < \infty$ be a real number such that $E(|X|^p) < \infty$. Let $q = \frac{p}{p-1}$, and suppose $E(|Y|^q) < \infty$. Then,*

$$E(|XY|) \le [E(|X|^p)]^{\frac{1}{p}}[E(|Y|^q)]^{\frac{1}{q}}.$$

(c) **(Minkowski's Inequality).** *Let X and Y be two random variables and $p \ge 1$ a real number such that $E(|X|^p), E(|Y|^p) < \infty$. Then,*

$$[E(|X + Y|^p)]^{\frac{1}{p}} \le [E(|X|^p)]^{\frac{1}{p}} + [E(|Y|^p)]^{\frac{1}{p}},$$

and, in particular, if $E(|X|)$ and $E(|Y|)$ are both finite, then

$$E(|X + Y|) \le E(|X|) + E(|Y|),$$

known as the triangular inequality.

(d) **(Lyapounov Inequality).** *Let X be a random variable and $0 < \alpha < \beta$ such that $E(|X|^{\beta}) < \infty$. Then,*

$$[E(|X|^{\alpha})]^{\frac{1}{\alpha}} \leq [E(|X|^{\beta})]^{\frac{1}{\beta}}.$$

Proof. Part (a) is a special case of part (b) on using $p = 2$ (and therefore $q = 2$). We prove part (b). We do not prove part (c) and part (d) here.

The proof simplifies notationally on assuming that $E(|X|^p) = E(|Y|^q) = 1$, so that Holder's inequality will amount to proving $E(|XY|) \leq \frac{1}{p} + \frac{1}{q} = 1$. This assumption that $E(|X|^p) = E(|Y|^q) = 1$ can be made without any loss of generality, for if it is not true to begin with, we can define new random variables $X^* = \frac{X}{[E(|X|^p)]^{1/p}}, Y^* = \frac{Y}{[E(|Y|^q)]^{1/q}}$, and the assumption will hold for X^*, Y^*. Furthermore, Holder's inequality holds for X, Y if and only if it holds for X^*, Y^*. So, we go ahead and assume that $E(|X|^p) = E(|Y|^q) = 1$.

The key fact we need is the following inequality for the exponential function:

$$e^{tx+(1-t)y} \leq te^x + (1-t)e^y \ \forall x, y, \text{ and } 0 \leq t \leq 1.$$

This is a consequence of the convexity of the function e^x on the real line. Consider now two positive numbers a and b, and $p > 1, q = \frac{p}{p-1}$. Note that the definition of q makes $\frac{1}{p} + \frac{1}{q} = 1$. Denote

$$t = \frac{1}{p}, \log(a^p) = x, \log(b^q) = y.$$

Then, by the exponential function inequality above,

$$\begin{aligned} e^{\frac{\log a^p}{p} + \frac{\log b^q}{q}} &\leq \frac{1}{p} e^{\log a^p} + \frac{1}{q} e^{\log b^q} \\ \Rightarrow ab &\leq \frac{a^p}{p} + \frac{b^q}{q}. \end{aligned}$$

Apply this last inequality to $a = |X(\omega)|$ and $b = |Y(\omega)|$ for fixed ω. This will result in

$$|X(\omega)Y(\omega)| \leq \frac{1}{p}|X(\omega)|^p + \frac{1}{q}|Y(\omega)|^q.$$

Since this holds for any fixed ω, it will also hold on averaging over ω; i.e.,

$$\begin{aligned} E(|XY|) &\leq \frac{1}{p} E(|X|^p) + \frac{1}{q} E(|Y|^q) \\ &= \frac{1}{p} + \frac{1}{q} = 1, \end{aligned}$$

which is what we needed to prove.

4.11.1 * Applying Moment Inequalities

Example 4.34 (Application of Lyapounov's Inequality). Here is a simple application of Lyapounov's inequality. Suppose X is a positive random variable and we only know that it has mean 5 and variance 4. What can we say about the third moment of X? By Lyapounov's inequality, we can say that $[E(X^3)]^{(1/3)} \geq E(X) = 5$, and therefore $E(X^3) \geq 5^3 = 125$. Also, since X has mean 5 and variance 4, we have $E(X^2) = \text{Var}(X) + [E(X)]^2 = 4 + 25 = 29$. Again, by Lyapounov's inequality, $[E(X^3)]^{(1/3)} \geq [E(X^2)]^{(1/2)} = \sqrt{29}$. This gives $E(X^3) \geq (29)^{3/2} = 156.17$. This is a better bound than $E(X^3) \geq 125$. Therefore, by using the better bound, we can assert that $E(X^3) \geq 156.17$.

Example 4.35 (Application of Cauchy-Schwarz Inequality). The most useful applications of Holder's inequality and the Cauchy-Schwarz inequality are to continuous random variables, which we have not discussed yet. We give a simple application of the Cauchy-Schwarz inequality to a dice problem.

Suppose X and Y are the maximum and the minimum of two rolls of a fair die. Also let X_1 be the first roll and X_2 be the second roll. Note that $XY = X_1X_2$. Therefore,

$$\begin{aligned} E(\sqrt{X}\sqrt{Y}) &= E(\sqrt{XY}) = E(\sqrt{X_1X_2}) \\ &= E(\sqrt{X_1}\sqrt{X_2}) = E(\sqrt{X_1})E(\sqrt{X_2}) \\ &= [E(\sqrt{X_1})]^2 = \frac{1}{36}(\sqrt{1} + \cdots + \sqrt{6})^2 \\ &= \frac{1}{36} \times (10.83)^2 = 3.26. \end{aligned}$$

Therefore, by the Cauchy-Schwarz inequality,

$$\sqrt{E(X)}\sqrt{E(Y)} \geq 3.26$$

$$\Rightarrow \sqrt{E(X)}\sqrt{7 - E(X)} \geq 3.26$$

(since $E(X) + E(Y) = E(X_1) + E(X_2) = 7$)

$$\Rightarrow \sqrt{m(7-m)} \geq 3.26$$

(writing m for $E(X)$)

$$\Rightarrow m(7-m) \geq 10.63$$
$$\Rightarrow m \leq 4.77$$

because the quadratic $m(7-m) - 10.63 = 0$ has the two roots $m = 2.23, 4.77$.

It is interesting that this bound is reasonably accurate, as the exact value of $m = E(X)$ *is* $\frac{161}{36} = 4.47$.

4.12 Truncated Distributions

In some applications, an underlying random variable X is observed only when X belongs to a particular set A. For example, in planetary detection studies, due to the current limitations of observational astronomy, a planet's size can be observed only if it is sufficiently large. As another example, suppose automobile accidents are supposed to be reported to the state motor vehicle bureau with estimates of the total amount of damage, but only if the damage exceeds some amount, say 500 dollars. The mathematical formulation of such a problem is that we observe a random variable Y that has the distribution of a latent random variable X *conditioned on the event that X belongs to some set A*. This can be discussed when X is any type of random variable, but we will only discuss the discrete case here.

Definition 4.12. Let X be a discrete random variable with pmf $p(x)$, and let A be a fixed subset of values of the random variable X. The distribution of X conditional on the event that X belongs to A is called *the distribution of X truncated to A* and has the pmf

$$p_A(y) = \frac{p(y)}{P(X \in A)}, y \in A;$$
$$= 0, \text{ if } y \notin A.$$

The mean and the variance of the distribution truncated to A have the following expressions:

mean of the distribution truncated to $A = \mu_A = \frac{\sum_{y \in A} y p(y)}{\sum_{y \in A} p(y)}$,

variance of the distribution truncated to $A = \sigma_A^2 = \frac{\sum_{y \in A} (y - \mu_A)^2 p(y)}{\sum_{y \in A} p(y)}$.

Example 4.36. Suppose a random variable X has the pmf $P(X = n) = \frac{1}{2^n}, n = 1, 2, 3, \ldots,$ but we can observe X only when $X \leq 5$. Thus, the truncation set is $A = \{1, 2, 3, 4, 5\}$. The truncated distribution then has the pmf $p_A(y) = \frac{1/2^y}{\sum_{y=1}^5 1/2^y} = \frac{1/2^y}{31/32} = \frac{2^{5-y}}{31}, y = 1, 2, \ldots, 5$. Thus, the truncated distribution has the pmf

$$p_A(1) = \frac{16}{31}, p_A(2) = \frac{8}{31}, p_A(3) = \frac{4}{31}, p_A(4) = \frac{2}{31}, p_A(5) = \frac{1}{31}.$$

The mean of the truncated distribution is

$$1 \times 16/31 + 2 \times 8/31 + 3 \times 4/31 + 4 \times 2/31 + 1 \times 1/31 = \frac{53}{31} = 1.71;$$

in contrast, X itself has the expected value $\sum_{n=1}^{\infty} n \times \frac{1}{2^n} = 2$. The truncated distribution has a smaller mean because X has been truncated to the small values.

It can be shown that a truncated distribution has a smaller variance than the original distribution. This makes intuitive sense because truncation makes the distribution less dispersed. Theorem 4.10 gives a precise result.

Theorem 4.10 (Chow-Studden Inequality). *Let X be a random variable and a and b any finite real constants. Let $U = \min(X, a), V = \max(X, b)$. Then,*

$$\text{Var}(U) \leq \text{Var}(X); \text{Var}(V) \leq \text{Var}(X).$$

A proof can be found in Chow and Studden (1969).

4.13 Synopsis

(a) For a discrete random variable X taking values $x_1, x_2, \ldots$, the pmf is defined as $p(x) = P(X = x), x = x_1, x_2, \ldots$, and zero otherwise. For any pmf, one must have $p(x) \geq 0$ for any x, and $\sum_i p(x_i) = 1$.

(b) The CDF of a random variable X is defined as $F(x) = P(X \leq x)$. Any CDF is monotonically nondecreasing in x, and $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$. Any CDF is necessarily a right continuous function but need not be left continuous. In fact, the CDF of any discrete random variable X is discontinuous and has jumps at the values of X. The magnitudes of the jumps are the probabilities at those values.

(c) The expected value of a discrete random variable X equals $E(X) = \mu = \sum_i x_i p(x_i)$. The expected value of a function $g(X)$ equals $\sum_i g(x_i) p(x_i)$.

(d) The method of indicator variables and the tail sum method are often extremely effective in calculating expectations of complicated discrete random variables. In particular, the tail sum method says that, for a nonnegative integer-valued random variable X, $E(X) = \sum_{n=0}^{\infty} P(X > n)$.

(e) The variance of a random variable X equals $\text{Var}(X) = \sigma^2 = E(X - \mu)^2 = E(X^2) - \mu^2$. $E(X^2)$ is the second moment of X; more generally, $E(X^k)$ is the kth moment of X. If a particular moment of X exists, then *all* lower-order moments also must exist.

(f) If $X_1, X_2, \ldots, X_k$ are k discrete random variables, then $X_1, X_2, \ldots, X_k$ are called independent if $P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = P(X_1 = x_1) P(X_2 = x_2) \cdots P(X_k = x_k)$ for all $x_1, x_2, \ldots, x_k$. If X and Y are independent random variables, then any function of X and any function of Y are also independent random variables.

(g) For any k random variables $X_1, X_2, \ldots, X_k$, and constants $c_1, c_2, \ldots, c_k$,

$$E\left(\sum_i c_i X_i\right) = \sum_i c_i E(X_i).$$

If $X_1, X_2, \ldots, X_k$ are independent, then

$$\text{Var}\left(\sum_i c_i X_i\right) = \sum_i c_i^2 \text{Var}(X_i).$$

(h) If $X_1, X_2, \ldots, X_k$ are independent, then $E(X_1 X_2 \ldots X_k) = E(X_1)E(X_2)\ldots E(X_k)$.

(i) The square root of the variance is called the standard deviation. Neither the variance nor the standard deviation of a random variable can be negative.
The skewness and the kurtosis of a random variable X are defined as

$$\beta = \frac{E[(X-\mu)^3]}{\sigma^3}; \quad \gamma = \frac{E[(X-\mu)^4]}{\sigma^4} - 3.$$

(j) Four important inequalities are:

$$\text{Markov's inequality for nonnegative random variables } P(X \geq c) \leq \frac{\mu}{c};$$

$$\text{Chebyshev's inequality } P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2};$$

$$\text{the Cauchy-Schwarz inequality } E(|XY|) \leq \sqrt{E(X^2)}\sqrt{E(Y^2)};$$

$$\text{Lyapounov's inequality for } 0 < \alpha < \beta, [E(|X|^\alpha)]^{\frac{1}{\alpha}} \leq [E(|X|^\beta)]^{\frac{1}{\beta}}.$$

4.14 Exercises

Exercise 4.1. Find the pmf and the CDF of the number of heads obtained in four tosses of a fair coin and plot the CDF.

Exercise 4.2. Suppose a fair die is rolled twice and that X is the absolute value of the difference of the two rolls. Find the pmf and the CDF of X and plot the CDF. Find a median of X. Is the median unique?

Exercise 4.3. A motorist encounters four consecutive traffic lights, each equally likely to be red or green. Let X be the number of green lights passed by the motorist before being stopped by a red light. What is the pmf of X?

Exercise 4.4. * **(A Two-Stage Experiment).** Suppose a fair die is rolled once and the number observed is N. Then a fair coin is tossed N times. Let X be the number of heads obtained. Find the pmf, the CDF, and the expected value of X. Does the expected value make sense intuitively?

Exercise 4.5. By using the indicator variable method, find the expected value of the total number of hearts in the hands of North and South combined in a bridge game. Does the expected value make sense intuitively?

Exercise 4.6. * **(Longest Run).** Suppose a fair coin is tossed five times. Find the pmf of the longest run, either of heads or of tails, and compare it with the pmf of the longest head run worked out in the text. Make a comment on what you learn from the comparison.

Exercise 4.7. Suppose X has the pmf $P(X = x) = \frac{c}{1+x^2}, x = \pm 3, \pm 2, \pm 1, 0$, where $c = \frac{5}{13}$. Find the pmf and the expected value of

(a) $h(X) = -1$ if $X < 0; = 0$ if $X = 0; = 1$ if $X > 0$;

(b) $g(X) = \cos\left(\frac{\pi}{2}X\right)$.

Exercise 4.8. * Find a discrete random variable X such that $E(X) = E(X^3) = 0$, $E(X^2) = E(X^4) = 1$.

Exercise 4.9. * **(Waiting Time).** An urn contains four red and four green balls, which are taken out without replacement, one at a time, at random. Let X be the first draw at which a green ball is taken out. Find the pmf and the expected value of X.

Hint: Try to use the tail sum formula or the indicator variable method.

Exercise 4.10. * Suppose a fair die is rolled n times. By using the indicator variable method, find the expected value of the number of faces that appear exactly twice. Now compute the expected value with $n = 2, 3, 5, 10$, and 20.

Exercise 4.11. A prisoner is trapped in a cell containing three doors. The first door leads to a tunnel that returns him to his cell after two days' travel. The second leads to a tunnel that returns him to his cell after four days' travel. The third door leads to freedom after one day of travel. If it is assumed that the prisoner will always select doors 1, 2, and 3 with respective probabilities .3, .5, and .2, what is the expected number of days until the prisoner reaches freedom?

Exercise 4.12 (Strangers' Psychology). A local tavern has six bar stools. The bartender predicts that if two strangers come into the bar, they will sit in such a way as to leave at least two stools between them.

(a) If two strangers do come in but choose their seats at random, what is the probability of the bartender's prediction coming true?

(b) Compute the expected value of the number of stools between the two customers.

Exercise 4.13. Using the general formula given in the text, find the pmf, and hence the expected value of the number of children who do not receive any cookies, if ten cookies are distributed independently to six children.

Exercise 4.14. Suppose a fair die is tossed repeatedly until the first six occurs. Let X be the roll at which the first six is obtained. Find the expected value of X.

Exercise 4.15. * **(Runs).** Suppose a fair die is rolled n times. By using the indicator variable method, find the expected number of times that a six is followed by at least two other sixes. Now compute the value when $n = 100$.

Exercise 4.16. * Suppose ten hunters each target one of 20 ducks flying by, independently, and choose the target duck at random. Each hunter has a 50% probability of actually shooting a duck that is targeted. By using the indicator variable method, show that the expected number of ducks that escape unhurt is about 15.5.

Exercise 4.17. * **(A Problem of Daniel Bernoulli).** Suppose that a jar contains $2N$ cards, two of them marked 1, two marked 2, two marked 3, and so on. Draw out m cards at random. What is the expected number of pairs that still remain in the jar?

(Bernoulli proposed the above as a possible probabilistic model for determining the number of marriages that remain intact when there are a total of m deaths among the N married couples.)

Exercise 4.18. Suppose X can take only the values -1, 0, and 1. If you want to make the mean of X zero and make the variance of X as large as possible, what must be the pmf of X?

Exercise 4.19. * **(Smart Guesses).** A fair die will be rolled once. Before the die is rolled, you have to guess which face will show up. If you underpredict the face by some number n, you will lose $2n$ dollars; if you overpredict the face by n, you will lose n dollars. To minimize your expected loss, what should your guess be?

Exercise 4.20. Suppose a town has ten taxicabs, with license plate numbers $1, 2, \ldots, 10$. You have observed the plate numbers of five of these ten cabs. Let X be the maximum of these five license plate numbers. Find the expected value and the variance of X. State your assumptions.

Exercise 4.21. * **(A Variant of the Birthday Problem).** Guests are pouring in at a party, and someone is writing down each guest's birthday. Let X be the first time a guest is checked in whose birthday is the birthday of someone who has already entered. By using the tail sum formula, prove that the expected value of X is about 24.6. Compare this number with the answer in the birthday problem discussed in the text.

Exercise 4.22 (Random Digit Dialing). The first digit in a ten-digit telephone number is one of $1, 2, \ldots, 9$, and the others can be any of $0, 1, 2, \ldots, 9$. A mischievous teen is dialing one telephone number at random every 30 seconds. He is dialing only within the 555 area code. Your phone number is 555 463 1482. What is the expected number of times the teen will call your number if he keeps dialing for one year for four hours every day?

Exercise 4.23. Suppose X has pmf $P(X = \frac{1}{n}) = \frac{1}{2^n}, n \geq 1$. Find the mean of X.

Exercise 4.24. * Suppose X has pmf $P(X = \frac{1}{n}) = \frac{1}{2^{n+1}}, P(X = n) = \frac{1}{2^{n+1}}$, $n \geq 1$. Find the mean of X.

Exercise 4.25 (A Calculus Calculation). The best quadratic predictor of some random variable Y is $a + bX + cX^2$, where a, b, and c are chosen to minimize $E[(Y - (a + bX + cX^2))^2]$. Determine a, b, and c.

Exercise 4.26. * Suppose a couple will have children until they have at least two children of each sex. By using the tail sum formula, find the expected value of the number of children the couple will have.

Exercise 4.27 (An Important Property of the Mean). Suppose X is a random variable with a finite variance. Show that $E[(X-a)^2]$ is minimized when $a = \mu$, the mean of X.

Exercise 4.28. * (Tail Sum Formula for the Second Moment). Let X be a nonnegative integer-valued random variable. Show that $E(X^2) - E(X) = 2\sum_{n=1}^{\infty} nP(X > n)$.

Exercise 4.29 (Discrete Uniform Distribution). Suppose X has the pmf $P(X = x) = \frac{1}{n}, x = 1, 2, \ldots, n$. Find the mean and variance of X. What are all the medians of X? Is the median unique?

Remark. If $n = 6$, an example of the discrete uniform distribution is the distribution of the number obtained when a fair die is rolled once.

Exercise 4.30. Suppose X is a nonnegative random variable and p is any positive integer. Show that $E(X^p) \geq (E(X))^p$. Can they be equal?

Exercise 4.31. Suppose the IQ scores of a million individuals have a mean of 100 and standard deviation 10.

(a) Without making any further assumptions about the distribution of the scores, find an upper bound on the number of people with an IQ score exceeding 130.
(b) Find a smaller upper bound on the number of scores exceeding 130, assuming the distribution of scores is symmetric about 100.

Exercise 4.32. * (Obtaining Equality in Chebyshev's Inequality). Consider a discrete random variable X with the pmf $P(X = \pm k) = p, P(X = 0) = 1 - 2p$, where k is a fixed positive number, and $0 < p < \frac{1}{2}$.

(a) Find the mean and variance of X.
(b) Find $P(|X - \mu| \geq k\sigma)$.
(c) Can you now choose p in such a way that $P(|X - \mu| \geq k\sigma)$ becomes equal to $\frac{1}{k^2}$?

Exercise 4.33 (A Consequence of the Paley-Zygmund Inequality). Suppose X is a nonnegative random variable, bounded above by some finite positive number M. Prove that

$$P\left(X > \frac{\mu}{2}\right) \geq \frac{\mu}{4M}.$$

Exercise 4.34 (Existence of Some Moments, But Not All). Give an example of a random variable X taking the values $1, 2, 3, \ldots$ such that $E(X^k) < \infty$ for any $k < p$ (p is specified) but $E(X^p) = \infty$.

Exercise 4.35 (Ordering Between Mean and Variance). Give an example of each of the following scenarios:

(a) $\mu > \sigma^2$,
(b) $\mu = \sigma^2$,
(c) $\mu < \sigma^2$.

Exercise 4.36. * **(Standard Deviation vs. Mean Absolute Deviation).** For the *discrete uniform distribution* taking the values $1, 2, \ldots, n$, find the mean absolute deviation and the standard deviation, and plot the standard deviation against the mean absolute deviation by varying the value of n; take $n = 1, 2, \ldots$, and verify that the graph lies above the straight line $y = x$.

Exercise 4.37. Suppose X_1 and X_2 are independent random variables. Show that, for any functions f, g, $f(X_1)$ and $g(X_2)$ are also independent random variables.

Exercise 4.38. * **(Variance of a Product).** Suppose X_1 and X_2 are independent random variables. Give a sufficient condition for it to be true that $\mathrm{Var}(X_1 X_2) = \mathrm{Var}(X_1)\mathrm{Var}(X_2)$.

Exercise 4.39 (Variance of the Number of Heads). By using the formula for the variance of the sum of independent random variables, show that if a coin with probability p for heads in a single toss is tossed n times, then the variance of the number of heads obtained is $np(1 - p)$.

Exercise 4.40 (Use Your Computer). Simulate the elevator stops problem, with 100 floors and 20 passengers. Perform the simulation 500 times, and record the first floor at which the elevator stops to let someone off. Compare your simulation with the theoretical expected value, which was worked out in the text.

Exercise 4.41 (Use Your Computer). Simulate the experiment of dropping $n = 25$ balls independently into $m = 12$ cells. Perform the simulation 500 times, and record the number of cells that remained empty. Compare your simulation with the theoretical distribution, which was worked out in the text. Typically, about how many cells tend to remain empty?

References

Alon, N. and Spencer, J. (2000). *The Probabilistic Method*, Wiley, New York.

Chow, Y. and Studden, W. (1969). Monotonicity of the variance under truncation and variations of Jensen's inequality, Ann. Math. Statist., 40, 1106–1108.

Erdös, P. and Rényi, A. (1970). On a new law of large numbers, J. Anal. Math., 23, 103–111.

Erdös, P. and Révész, P. (1975). On the length of the longest head-run, Colloq. Janos Bolyai Math. Soc., Topics Inform. Theory I., 16, 219–228.

Paley, R.E. and Zygmund, A. (1932). A note on analytic functions in the unit circle, Proc. Cambridge Philos. Soc., 28, 266–272.

Rao, C.R. (1973). *Linear Statistical Inference and Applications*, Wiley, New York.

Chapter 5
Generating Functions

Studying distributions of random variables and their basic quantitative properties, such as expressions for moments, occupies a central role in both statistics and probability. It turns out that a function called the probability generating function is often a very useful mathematical tool in studying distributions of random variables. It is useful to derive formulas for moments and for the pmf of random variables that appear too complicated at first glance. In this chapter, we introduce the probability generating function, study some of its properties, and apply it to a selection of examples. The moment generating function, which is related to the probability generating function, is also extremely useful as a mathematical tool in numerous problems and is also introduced in this chapter. Both the generating function and the moment generating function should be primarily treated as useful tools. They help us solve important problems, and therefore they are useful as mathematical tools.

5.1 Generating Functions

Definition 5.1. The *probability generating function* (pgf), also called simply the *generating function*, of a nonnegative integer-valued random variable X is defined as $G(s) = G_X(s) = E(s^X) = \sum_{x=0}^{\infty} s^x P(X = x)$, provided the expectation is finite.

In this definition, 0^0 is to be understood as being equal to 1. Note that $G(s)$ is always finite for $|s| \le 1$, but it could be finite over a larger interval, depending on the specific random variable X.

Two basic properties of the generating function are the following.

Theorem 5.1.

(a) *Suppose $G(s)$ is finite in some open interval containing the origin. Then, $G(s)$ is infinitely differentiable in that open interval, and $P(X = k) = \frac{G^{(k)}(0)}{k!}, k \ge 0$, where $G^{(0)}(0)$ means $G(0)$.*

(b) *If $\lim_{s\uparrow 1} G^{(k)}(s)$ is finite, then $E[X(X-1)\ldots(X-k+1)]$ exists and is finite, and $G^{(k)}(1) = \lim_{s\uparrow 1} G^{(k)}(s) = E[X(X-1)\ldots(X-k+1)]$.*

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_5,

Proof. The infinite differentiability is a fact from the theory of power series that converge in some nonempty open interval. The power series can be differentiated infinitely many times term by term in that open interval. That $P(X = k) = \frac{G^{(k)}(0)}{k!}$ follows on differentiating $G(s)$ term by term k times and setting $s = 0$, while part (b) follows on differentiating $G(s)$ k times and letting $s \to 1$.

Definition 5.2. $E[X(X-1)\cdots(X-k+1)]$ is called the kth *factorial moment* of X.

Remark. The kth factorial moment of X exists if and only if the kth moment $E(X^k)$ exists.

One of the most important properties of generating functions is the following.

Theorem 5.2. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with generating functions $G_1(s), G_2(s), \ldots, G_n(s)$. Then the generating function of $X_1 + X_2 + \cdots + X_n$ equals*

$$G_{X_1+X_2+\cdots+X_n}(s) = \prod_{i=1}^{n} G_i(s).$$

Proof. By definition,

$$\begin{aligned} G_{X_1+X_2+\cdots+X_n}(s) &= E[s^{X_1+X_2+\cdots+X_n}] = E[s^{X_1}s^{X_2}\cdots s^{X_n}] \\ &= E[s^{X_1}]E[s^{X_2}]\cdots E[s^{X_n}] = \prod_{i=1}^{n} G_i(s) \end{aligned}$$

by virtue of the independence of $X_1, X_2, \ldots, X_n$, which would imply the independence of $s^{X_1}, s^{X_2}, \ldots, s^{X_n}$.

One reason that the generating function is useful as a tool is its *distribution determining property*, in the following sense.

Theorem 5.3. *Let $G(s)$ and $H(s)$ be the generating functions of two random variables X and Y. If $G(s) = H(s)$ in any nonempty open interval, then X and Y have the same distribution.*

Proof. Let $P(X = n) = p_n, P(Y = n) = q_n, n \geq 0$. Then, $G(s) = \sum_{n=0}^{\infty} s^n p_n$, and $H(s) = \sum_{n=0}^{\infty} s^n q_n$. If there is a nonempty open interval in which $\sum_{n=0}^{\infty} s^n p_n = \sum_{n=0}^{\infty} s^n q_n$, then from the theory of power series, $p_n = q_n \; \forall n \geq 0$, and therefore X and Y have the same distribution.

Summarizing, then, one can find from the generating function of a nonnegative integer-valued random variable X the pmf of X and every moment of X, including the moments that are infinite.

Example 5.1 (Discrete Uniform Distribution). Suppose X has the discrete uniform distribution on $\{1, 2, \ldots, n\}$. Then, its generating function is

$$G(s) = E[s^X] = \sum_{x=1}^{n} s^x P(X = x) = \frac{1}{n} \sum_{x=1}^{n} s^x = \frac{s(s^n - 1)}{n(s - 1)}$$

by summing the geometric series $\sum_{x=1}^{n} s^x$. As a check, if we differentiate $G(s)$ once, we get

$$G'(s) = \frac{1 + s^n[n(s - 1) - 1]}{n(s - 1)^2}.$$

On applying L'Hospital's rule, we get that $G'(1) = \frac{n+1}{2}$, which therefore is the mean of X.

Example 5.2. Let $G(s) = \frac{(1+s)^n}{2^n}$. Then, by just expanding $(1 + s)^n$ using the binomial theorem, we have

$$G(s) = \frac{1}{2^n} \sum_{x=0}^{n} \binom{n}{x} s^x.$$

We now recognize that the coefficients $\frac{\binom{n}{x}}{2^n}, x = 0, 1, 2, \ldots n$, are all nonnegative and that they do add to one. Therefore, $G(s) = \frac{(1+s)^n}{2^n}$ is a generating function, and indeed it is the generating function of the random variable X with the pmf $P(X = x) = \frac{\binom{n}{x}}{2^n}, x = 0, 1, 2, \ldots n$, which is the binomial random variable with parameters n and $\frac{1}{2}$.

Example 5.3 (The Poisson Distribution). Consider a nonnegative integer-valued random variable X with the pmf $p(x) = e^{-1}\frac{1}{x!}, x = 0, 1, 2, \ldots$. This is indeed a valid pmf. First, it is clear that $p(x) \geq 0$ for any x. Also,

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} e^{-1} \frac{1}{x!} = e^{-1} \sum_{x=0}^{\infty} \frac{1}{x!} = e^{-1} e = 1.$$

We will find the generating function of this distribution. The generating function is

$$G(s) = E[s^X] = \sum_{x=0}^{\infty} s^x e^{-1} \frac{1}{x!}$$

$$= e^{-1} \sum_{x=0}^{\infty} \frac{s^x}{x!} = e^{-1} e^s = e^{s-1}.$$

The first derivative of $G(s)$ is $G'(s) = e^{s-1}$, and therefore $G'(1) = e^0 = 1$. From our theorem above, we conclude that $E(X) = 1$. *Indeed, the pmf that we have in this example is the pmf of the so-called Poisson distribution with mean one.* The pmf

of the Poisson distribution with a general mean λ is $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots$. The Poisson distribution is an extremely important distribution in probability theory and will be studied in detail in Chapter 6.

Example 5.4 (The Exact Distribution of the Sum of n Dice Rolls). This is a more advanced example. The distribution of the sum of two dice rolls was found in Chapter 2 by direct enumeration. It is interesting that the distribution of the sum of n dice rolls for a general n can be found by using the generating function of the sum. The derivation needs quite a bit of clever algebraic manipulation, and it is really impressive that de Moivre derived this distribution by using the generating function. The example also demonstrates the power of the generating function as a problem-solving tool.

Let X_i be the number obtained on the ith roll of the die. Each X_i has a discrete uniform distribution on $\{1, 2, \ldots, 6\}$, and therefore, using the discrete uniform example above, the generating function of $X_1 + X_2 + \cdots + X_n$ is

$$G(s) = \frac{s^n(s^6 - 1)^n}{6^n(s-1)^n} = \frac{s^n(1 + s + s^2 + \cdots + +s^5)^n}{6^n}.$$

On expanding $(1 + s + s^2 + \cdots + +s^5)^n$ and collecting terms, after some algebra, one gets

$$G(s) = \frac{s^n \sum_{i=0}^{n}(-1)^i \binom{n}{i} s^{6i} \times \sum_{j=0}^{\infty} \binom{n+j-1}{j} s^j}{6^n}.$$

Now make the change of variable $k = n + 6i + j$ in the numerator of this expression. Then, $G(s)$ reduces to

$$G(s) = \frac{1}{6^n} \sum_{k=n}^{6n} s^k a_k,$$

where $a_k = \sum_{i=0}^{\lfloor (k-n)/6 \rfloor}(-1)^i \binom{n}{i}\binom{k-6i-1}{n-1}$, and therefore, for $n \le k \le 6n$, $P(X_1 + X_2 + \cdots + X_n = k)$ must be

$$p_k = p_{k,n} = \frac{1}{6^n} \sum_{i=0}^{\lfloor (k-n)/6 \rfloor} (-1)^i \binom{n}{i}\binom{k-6i-1}{n-1},$$

where the notation $\lfloor\ \rfloor$ denotes the integer part function. If $k < n$ or $> 6n$, then $P(X_1 + X_2 + \cdots + X_n = k)$ is of course zero.

It is quite remarkable that an exact formula for the distribution of the sum can be found for an arbitrary number of rolls of the die, and the formula, due to de Moivre, is completely classic.

Using de Moivre's formula for the pmf of the sum of n dice rolls, we plot in Figure 5.1 the pmf when five dice are rolled. The pmf has a beautiful bell shape. We will later see a connection of this visual finding to a result known as the *central limit*

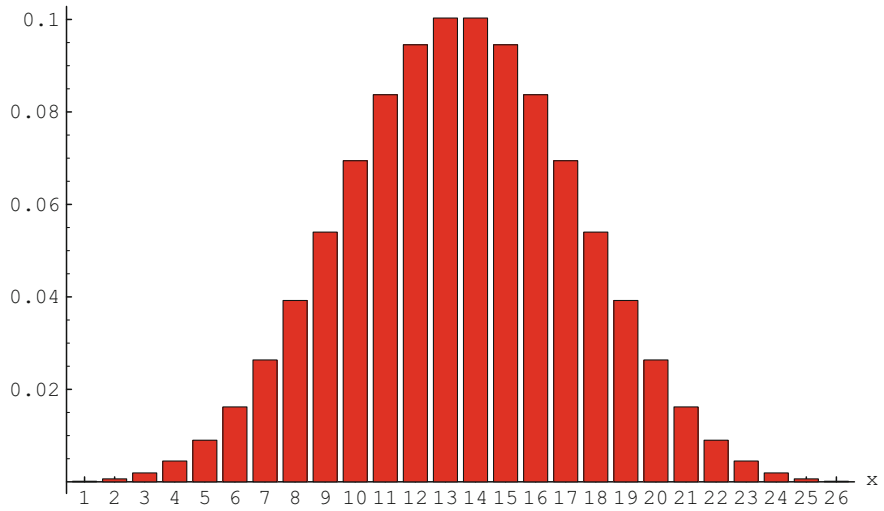


Fig. 5.1 pmf of sum of five dice; $x = 1$ in plot means sum$= 5$

theorem. It should be noted that the x values in the plot start at one, an idiosyncrasy of the software; thus $x = 1$ in the plot corresponds to a sum of 5, and $x = 2$ corresponds to a sum of 6, etc. It is the bell shape that we want to emphasize in this example.

As an illustration, suppose now that we want to know the chances that the sum of 20 rolls of a fair die will be between 64 and 76. Then, applying the exact formula above with $n = 20$, the required probability is $\sum_{k=64}^{76} p_{k,20} = .6027$. It would of course be impossible to do this by direct enumeration.

5.2 Moment Generating Functions and Cumulants

We have defined the probability generating function only for nonnegative integer-valued random variables. The moment generating function is usually discussed in the context of general random variables, not necessarily integer-valued, or discrete. The two functions are connected. Here is the formal definition.

Definition 5.3. Let X be a real-valued random variable. The moment generating function (mgf) of X is defined as

$$\psi_X(t) = \psi(t) = E[e^{tX}],$$

whenever the expectation is finite.

Note that the mgf $\psi(t)$ of a random variable X *always* exists and is finite if $t = 0$ and $\psi(0) = 1$. It may or may not exist when $t \neq 0$. If it does exist for t in a

nonempty open interval containing zero, then many properties of X can be derived by using the mgf $\psi(t)$; it is an extremely useful tool. If X is a nonnegative integer-valued random variable, then writing s^X as $e^{X \log s}$, it follows that the (probability) generating function $G(s)$ is equal to $\psi(\log s)$ whenever $G(s) < \infty$. Thus, the two generating functions, namely the probability generating function and the moment generating function, are connected.

The following theorem explains the name moment generating function.

Theorem 5.4.

(a) *Suppose the mgf $\psi(t)$ of a random variable X is finite in some open interval containing zero. Then, $\psi(t)$ is infinitely differentiable in that open interval, and for any $k \geq 1$*

$$E(X^k) = \psi^{(k)}(0).$$

(b) **(Distribution Determining Property).** *If $\psi_1(t)$ and $\psi_2(t)$ are the mgfs of two random variables X and Y, and if $\psi_1(t) = \psi_2(t)$ in some nonempty open interval containing zero, then X and Y have the same distribution.*

(c) *If $X_1, X_2, \ldots, X_n$ are independent random variables, and if each X_i has an mgf $\psi_i(t)$ existing in some open interval around zero, then $X_1 + X_2 + \cdots + X_n$ also has an mgf in that open interval, and*

$$\psi_{X_1+X_2+\cdots+X_n}(t) = \prod_{i=1}^{n} \psi_i(t).$$

Proof. The proof is formally similar to the proof of the corresponding result for (probability) generating functions. If $\psi(t)$ exists in an open interval, then it admits a power series expansion in that open interval, and it can be differentiated infinitely often in that open interval. Furthermore,

$$\frac{d^k}{dt^k} E[e^{tX}] = E\left[\frac{d^k}{dt^k} e^{tX}\right] = E[X^k e^{tX}]$$

$\Rightarrow \psi^{(k)}(0) = E[X^k]$ for any $k \geq 1$. The formal proofs of parts (b) and (c) are similar to the corresponding proofs for generating functions and are therefore omitted. However, a mathematically rigorous proof of this theorem requires the use of complex analysis, a branch of more advanced mathematics.

Let us now see a few examples.

Example 5.5 (Discrete Uniform Distribution). Let X have the pmf $P(X = x) = \frac{1}{n}, x = 1, 2, \ldots, n$. Then, its mgf is

$$\psi(t) = E[e^{tX}] = \frac{1}{n} \sum_{k=1}^{n} e^{tk} = \frac{e^t(e^{nt} - 1)}{n(e^t - 1)}.$$

By direct differentiation,

$$\psi'(t) = \frac{e^t(1 + ne^{(n+1)t} - (n+1)e^{nt})}{n(e^t - 1)^2}.$$

On applying L'Hospital's rule twice, we get the previously derived fact that $E(X) = \frac{n+1}{2}$.

Example 5.6. Suppose X takes only two values, 0 and 1, with $P(X=1)=p$, $P(X = 0) = 1 - p, 0 < p < 1$. Thus, X is a Bernoulli variable with parameter p. Then, the mgf of X is

$$\psi(t) = E[e^{tX}] = pe^t + (1 - p).$$

If we differentiate this, we get $\psi'(t) = pe^t, \psi''(t) = pe^t$. Therefore, $\psi'(0) = pe^0 = p$, and also $\psi''(0) = p$. From the general properties of mgfs, it then follows that $E(X) = \psi'(0) = p$ and $E(X^2) = \psi''(0) = p$. Now go back to the pmf of X that we started with in this example, and note that indeed, by direct calculation, $E(X) = E(X^2) = p$.

Example 5.7. Suppose X is the sum of two rolls of a fair die. Then, X can be written as $X = X_1 + X_2$, where X_1 and X_2 are the numbers obtained on the two rolls, respectively. The mgf of each of X_1 and X_2 is obtained from the general mgf for the discrete uniform distribution worked out above using $n = 6$. By part (c) of the preceding theorem, we have

$$\psi_X(t) = \left[\frac{e^t(e^{6t} - 1)}{6(e^t - 1)}\right]^2 = \frac{e^{2t}(e^{6t} - 1)^2}{36(e^t - 1)^2}.$$

5.2.1 ∗ Cumulants

Closely related to the moments of a random variable are certain quantities known as *cumulants*. Cumulants arise in accurate approximations of the distribution of sums of independent random variables. They are also used for statistical modeling purposes in some applied sciences. The name *cumulant* was coined by Sir Ronald Fisher (1929), although it was discussed in the literature by others prior to Fisher's coining of the cumulant term. We will define and describe some basic facts about cumulants below; this material is primarily for reference purposes and may be omitted at first reading.

We need to define *central moments* of a random variable first because cumulants are related to them.

Definition 5.4. Let a random variable X have a finite jth moment for some specified $j \geq 1$. The jth *central moment* of X is defined as $\mu_j = E[(X-\mu)^j]$, where $\mu = E(X)$.

Remark. Note that $\mu_1 = E(X-\mu) = 0$, and $\mu_2 = E(X-\mu)^2 = \sigma^2$, the variance of X. If X has a distribution *symmetric about zero*, then every odd-order central moment, $E[(X-\mu)^{2k+1}]$, is easily proved to be zero, provided it exists.

Definition 5.5. Let X have a finite mgf $\psi(t)$ in some neighborhood of zero, and let $K(t) = \log \psi(t)$ when it exists. The rth cumulant of X is defined as $\kappa_r = \frac{d^r}{dt^r} K(t)|_{t=0}$. Equivalently, the cumulants of X are the coefficients in the power series expansion $K(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}$ within its radius of convergence.

Note that $K(t) = \log \psi(t)$ implies that $e^{K(t)} = \psi(t)$. By equating coefficients in the power series expansion of $e^{K(t)}$ with those in the power series expansion of $\psi(t)$, it is easy to express the first few moments (and therefore the first few central moments) in terms of the cumulants. Indeed, denoting $c_i = E(X^i), \mu = E(X) = c_1, \mu_i = E(X-\mu)^i, \sigma^2 = \mu_2$, one obtains the expressions

$$c_1 = \kappa_1; c_2 = \kappa_2 + \kappa_1^2; c_3 = \kappa_3 + 3\kappa_1\kappa_2 + \kappa_1^3;$$
$$c_4 = \kappa_4 + 4\kappa_1\kappa_3 + 3\kappa_2^2 + 6\kappa_1^2\kappa_2 + \kappa_1^4.$$

The corresponding expressions for the central moments are much simpler:

$$\sigma^2 = \kappa_2; \mu_3 = \kappa_3; \mu_4 = \kappa_4 + 3\kappa_2^2.$$

In general, the cumulants satisfy the recursion relations

$$\kappa_n = c_n - \sum_{k=1}^{n-1} \binom{n-1}{k-1} c_{n-k}\kappa_k.$$

These result in the specific expressions

$$\kappa_2 = \mu_2; \kappa_3 = \mu_3; \kappa_4 = \mu_4 - 3\mu_2^2.$$

High-order cumulants have quite complex expressions in terms of the central moments μ_j; the corresponding expressions in terms of the c_j are even more complex.

The derivations of these expressions stated above involve straight differentiation. We will not present the algebra. It is useful to know these expressions for some problems in statistics.

5.3 Synopsis

(a) The generating function of a nonnegative integer-valued random variable X is defined as $G_X(s) = E(s^X)$. It exists for $|s| \leq 1$ but may or may not exist for $|s| > 1$.
(b) If two random variables X and Y have the same generating function in an open interval containing zero, then they must have the same distribution.
(c) For a nonnegative integer-valued random variable X, $P(X = k) = \frac{G^{(k)}(0)}{k!}$, $k \geq 0$.
(d) If $X_1, X_2, \ldots, X_n$ are independent random variables, then the generating function of $X_1 + X_2 + \cdots + X_n$ equals $\prod_{i=1}^{n} G_i(s)$.
(e) The mgf of a real-valued random variable X is defined as $\psi_X(t) = E[e^{tX}]$. It exists when $t = 0$ and always $\psi_X(0) = 1$. It may or may not exist for $t \neq 0$.
(g) If two random variables X and Y have the same mgf in an open interval containing zero, then they must have the same distribution.
(g) If the mgf $\psi(t)$ of a random variable X is finite in some open interval containing zero, then $E(X^k) = \psi^{(k)}(0)$.
(h) If $X_1, X_2, \ldots, X_n$ are independent random variables, and if each X_i has an mgf $\psi_i(t)$, then the mgf of $X_1 + X_2 + \cdots + X_n$ equals $\psi_{X_1+X_2+\cdots+X_n}(t) = \prod_{i=1}^{n} \psi_i(t)$.

5.4 Exercises

Exercise 5.1. Find the generating function and the mgf of the random variable X with the pmf $P(X = n) = \frac{1}{2^n}, n = 1, 2, 3, \ldots$.

Exercise 5.2. * Give an example of a function $G(s)$ such that $G(0) \geq 0$, $G'(1) > 0$, but $G(s)$ is not the generating function of any nonnegative integer-valued random variable.

Exercise 5.3 (Generating Function of a Linear Function). Suppose X has the generating function $G(s)$. What are the generating functions of $X \pm 1$? Of $2X$?

Exercise 5.4 (MGF of a Linear Function). Suppose X has the mgf $\psi(t)$. Find an expression for the mgf of $aX + b$, where a and b are real constants.

Exercise 5.5. * Suppose X is a nonnegative random variable with a finite mgf at some point t. Prove or disprove that $\sqrt{X}$ also has a finite mgf at that point t.

Exercise 5.6. * Give an example of a random variable X such that X has a finite mgf at any t but X^2 does not have a finite mgf at any $t > 0$.

Exercise 5.7 (Generating Function and Moments). Suppose X has the generating function $G(s)$. Express the variance and the third moment of X in terms of $G(s)$ and its derivatives.

Exercise 5.8. Suppose $G(s)$ and $H(s)$ are both generating functions. Show that $pG(s) + (1-p)H(s)$ is also a valid generating function for any p in $(0,1)$. What is an interesting interpretation of the distribution that has $pG(s) + (1-p)H(s)$ as its generating function?

Exercise 5.9 (Convexity of the MGF). Suppose X has the mgf $\psi(t)$, finite in some open interval. Show that $\psi(t)$ is convex in that open interval.

Exercise 5.10. Find the first four moments, the first four central moments, and the first four cumulants of X, where X is the number of heads obtained in three tosses of a fair coin, and verify all the interrelationships between them stated in the text.

Exercise 5.11. * **(Cumulants of a Bernoulli Variable).** Suppose X has the pmf $P(X=1) = p$, $P(X=0) = 1-p$. What are the first four cumulants of X?

Exercise 5.12. Suppose X has a symmetric distribution $P(X = \pm 1) = p$, $P(X=0) = 1-2p$. What are its first four cumulants?

References

Fisher, R.A. (1929). Moments and product moments of sampling distributions, Proc. London Math. Soc., 2, 199–238.

Chapter 6
Standard Discrete Distributions

A few special discrete distributions arise very frequently in applications. Either the underlying probability mechanism of a problem is such that one of these distributions is truly the correct distribution for that problem or the problem may be such that one of these distributions is a very good choice to model that problem. We present these distributions and study their basic properties in this chapter; they deserve the special attention because of their importance in applications. The special distributions we present are the discrete uniform, binomial, geometric, negative binomial, hypergeometric, and Poisson. Benford's distribution is also covered briefly. A few other special distributions are covered in the chapter exercises.

6.1 Introduction to Special Distributions

We first provide the pmfs of these special distributions and a quick description of the contexts where they are relevant. We will then study these distributions in detail in later sections.

The Discrete Uniform Distribution. The discrete uniform distribution represents a finite number of equally likely values. The simplest real-life example is the face obtained when a fair die is rolled once. It can also occur in some other physical phenomena, particularly when the number of possible values is small and the scientist feels that they are just equally likely. If we let the values of the random variable be $1, 2, \ldots, n$, then the pmf of the discrete uniform distribution is $p(x) = \frac{1}{n}, x = 1, 2, \ldots, n$, We sometimes write $X \sim Unif\{1, 2, \ldots, n\}$.

The Binomial Distribution. The binomial distribution represents a sequence of independent coin-tossing experiments. Suppose a coin with probability $p, 0 < p < 1$, for heads in a single trial is tossed independently a prespecified number of times, say n times, $n \geq 1$. Let X be the number of times in these n tosses that a head is obtained. Then the pmf of X is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \ldots, n,$$

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_6,

the $\binom{n}{x}$ term giving the choice of the x tosses out of the n tosses in which the heads occur.

Coin tossing, of course, is just an artifact. Suppose a trial can result in only one of two outcomes, called a *success* (S) or a *failure* (F), the probability of obtaining a success being p in any trial. Such a trial is called a *Bernoulli trial*. Suppose a Bernoulli trial is repeated independently a prespecified number of times, say n times. Let X be the number of times in the n trials that a success is obtained. Then X has the pmf given above, and we say that X has a *binomial distribution with parameters n and p* and write $X \sim Bin(n, p)$.

The Geometric Distribution. Suppose a coin with probability $p, 0 < p < 1$ for heads in a single trial is tossed repeatedly until a head is obtained for the first time. Assume that the tosses are independent. Let X be the number of the toss at which the very first head is obtained. Then the pmf of X is

$$P(X = x) = p(1 - p)^{x-1}, x = 1, 2, 3, \ldots.$$

We say that X has a *geometric distribution with parameter p*, and we will write $X \sim Geo(p)$. The distinction between the binomial distribution and the geometric distribution is that in the binomial case the number of tosses is *prespecified*, but in the geometric case the number of tosses actually performed when the experiment ends is a random variable. A geometric distribution measures a waiting time for the first success in a sequence of independent Bernoulli trials, each with the same success probability p; i.e., the coin cannot change from one toss to another.

The Negative Binomial Distribution. The negative binomial distribution is a generalization of a geometric distribution when we repeatedly toss a coin with probability p for heads, independently, until a total number of r heads has been obtained, where r is some fixed integer ≥ 1. The case $r = 1$ corresponds to the geometric distribution. Let X be the number of the first toss at which the rth success is obtained. Then the pmf of X is

$$P(X = x) = \binom{x-1}{r-1} p^r (1 - p)^{x-r}, x = r, r + 1, \ldots,$$

the term $\binom{x-1}{r-1}$ simply giving the choice of the $r - 1$ tosses among the first $x - 1$ tosses where the first $r - 1$ heads were obtained. We say that X has a *negative binomial distribution with parameters r and p*, and we will write $X \sim NB(r, p)$.

The Hypergeometric Distribution. The hypergeometric distribution also represents the number of successes in a prespecified number of Bernoulli trials, but the trials happen to be dependent. A typical example is that of a finite population in which there are in all N objects, of which some D are of type I and the other $N - D$ are of type II. A *sample without replacement* of size $n, 1 \leq n < N$, is chosen at random from the population. Thus, the selected sampling units are necessarily different. Let

X be the number of units or individuals of type I among the n units chosen. Then the pmf of X is

$$P(X = x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}},$$

$n - N + D \leq x \leq D$; note that, trivially, x is also ≥ 0 and $\leq n$. An example would be that of a pollster polling $n = 100$ people from a population of 10,000 people, where $D = 5500$ are in favor of some proposition and the remaining $N - D = 4500$ are against it. The number of individuals in the sample who are in favor of the proposition then has the pmf above. We say that such an X has a *hypergeometric distribution with parameters* n, D, N, and we will write $X \sim Hypergeo(n, D, N)$.

The Poisson Distribution. The Poisson distribution is perhaps the most used and useful distribution for modeling nonnegative integer-valued random variables. Unlike the first four distributions above, we cannot say that a Poisson distribution is necessarily the correct distribution for some integer-valued random variable. Rather, a Poisson distribution is chosen by a scientist as his or her model for the distribution of an integer-valued random variable. But the choice of the Poisson distribution as a model is frequently extremely successful in describing and predicting how the random variable behaves. The Poisson distribution also arises, as a mathematical fact, as the *limiting distribution* of numerous integer-valued random variables when in some sense a sequence of Bernoulli trials makes it increasingly harder to obtain a success; i.e., the number of times a very rare event happens if we observe the process for a long time *often* has an approximately Poisson distribution.

The pmf of a *Poisson distribution with parameter* λ is

$$P(X = x), \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots;$$

by using the power series expansion of $e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$, it follows that this is indeed a valid pmf.

Three specific situations where a Poisson distribution is almost routinely adopted as a model are the following:

(a) The number of times a specific event happens in a specified period of time; e.g., the number of phone calls received by someone over a 24 hour period.
(b) The number of times a specific event or phenomenon is observed in a specified amount of area or volume; e.g., the number of bacteria of a certain kind in one liter of a sample of water, the number of misprints per page of a book, etc.
(c) The number of times a success is obtained when a Bernoulli trial with success probability p is repeated independently n times, with p being small and n being large, such that the product np has a *moderate value*, say between .5 and 10.

We now treat these distributions in greater detail one at a time.

6.2 Discrete Uniform Distribution

Definition 6.1. The discrete uniform distribution on $\{1, 2, \ldots, n\}$ is defined by the pmf $P(X = x) = \frac{1}{n}, x = 1, 2, \ldots, n$, and zero otherwise. Of course, the set of values can be any finite set; we take the values to be $1, 2, \ldots, n$ for convenience.

Clearly, for any given integer $k, 1 \le k \le n, F(k) = P(X \le k) = \frac{k}{n}$. The first few moments are found easily. For example,

$$\mu = E(X) = \sum_{x=1}^{n} xp(x) = \sum_{x=1}^{n} x\frac{1}{n} = \frac{1}{n}\sum_{x=1}^{n} x$$
$$= \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}.$$

Similarly,

$$E(X^2) = \sum_{x=1}^{n} x^2 p(x) = \frac{1}{n}\sum_{x=1}^{n} x^2$$
$$= \frac{1}{n}\frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}.$$

Therefore,

$$\sigma^2 = \text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}.$$

It follows from the trivial symmetric nature of the discrete uniform distribution that $E(X - \mu)^3 = 0$. We can also find $E(X - \mu)^4$ in closed form. For this, the only additional fact that we need is that $\sum_{x=1}^{n} x^4 = (\frac{n(n+1)}{2})^2$. Then, by expanding $(X - \mu)^4$, after some algebra it follows that

$$E(X - \mu)^4 = \frac{(3n^2-7)(n^2-1)}{240}.$$

The moment information about the discrete uniform distribution is collected together in the theorem below.

Theorem 6.1. *Let $X \sim Unif\{1, 2, \ldots, n\}$. Then,*

$$\mu = E(X) = \frac{n+1}{2}; \sigma^2 = \text{Var}(X) = \frac{n^2-1}{12}; E(X-\mu)^3 = 0;$$
$$E(X - \mu)^4 = \frac{(3n^2-7)(n^2-1)}{240}.$$

Corollary 6.1. *The skewness and the kurtosis of the discrete uniform distribution are*

$$\beta = 0; \gamma = -\frac{6}{5}\frac{n^2+1}{n^2-1}.$$

6.3 Binomial Distribution

We start with a few examples.

Example 6.1 (Heads in Coin Tosses). Suppose a fair coin is tossed ten times, independently, and suppose X is the number of times in the ten tosses that a head is obtained. Then $X \sim Bin(n, p)$ with $n = 10$, $p = \frac{1}{2}$. Therefore,

$$P(X = x) = \binom{10}{x}\left(\frac{1}{2}\right)^{10}, x = 0, 1, 2, \ldots, 10.$$

Converting to decimals, the pmf of X is

x	0	1	2	3	4	5	6	7	8	9	10
$P(X = x)$	.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010

Note that the pmf is *symmetric* about $x = 5$ and that $P(X = x)$ increases from $x = 0$ to $x = 5$ and then decreases from $x = 5$ to $x = 10$ symmetrically.

Example 6.2 (Guessing on a Multiple-Choice Exam). A multiple-choice test with 20 questions has five possible answers for each question. A completely unprepared student picks the answer for each question at random and independently. Suppose X is the number of questions that the student answers correctly.

We identify each question with a Bernoulli trial and a correct answer as a success. Since there are 20 questions and the student picks his answer at random from five choices, $X \sim Bin(n, p)$, with $n = 20, p = \frac{1}{5} = .2$. We can now answer any question we want about X.

For example,

$$P(\text{The student gets every answer wrong}) = P(X = 0) = .8^{20} = .0115,$$

while

$$P(\text{The student gets every answer right}) = P(X = 20) = .2^{20} = 1.05 \times 10^{-14},$$

a near impossibility. Suppose the instructor has decided that it will take at least 13 correct answers to pass this test. Then,

$$P(\text{The student will pass}) = \sum_{x=13}^{20} \binom{20}{x} .2^x .8^{20-x} = .000015,$$

still a very small probability.

Example 6.3 (To Cheat or Not to Cheat). Ms. Smith drives into town once a week to buy groceries. In the past she parked her car at a lot for five dollars, but she decided that for the next five weeks she will park at the fire hydrant and risk getting tickets with fines of 25 dollars per offense. If the probability of getting a ticket is .1, what is the probability that she will pay more in fines in five weeks than she would pay in parking fees if she had opted not to park by the fire hydrant?

Suppose that X is the number of weeks among the next five weeks in which she gets a ticket. Then, $X \sim Bin(5, .1)$. Ms. Smith's parking fees would have been 25 dollars for the five weeks combined if she did not park by the hydrant. Thus, the required probability is

$$\begin{aligned} P(25X > 25) = P(X > 1) &= 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - \left[.9^5 + \binom{5}{1} .1(.9)^4 \right] = .0815. \end{aligned}$$

So the chances are quite low that Ms. Smith will pay more in tickets by breaking the law than she would pay by paying the parking fees.

Example 6.4. Suppose a fair coin is tossed $n = 2m$ times. What is the probability that the number of heads obtained will be an even number?

Since $X =$ the number of heads $\sim Bin(2m, \frac{1}{2})$, we want to find

$$P(X = 0) + P(X = 2) + \cdots + P(X = 2m) = \sum_{x=0}^{m} \binom{2m}{2x} \Big/ 2^{2m} = 2^{2m-1}/2^{2m}$$

$= \frac{1}{2}$ on using the identity that, for any n,

$$\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots = 2^{n-1}.$$

Thus, with a fair coin, the chances of getting an even number of heads in an even number of tosses are $\frac{1}{2}$. The same is true also if the number of tosses is odd and is proved similarly.

Example 6.5 (Flush in Poker). A flush in five-card poker is five cards of the same suit but not in a sequence. We saw in Chapter 1 that the probability of obtaining a flush in five-card poker is .00197.

Suppose someone plays poker once a week every week for a year, and each time that he plays, he plays four deals. Let X be the number of times he obtains a flush during the year. Assuming that decks are always well shuffled between plays, $X \sim Bin(n, p)$, where $n = 52 \times 4 = 208$ and $p = .00197$. Then, $P(X \geq 1) = 1 - (1 - .00197)^{208} = .3365$. So there is about a one in three chance that the player will obtain a flush within a year.

In this example, n was large and p was small. In such cases, the $Bin(n, p)$ distribution can be well approximated by a Poisson distribution with $\lambda = np$. If we do the approximation, we will get $P(X \geq 1) \approx 1 - e^{-208 \times .00197} = 1 - e^{-.40976} = .3362$, clearly a very close approximation to the exact value .3365. We will discuss Poisson approximations of binomials in greater detail later in this chapter.

Example 6.6 (A Stock Inventory Example). This example takes a little more careful reading because the formulation is a little harder. Here is the problem. Professor Rubin loves diet soda. Twice a day he drinks an 8 oz. can of diet soda, and each time he reaches at random into one of two brown bags containing Diet Coke and Diet Pepsi, respectively. One box of soda picked up at a supermarket has six soda cans. How many boxes of each type of soda should professor Rubin buy per to be 90% sure that he will not find a brown bag empty when he reaches into it?

Let $X =$ the number of times Professor Rubin reaches to find a Diet Coke; then, $X \sim Bin(n, p)$ with $n = 14$ and $p = .5$. Since $p = .5, n - X$ is also distributed as the same binomial, namely $Bin(n, p)$, with $n = 14$ and $p = .5$. Suppose Professor Rubin has N sodas of each type in stock. We want $P(X > N) + P(n - X > N) \leq .1$. Now,

$$\begin{aligned} P(X > N) + P(n - X > N) &= 2 \sum_{x=N+1}^{n} \binom{n}{x} (.5)^n \\ &= 2 \sum_{x=N+1}^{14} \binom{14}{x} (.5)^{14} = g(N), \end{aligned}$$

say. By computing it, we find that $g(9) = .18$ and $g(10) = .06 < .1$. Therefore, Professor Rubin needs to have ten sodas of each type (that is, two boxes of each type of soda) in stock each week.

Example 6.7 (Flukes are Easier in the Short Run). Suppose two tennis players, A and B, will play an odd number of games, and whoever wins a majority of the games will be the winner. Suppose that A is a better player, and A has a probability of .6 of winning any single game. If B were to win this tournament, it might be considered a fluke.

Suppose that they were to play three games. Let X be the number of games won by B. Under the usual assumptions of independence, $X \sim Bin(n, p)$ with $n = 3, p = .4$. Thus, the chances of B winning the tournament are

$$P(X \geq 2) = 3(.4)^2(.6) + .4^3 = .352.$$

Suppose next that they were to play nine games. Now, $X \sim Bin(n, p)$ with $n = 9$, $p = .4$, so the chances of B winning the tournament are

$$P(X \geq 5) = \sum_{x=5}^{9} \binom{9}{x} (.4)^x (.6)^{9-x} = .2665.$$

We see that the chances of B winning the tournament go down when they play more games. This is because a weaker player can get lucky in the short run, but the luck will run out in the long run.

Some key mathematical facts about a binomial distribution are given in the following theorem.

Theorem 6.2. *Let $X \sim Bin(n, p)$. Then,*

(a) $\mu = E(X) = np; \sigma^2 = Var(X) = np(1-p)$.
(b) *The mgf of X equals $\psi(t) = (pe^t + 1 - p)^n$ at any t.*
(c) $E[(X-\mu)^3] = np(1 - 3p + 2p^2)$.
(d) $E[(X-\mu)^4] = np(1-p)[1 + 3(n-2)p(1-p)]$.

Proof. By writing X as $X = \sum_{i=1}^{n} I_{A_i}$, where A_i is the event of a success on the ith Bernoulli trial, it follows readily that $E(X) = \sum_{i=1}^{n} P(A_i) = np$ and $\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(I_{A_i}) = \sum_{i=1}^{n} P(A_i)(1 - P(A_i)) = np(1-p)$.

The mgf expression also follows immediately from this representation using the indicator variables I_{A_i}, as each indicator variable has the mgf $(pe^t + 1 - p)$, and they are independent.

Parts (c) and (d) follow on differentiating $\psi(t)$ three and four times, respectively, thus obtaining $E(X^3)$ and $E(X^4)$ as the third and fourth derivatives of $\psi(t)$ at zero, and finally plugging them into the binomial expansion $E[(X-\mu)^3] = E(X^3) - 3\mu E(X^2) + 2\mu^3$ and a similar expansion for $E[(X-\mu)^4]$. This tedious algebra is omitted.

Corollary 6.2. *Let $\beta = \beta(n, p)$ be the* skewness *and $\gamma = \gamma(n, p)$ be the* kurtosis *of X. Then $\beta, \gamma \to 0$ for any p as $n \to \infty$.*

The corollary follows by directly using the definitions $\beta = \frac{E[(X-\mu)^3]}{\sigma^3}$ and $\gamma = \frac{E[(X-\mu)^4]}{\sigma^4} - 3$ and plugging in the formulas from the theorem above.

Thus, whatever $p, 0 < p < 1$, the binomial distribution becomes nearly symmetric and *normal-like* as n gets large.

Mean absolute deviations, whenever they can be found in closed form, are appealing measures of variability. Remarkably, an exact formula for the mean absolute deviation of a general binomial distribution exists and is quite classic. Several different versions of it have been derived by various authors, including Poincaré (1896) and Feller (1968); Diaconis and Zabell (1991) is an authoritative exposition of the problem. Another interesting question is, which value in a general binomial distribution has the largest probability? That is, what is the *mode* of the distribution? The next result summarizes the answers to these questions.

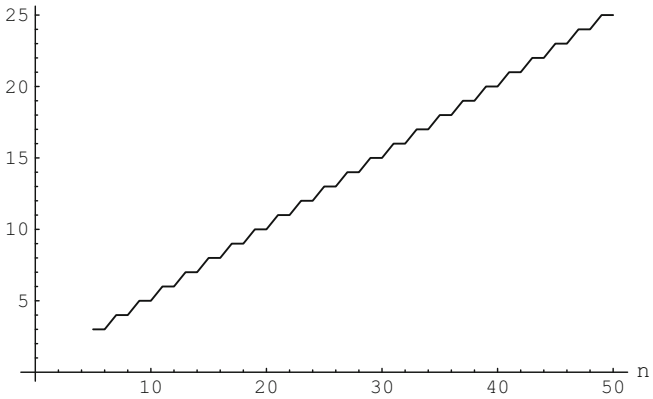


Fig. 6.1 The oscillatory nature of the mode of $Bin(n, .5)$ distribution

Theorem 6.3 (∗ Mean Absolute Deviation and Mode). *Let $X \sim Bin(n, p)$. Let ν denote the smallest integer $> np$ and let $m = \lfloor np + p \rfloor$. Then,*

(a) $E|X - np| = 2\nu(1-p)P(X = \nu)$.
(b) *The mode of X equals m. In particular, if np is an integer, then the mode is exactly np; if np is not an integer, then the mode is one of the two integers just below and just above np.*

Proof. Suppose first that $m \geq 1$. Part (b) can be proved by looking at the ratio $\frac{P(X=k+1)}{P(X=k)}$ and on observing that this ratio is ≥ 1 for $k \leq m-1$. If n, p are such that m is zero, then $P(X = k)$ can be directly verified to be maximized at $k = 0$. This is a standard technique for finding the maximum of a *unimodal function* of an integer argument. Part (a) requires nontrivial calculations; see Diaconis and Zabell (1991).

Remark 6.1. It follows from this theorem that the mode of a binomial distribution need not be the integer closest to the mean np. The modal value maintains a gentle oscillatory nature as n increases and p is held fixed; a plot when $p = .5$ is given in Figure 6.1 to illustrate this oscillation.

6.4 Geometric and Negative Binomial Distributions

Again, it is helpful to begin with some examples.

Example 6.8 (Family Planning). In some economically disadvantaged countries, a male child is considered necessary to help with physical work and family finances. Suppose a couple will have children until they have had two boys. Let X be the number of children they will have. Then, $X \sim NB(r, p)$, with $r = 2, p = .5$ (assumed). Thus, X has the pmf

$$P(X = x) = (x-1)(.5)^x, x = 2, 3, \ldots.$$

For example, $P(\text{The couple will have at least one girl}) = P(X \geq 3) = 1 - P(X = 2) = 1 - .25 = .75$. The probabilities of some values of X are given in the following table:

x	2	3	4	5	6	7	8
$P(X = x)$	.25	.25	.1875	.125	.0781	.0469	.0273

For example, $P(X \geq 6) = 1 - P(X \leq 5) = 1 - (.25 + .25 + .1875 + .125) \approx .19$. It is surprising that nearly 19% of such couples will have six or more children!

Example 6.9 (Meeting Someone with the Same Birthday). Suppose you were born on October 15. How many different people do you have to meet before you find someone who was also born on October 15? Under the usual conditions of equally likely birthdays and independence of the birthdays of all people you will meet, the number of people X you have to meet to find the first person with the same birthday as yours is geometric; i.e., $X \sim Geo(p)$ with $p = \frac{1}{365}$. The pmf of X is $P(X = x) = p(1-p)^{x-1}$. Thus, for any given k,

$$P(X > k) = \sum_{x=k+1}^{\infty} p(1-p)^{x-1} = p \sum_{x=k}^{\infty} (1-p)^x = (1-p)^k.$$

For example, the chance that you will have to meet more than 1000 people to find someone with the same birthday as yours is $(364/365)^{1000} = .064$. But, of course, you will usually not ask people you meet what their birthday is, so it may be hard to verify experimentally that you should not need to meet 1000 people.

Example 6.10. Suppose a door-to-door salesman makes an actual sale in 25% of the visits he makes. He is supposed to make at least two sales per day. How many visits should he plan on making to be 90% sure of making at least two sales?

Let X be the visit at which the second sale is made. Then, $X \sim NB(r, p)$ with $r = 2, p = .25$. Therefore, X has the pmf $P(X{=}x){=}(x-1)(.25)^2(.75)^{x-2}, x = 2, 3, \ldots$. Summing, for any given k, $P(X > k){=}\sum_{x=k+1}^{\infty}(x-1)(.25)^2(.75)^{x-2}$ $={\frac{k+3}{3}}(3/4)^k$ (try to derive this). We want $\frac{k+3}{3}(3/4)^k \leq .1$. By computing this directly, we find that $P(X > 15) < 1$ but $P(X > 14) > .1$. So, the salesman should plan on making 15 visits.

Example 6.11 (Lack of Memory of Geometric Distribution). Let $X \sim Geo(p)$, and suppose m and n are given positive integers. Then, X has the interesting property

$$P(X > m + n | X > n) = P(X > m).$$

That is, suppose you are waiting for some event to happen for the first time. You have tried, say, 20 times, and you still have not succeeded. You may feel that it is due anytime now. The lack of memory property would say that $P(X > 30 | X > 20) = P(X > 10)$. That is, the chance that it will take another ten tries is the same as what it would be if you had just started, and *forget* that you have already been patient for a long time and have tried hard for a success.

The proof is simple. Indeed,

$$P(X > m+n|X > n) = \frac{P(X > m+n)}{P(X > n)} = \frac{\sum_{x>m+n} p(1-p)^{x-1}}{\sum_{x>n} p(1-p)^{x-1}}$$
$$= \frac{(1-p)^{m+n}}{(1-p)^n} = (1-p)^m = P(X > m).$$

We now give some important formulas for the geometric and negative binomial distributions.

Theorem 6.4.
(a) *Let* $X \sim Geo(p)$. *Let* $q = 1 - p$. *Then,*

$$E(X) = \frac{1}{p}; \text{Var}(X) = \frac{q}{p^2}.$$

(b) *Let* $X \sim NB(r, p), r \geq 1$. *Then,*

$$E(X) = \frac{r}{p}; \text{Var}(X) = \frac{rq}{p^2}.$$

Furthermore, the mgf and the (probability) generating function of X *equal*

$$\psi(t) = \left(\frac{pe^t}{1-qe^t}\right)^r, t < \log\left(\frac{1}{q}\right);$$
$$G(s) = \left(\frac{ps}{1-qs}\right)^r, s < \frac{1}{q}.$$

Proof. The formula for the mean and the variance of the geometric distribution follows by simply performing the sums. For example,

$$E(X) = \sum_{x\geq 1} xpq^{x-1} = p\sum_{x\geq 1} xq^{x-1} = p \times \frac{1}{(1-q)^2} = p \times \frac{1}{p^2} = \frac{1}{p}.$$

To find the variance, find the second moment by summing $\sum_{x\geq 1} x^2 pq^{x-1}$, and then plug into the variance formula $\text{Var}(X) = E(X^2) - [E(X)]^2$. It would be easier to find the second moment by first finding the factorial moment $E[X(X-1)]$ and then use the fact that $E(X^2) = E[X(X-1)] + E(X)$. We omit the algebra.

The mean and the variance for the general negative binomial follow from the geometric case on using the very useful representation

$$X = X_1 + X_2 + \cdots + X_r,$$

where X_i is the geometric random random variable measuring the number of additional trials needed to obtain the ith success after the $(i-1)$th success has been obtained. Thus, the X_i are independent, and each is distributed as $Geo(p)$. So, their

variance can be obtained by summing the variances of $X_1, X_2, \ldots, X_r$, which gives $\text{Var}(X) = \sum_{i=1}^{r} \frac{q}{p^2} = \frac{rq}{p^2}$, and the expectation of course also adds up, to give $E(X) = \frac{r}{p}$.

The formula for the mgf of the geometric distribution is immediately obtained by summing $\sum_{x\geq 1} e^{tx} pq^{x-1} = \frac{p}{q} \sum_{x\geq 1} (qe^t)^x = \frac{p}{q} \frac{qe^t}{1-qe^t} = \frac{pe^t}{1-qe^t}$. The formula for the negative binomial distribution follows from this formula by representing X as $X_1 + X_2 + \cdots + X_r$ as above. Finally, the (probability) generating function is derived by following exactly the same steps.

6.5 Hypergeometric Distribution

As we mentioned, the hypergeometric distribution arises when sampling without replacement from a finite population consisting of elements of just two types. Here are some illustrative examples.

Example 6.12 (Gender Discrimination). From a pool of five male and five female applicants, three were selected and all three happened to be men. Is there a priori evidence of gender discrimination?

If we let X be the number of female applicants selected, then $X \sim Hypergeo(n, D, N)$, with $n = 3, D = 5, N = 10$. Therefore,

$$P(X = 0) = \binom{D}{0}\binom{N-D}{n} \Big/ \binom{N}{n} = \binom{5}{3} \Big/ \binom{10}{3} = \frac{1}{12}.$$

So, if selection was done at random, which should be the policy if all applicants are equally qualified, then selecting no women is a low-probability event. There might be some a priori evidence of gender discrimination.

Example 6.13 (Bridge). Suppose North and South together received no aces at all in three consecutive bridge plays. Is there a reason to suspect that the distribution of cards is not being done at random?

Let X be the number of aces in the hands of North and South combined in one play. Then,

$$P(X = 0) = \frac{\binom{48}{13}\binom{35}{13}}{\binom{52}{13}\binom{39}{13}} = \frac{46}{833} = .0552.$$

Therefore, the probability of North and South not receiving any aces for three consecutive plays is $(.0552)^3 = .00017$, which is very small. Either an extremely rare event has happened or the distribution of cards has not been random. Statisticians call this sort of calculation a *p-value calculation* and use it to assess doubt about some proposition, in this case randomness of the distribution of the cards.

Example 6.14 (A Classic Example: Capture-Recapture). An ingenious use of the hypergeometric distribution in estimating the size of a finite population is the *capture-recapture* method. It was originally used for estimating the total number of fish in a body of water, such as a pond. Let N be the number of fish in the pond. In this method, a certain number of fish, say D of them are initially captured and tagged with a safe mark or identification device and then returned to the water. Then, a second sample of n fish is recaptured from the water. Assuming that the fish population has not changed in any way in the intervening time and that the initially captured fish remixed with the fish population homogeneously, the number of fish in the second sample, say X, that bear the mark is a hypergeometric random variable, namely $X \sim Hypergeo(n, D, N)$. We will shortly see that the expected value of a hypergeometric random variable is $n\frac{D}{N}$. If we set as a formalism $X = n\frac{D}{N}$ and *solve* for N, we get $N = \frac{nD}{X}$. This is an estimate of the total number of fish in the pond. Although the idea is extremely original, this estimate can run into various kinds of difficulties if, for example, the first catch of fish clusters around after being returned, hides, or if the fish population has changed between the two catches due to death or birth, and of course if X turns out to be zero. Modifications of this estimate (known as the *Petersen estimate*) are widely used in wildlife estimation, taking a census, and by the government for estimating tax fraud and the number of people afflicted with some infection.

The mean and variance of a hypergeometric distribution are given in the next result.

Theorem 6.5. *Let $X \sim Hypergeo(n, D, N)$ and let $p = \frac{D}{N}$. Then,*

$$E(X) = np; \text{Var}(X) = np(1-p)\left(\frac{N-n}{N-1}\right).$$

We will not prove this result, as it involves the standard indicator variable argument we are familiar with and some routine algebra. Two points worth mentioning are that although sampling is without replacement in the hypergeometric case, so the Bernoulli trials are not independent, the same formula for the mean as in the binomial case holds. But the variance is smaller *than in the binomial case because the extra factor $\frac{N-n}{N-1} < 1$. Sampling without replacement makes the composition of the sample more like the composition of the entire population, and this reduces the variance around the population mean. The factor $\frac{N-n}{N-1}$ is often called the* finite population correction factor.

Problems that should truly be modeled as hypergeometric distribution problems are often analyzed as if they were binomial distribution problems. That is, the fact that samples have been taken without replacement is ignored, and one pretends that the successive draws are independent. When does it not matter that the dependence between the trials is ignored? Intuitively, we would think that if the population size N was large and neither D nor $N - D$ was small, the trials would act like they are independent. The following theorem justifies this intuition.

Theorem 6.6 (Convergence of Hypergeometric to Binomial). *Let $X = X_N \sim Hypergeo(n, D, N)$, where $D = D_N$ and N are such that $N \to \infty, \frac{D}{N} \to p$, $0 < p < 1$. Then, for any fixed n and for any fixed x,*

$$P(X = x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \to \binom{n}{x} p^x (1-p)^{n-x}$$

as $N \to \infty$.

This is proved by using Stirling's approximation (which says that as $k \to \infty$, $k! \sim e^{-k} k^{k+1/2}\sqrt{2\pi}$) for each factorial term in $P(X = x)$ and then doing some algebra.

6.6 Poisson Distribution

As mentioned before, Poisson distributions arise as counts of events in fixed periods of time, fixed amounts of area or space, and as limits of binomial distributions for large n and small p. The first thing to note, before we can work out examples, is that the single parameter λ of a Poisson distribution is its mean; quite remarkably, λ is also the variance of the distribution. We will write $X \sim Poi(\lambda)$ to denote a Poisson random variable. The distribution was introduced by Siméon Poisson (1838).

Theorem 6.7. *Let $X \sim Poi(\lambda)$. Then,*

(a) $E(X) = Var(X) = \lambda$.
(b) $E(X - \lambda)^3 = \lambda; E(X - \lambda)^4 = 3\lambda^2 + \lambda$.
(c) *The mgf of X equals*

$$\psi(t) = e^{\lambda(e^t - 1)}.$$

Proof. Although parts (a) and (b) can be proved directly, it is most efficient to derive them from the mgf. So, we first prove part (c):

$$\psi(t) = E[e^{tX}] = \sum_{x=0}^{\infty} e^{tx} P(X = x) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} [\lambda e^t]^x / x! = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}.$$

Therefore,

$$\psi'(t) = e^{\lambda(e^t-1)}\lambda e^t,$$
$$\psi''(t) = \lambda e^t(1 + \lambda e^t)e^{\lambda(e^t-1)},$$
$$\psi^{(3)}(t) = \lambda e^t(1 + 3\lambda e^t + \lambda^2 e^{2t})e^{\lambda(e^t-1)},$$
$$\psi^{(4)}(t) = \lambda e^t(1 + 7\lambda e^t + 6\lambda^2 e^{2t} + \lambda^3 e^{3t})e^{\lambda(e^t-1)}.$$

From these, by using the fact that $E(X^k) = \psi^{(k)}(0)$, we get

$$E(X) = \lambda; E(X^2) = \lambda + \lambda^2; E(X^3) = \lambda(1 + 3\lambda + \lambda^2);$$
$$E(X^4) = \lambda(1 + 7\lambda + 6\lambda^2 + \lambda^3).$$

The formulas in parts (a) and (b) now follow by simply plugging in the expressions given above for the first four moments of X.

Corollary 6.3. *The skewness and kurtosis of X equal*

$$\beta = \frac{1}{\sqrt{\lambda}}; \gamma = \frac{1}{\lambda}.$$

The corollary follows immediately by using the definitions of skewness and kurtosis.

Let us now see some illustrative examples. The appendix gives a table of Poisson probabilities for λ between .5 and 5. These may be used instead of manually calculating the probabilities whenever the required probability can be obtained from the table given in the appendix.

Example 6.15 (Events over Time). April receives three phone calls at her home on average per day. On what percentage of days does she receive no phone calls? More than five phone calls?

Because the number of calls received in a 24 hour period counts the occurrences of an event in a fixed time period, we model $X =$ number of calls received by April on one day as a Poisson random variable with mean 3. Then,

$$P(X = 0) = e^{-3} = .0498; P(X > 5) = 1 - P(X \leq 5) = 1 - \sum_{x=0}^{5} e^{-3}3^x/x!$$
$$= 1 - .9161 = .0839.$$

Thus, she receives no calls on 4.98% of the days and more than five calls on 8.39% of the days. *It is important to understand that X has only been modeled as a Poisson random variable, and other models could also be reasonable.*

Example 6.16. Lengths of an electronic tape contain, on average, one defect per 100 ft. If we need a tape of 50 ft., what is the probability that it will be defect-free?

Let X denote the number of defects per 50 ft. of this tape. We can think of lengths of the tape as a window of *time*, although not in a literal sense. If we assume that the defective rate is *homogeneous* over the length of the tape, then we can model X as $X \sim Poi(.5)$. That is, if 100 ft. contain one defect on average, then 50 ft. of tape should contain half a defect on average. This can be made rigorous by using the concept of a *homogeneous Poisson process*.

Therefore,

$$P(X = 0) = e^{-.5} = .6065.$$

Example 6.17 (Events over an Area). Suppose a 14 inch circular pizza has been baked with 20 pieces of barbecued chicken. At a party, you were served a $4 \times 4 \times 2$ (in inches) triangular slice. What is the probability that you got at least one piece of chicken?

The area of a circle of radius 7 is $\pi \times 7^2 = 153.94$. The area of a triangular slice of lengths 4, 4, and 2 inches on a side is $\sqrt{s(s-a)(s-b)(s-c)} = \sqrt{5 \times 1 \times 1 \times 3} = \sqrt{15} = 3.87$, where a, b, c are the lengths of the three sides and $s = (a+b+c)/2$. Therefore, we model X, the number of pieces of chicken in the triangular slice, as $X \sim Poi(\lambda)$, where $\lambda = 20 \times 3.87/153.94 = .503$. Using the Poisson pmf,

$$P(X \geq 1) = 1 - e^{-.503} = .395.$$

Example 6.18 (A Hierarchical Model with a Poisson Base). Suppose a chick lays a $Poi(\lambda)$ number of eggs in some specified period of time, say a month. Each egg has a probability p of actually developing. We want to find the distribution of the number of eggs that actually develop during that period of time.

Let $X \sim Poi(\lambda)$ denote the number of eggs the chick lays and Y the number of eggs that develop. For example,

$$\begin{aligned} P(Y=0) &= \sum_{x=0}^{\infty} P(Y=0|X=x)P(X=x) = \sum_{x=0}^{\infty}(1-p)^x \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda(1-p))^x}{x!} = e^{-\lambda}e^{\lambda(1-p)} = e^{-p\lambda}. \end{aligned}$$

In general,

$$\begin{aligned} P(Y=y) &= \sum_{x=y}^{\infty}\binom{x}{y}p^y(1-p)^{x-y}\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \frac{(p/(1-p))^y}{y!}e^{-\lambda}\sum_{x=y}^{\infty}\frac{1}{(x-y)!}(1-p)^x\lambda^x \\ &= \frac{(p/(1-p))^y}{y!}e^{-\lambda}(\lambda(1-p))^y\sum_{n=0}^{\infty}\frac{(\lambda(1-p))^n}{n!} \\ &= \frac{(\lambda p)^y}{y!}e^{-\lambda}e^{\lambda(1-p)} = \frac{e^{-\lambda p}(\lambda p)^y}{y!}, \end{aligned}$$

on writing $n = x - y$ in the summation, so we recognize by inspection that $Y \sim Poi(\lambda p)$. What is interesting here is that the distribution of Y still remains Poisson under assumptions that seem to be very realistic physically.

Example 6.19 (Meteor Showers). Between the months of May and October, you can see a shooting star at the rate of about one per 20 minutes. If you sit on your patio

for one hour each evening, how many days would it be before you see ten or more shooting stars on the same day?

This example combines the Poisson and the geometric distributions in an interesting way. Let p be the probability of seeing ten or more shooting stars in one day. If we let N denote the number of shooting stars observed in one day and model N as $N \sim Poi(\lambda)$, with $\lambda = 3$ (since one hour is equal to three 20 minute intervals), then

$$p = P(N \geq 10) = \sum_{x=10}^{\infty} \frac{e^{-3}3^x}{x!} = .0011.$$

Now, if we let X denote the number of days that you have to watch the sky until you see this shower of ten or more shooting stars, then $X \sim Geo(p)$ and therefore $E(X) = \frac{1}{p} = 909.1$, which is about 30 months. You are observing for six months each year because there are six months between May and October (inclusive). So, you can expect that if you observe for about five years, you will see a shower of ten or more shooting stars on some evening.

Example 6.20 (Poisson Forest). It is a common assumption in forestry and ecology that the number of plants in a part of a forest is distributed according to a Poisson distribution with mean proportional to the area of the part of the forest.

Suppose on average there are ten trees per 100 square ft. in a forest. An entomologist is interested in estimating an insect population in a forest of size 10,000 square ft. The insects are found in the trees, and it is believed that there are 100 of them per tree. The entomologist will cover a 900 square ft. area and count the insects on all trees in that area. What are the chances that the entomologist will discover more than 9200 insects in this area?

Suppose X is the number of trees in the 900 square ft. area the entomologist covers, and let Y be the number of insects the entomologist discovers. We assume that $X \sim Poi(\lambda)$, with $\lambda = 90$. Then, because there are 100 insects per tree,

$$P(Y > 9200) = P(X > 92) = \sum_{x=93}^{\infty} \frac{e^{-90}(90)^x}{x!} \approx .3898.$$

The .3898 value was found by direct summation on a computer. A more realistic model will assume the number of insects per tree is a random variable rather than being constantly equal to 100. However, finding an answer to the question would then be much harder.

Example 6.21 (Gamma-Ray Bursts). Gamma-ray bursts are thought to be the most intense electromagnetic events observed in the sky, and they typically last a few seconds. While they are on, their intense brightness covers up any other gamma-ray source in the sky. They occur at the rate of about one episode per day. It was initially thought that they were events within the Milky Way galaxy, but most astronomers now believe that is not true or not entirely true.

The 2000th gamma-ray burst since 1991 was detected at the end of 1997 at NASA's Compton Gamma Ray Observatory. Are these data compatible with a model of a Poisson-distributed number of bursts with a rate of one per day?

Using a model of homogeneously distributed events, the number of bursts in a seven-year period is $Poi(\lambda)$ with $\lambda = 7 \times 365 \times 1 = 2555$. The observed number of bursts is 2000, less than the expected number of bursts. But is it so much less that the postulated model is in question? To assess this, we calculate $P(X \leq 2000)$, the probability that we could observe an observation as deviant from the expected one as we did just by chance. Statisticians call such a deviation probability a *p-value*. The p-value then equals

$$P(X \leq 2000) = \sum_{x=0}^{2000} \frac{e^{-2555}(2555)^x}{x!}.$$

Due to the large values of λ and the range of the summation, directly summing this is not recommended. But the sum can be approximated by using various other indirect means, including a theorem known as the *central limit theorem*, which we will later discuss in detail. The approximate p-value can be seen to be extremely small, virtually zero. So, the chance of such a deviant observation, if the Poisson model at the rate of one burst per day was correct, is very small. One would doubt the model in such a case. The bursts may not occur at a homogeneous rate of one per day.

6.6.1 Mean Absolute Deviation and the Mode

Similar to the binomial case, a closed-form formula is available for the mean absolute deviation $E[|X - \lambda|]$ of a Poisson distribution; we can also characterize the mode; i.e., the value with the largest probability. Again, see Diaconis and Zabell (1991) for these results.

Theorem 6.8 (Mean Absolute Deviation and Mode). *Let $X \sim Poi(\lambda)$. Then:*

(a) $E[|X - \lambda|] = 2\lambda P(X = \lfloor \lambda \rfloor)$.
(b) *A Poisson distribution is unimodal and* $P(X = k) \leq P(X = \lfloor \lambda \rfloor)\ \forall k \geq 0$.

Proof. Part (a) requires nontrivial calculations; see Diaconis and Zabell (1991). Part (b), however, is easy to prove. Consider the ratio

$$\frac{P(X = k+1)}{P(X = k)} = \frac{\lambda}{k+1},$$

and note that this is ≥ 1 if and only if $k + 1 \leq \lfloor \lambda \rfloor$, which proves that $\lfloor \lambda \rfloor$ is always a mode. If λ is an integer, then λ and $\lambda - 1$ will both be modes; that is, there would be two modes. If λ is not an integer, then $\lfloor \lambda \rfloor$ is the unique mode.

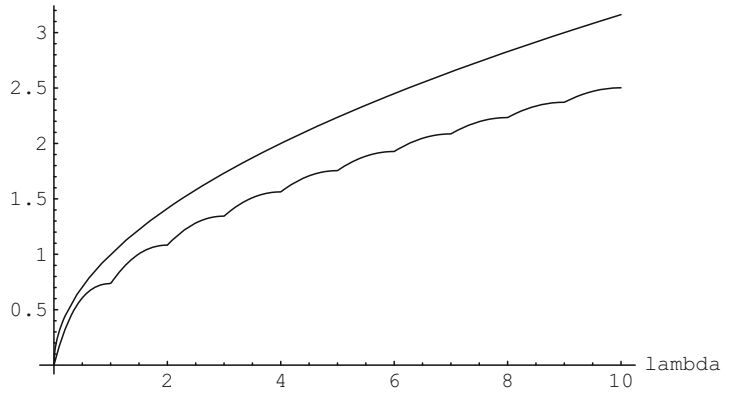


Fig. 6.2 Mean absolute deviation and standard deviation of a poisson distribution

We recall that the mean absolute deviation of a random variable is always smaller than (or equal to) the standard deviation of the random variable. It is interesting to see a plot of these two as a function of λ (see Figure 6.2). The mean absolute deviation is continuous, but not differentiable, and has a periodic component in it.

6.7 Poisson Approximation to Binomial

A binomial random variable is the sum of n indicator variables. When the expectation of these indicator variables, namely p, is small, and the number of summands n is large, the Poisson distribution provides a good approximation to the binomial. The Poisson distribution can also sometimes serve as a good approximation when the indicators are independent but have different expectations p_i, or when the indicator variables have some weak dependence. We will start with the Poisson approximation to the binomial when n is large and p is small.

Theorem 6.9. *Let $X_n \sim Bin(n, p_n), n \geq 1$. Suppose $np_n \to \lambda, 0 < \lambda < \infty$, as $n \to \infty$. Let $Y \sim Poi(\lambda)$. Then, for any given $k, 0 \leq k < \infty$,*

$$P(X_n = k) \to P(Y = k)$$

as $n \to \infty$.

Proof. For ease of explanation, let us first consider the case $k = 0$. We have

$$P(X_n = 0) = (1-p)^n = \left(1 - \frac{np}{n}\right)^n \sim \left(1 - \frac{\lambda}{n}\right)^n \sim e^{-\lambda}.$$

Note that we did not actually prove the claimed fact that $(1-\frac{np}{n})^n \sim (1-\frac{\lambda}{n})^n$, but it is true and is not hard to prove.

Now consider $k=1$. We have

$$P(X_n=1) = np(1-p)^{n-1} = (np)(1-p)^n \frac{1}{1-p} \sim \lambda(e^{-\lambda})(1) = \lambda e^{-\lambda}.$$

The same technique works for any k. Indeed, for a general k,

$$\begin{aligned} P(X_n=k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{1}{k!}[n(n-1)\cdots(n-k+1)]p^k(1-p)^n \left[\frac{1}{(1-p)^k}\right] \\ &= \frac{1}{k!}n^k \left[1\frac{n-1}{n}\cdots\frac{n-k+1}{n}\right] p^k(1-p)^n \left[\frac{1}{(1-p)^k}\right] \\ &= \frac{1}{k!}(np)^k \left[1\frac{n-1}{n}\cdots\frac{n-k+1}{n}\right] (1-p)^n \left[\frac{1}{(1-p)^k}\right] \\ &\sim \frac{1}{k!}(\lambda)^k[1]e^{-\lambda}[1] = \frac{e^{-\lambda}\lambda^k}{k!}, \end{aligned}$$

which is what the theorem says.

In fact, the convergence is not just pointwise for each fixed k but is *uniform* in k. This will follow from the following more general theorem, which we state for reference (see Le Cam, 1960; Barbour and Hall, 1984; Steele, 1994)

Theorem 6.10 (Le Cam, Barbour and Hall, Steele). *Let $X_n = B_1+B_2+\cdots+B_n$, where B_i are independent Bernoulli variables with parameters $p_i = p_{i,n}$. Let $Y_n \sim Poi(\lambda)$, where $\lambda = \lambda_n = \sum_{i=1}^n p_i$. Then,*

$$\sum_{k=0}^{\infty} |P(X_n=k) - P(Y_n=k)| \le 2\frac{1-e^{-\lambda}}{\lambda}\sum_{i=1}^n p_i^2.$$

Here are some more examples of the Poisson approximation to the binomial.

Example 6.22 (Lotteries). Consider a weekly lottery in which three numbers out of 25 are selected at random and a person holding exactly those three numbers is the winner of the lottery. Suppose the person plays for n weeks, for large n. What is the probability that he will win the lottery at least once? At least twice?

Let X be the number of weeks that the player wins. Then, assuming the weekly lotteries are independent, $X \sim Bin(n,p)$, where $p = 1/\binom{25}{3} = \frac{1}{2300} = .00043$. Since p is small and n is supposed to be large, $X \overset{approx.}{\sim} Poi(\lambda), \lambda = np = .00043n$. Therefore,

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - e^{-.00043n}$$

and

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) \approx 1 - e^{-.00043n} - .00043n e^{-.00043n}$$
$$= 1 - (1 + .00043n)e^{-.00043n}.$$

We can compute these for various n. If the player plays for five years,

$$1 - e^{-.00043n} = 1 - e^{-.00043 \times 5 \times 52} = .106$$

and

$$1 - (1 + .00043n)e^{-.00043n} = .006.$$

If he plays for ten years,

$$1 - e^{-.00043n} = 1 - e^{-.00043 \times 10 \times 52} = .200$$

and

$$1 - (1 + .00043n)e^{-.00043n} = .022.$$

We can see that the chances of any luck are at best moderate even after prolonged tries.

Example 6.23 (An Insurance Example). Suppose 5000 clients are each insured for one million dollars against fire damage in a coastal property. Each residence has a 1 in 10,000 chance of being damaged by fire in a 12 month period. How likely is it that the insurance company has to pay out as much as 3 million dollars in fire damage claims in one year? Four million dollars?

If X is the number of claims made during a year, then $X \sim Bin(n, p)$ with $n = 5000$ and $p = 1/10,000$. We assume that no one makes more than one claim and that the clients are independent. Then we can approximate the distribution of X by $Poi(np) = Poi(.5)$. We need

$$P(X \geq 3) = 1 - P(X \leq 2) \approx 1 - (1 + .5 + .5^2/2)e^{-.5} = .014$$

and

$$P(X \geq 4) = 1 - P(X \leq 3) \approx 1 - (1 + .5 + .5^2/2 + .5^3/6)e^{-.5} = .002.$$

These two calculations are done above by using the Poisson approximation, namely $\frac{e^{-.5}.5^k}{k!}$, for $P(X = k)$. The insurance company is quite safe being prepared for 3 million dollars in payout and very safe being prepared for 4 million dollars.

6.8 ∗ Miscellaneous Poisson Approximations

A binomial random variable is the sum of independent and identically distributed Bernoulli variables. Poisson approximations are also often accurate when the individual Bernoulli variables are independent but have small and different parameters p_i or when the Bernoulli variables have a weak dependence. A rule of thumb is that if the individual p_i's are small and their sum is moderate, then a $Poi(\sum p_i)$ approximation should be accurate. There are many rigorous theorems in this direction. There are the first-generation Poisson approximation theorems and the more modern Poisson approximation theorems, that go by the name of the *Stein-Chen method*. The Stein-Chen method is now regarded as the principal tool for approximating the distribution of sums of weakly dependent Bernoulli variables, with associated bounds on the error of the approximation. The two original papers are Stein (1972) and Chen (1975) . More recent sources with modern applications in a wide variety of fields are Barbour et al. (1992) and Diaconis and Holmes (2004).

We will first work out a formal Poisson approximation in some examples below.

Example 6.24 (Poisson Approximation in the Birthday Problem). In the birthday problem, n unrelated people gather around and we want to know if there is at least one pair of individuals with the same birthday. Defining $I_{i,j}$ as the indicator of the event that individuals i and j have the same birthday, we have

$$X = \text{number of different pairs of people who share a common birthday}$$
$$= \sum_{1 \le i < j \le n} I_{i,j}.$$

Each $I_{i,j} \sim Ber(p)$, where $p = 1/365$. Note, however, that the $I_{i,j}$ are definitely not independent. Now, the expected value of X is $\lambda = \binom{n}{2}/365$. This is moderate ($> .5$) if $n \ge 20$. So, a Poisson approximation may be accurate when n is about 20 or more.

If we use a Poisson approximation when $n = 23$, we get

$$P(X > 0) \approx 1 - e^{-\binom{23}{2}/365} = 1 - e^{-.693151} = .500002,$$

which is almost exactly equal to the true value of the probability that there will be a pair of people with the same birthday in a group of 23 people; this was previously discussed in Chapter 2.

Example 6.25 (Three People with the Same Birthday). Consider again a group of n unrelated people, and ask what the chances are that we can find three people in the group with the same birthday. We proceed as in the preceding example. Define $I_{i,j,k}$ as the indicator of the event that individuals i, j, k have the same birthday. Then, $I_{i,j,k} \sim Ber(p)$, $p = 1/(365)^2$. Let

$$X = \sum_{1 \le i < j < k \le n} I_{i,j,k}$$

$$= \text{number of different triplets of people who share a common birthday.}$$

The expected value of X is $\lambda = \binom{n}{3}/365^2$. We want to approximate $P(X \geq 1)$. If we use the Poisson approximation $X \sim Poi(\binom{n}{3}/365^2)$, we get with $n = 84$

$$P(X \geq 1) \approx 1 - e^{-\binom{84}{3}/365^2} = 1 - e^{-.715211} = .5109.$$

In fact, $n = 84$ is truly the first n for which the probability that we can find three people with the same birthday exceeds .5. We again see the effectiveness of the Poisson approximation in approximating sums of dependent Bernoulli variables. Note how much *harder it is to find three people with the same birthday than it was to find two*!

A reasonably simple first-generation theorem on the validity of the Poisson approximation for suitable sums of (not necessarily independent) Bernoulli variables can be described by using the so-called *binomial moments* of the sum. We will first define the term binomial moment.

Definition 6.2. Let X be a nonnegative integer-valued random variable with a finite jth moment for a given $j \geq 1$. The jth binomial moment of X is defined as $M_j = E[\binom{X}{j}] = \sum_{x=j}^{\infty} \binom{x}{j} P(X = x)$.

Remark. Note that the binomial moments and the factorial moments are related as $M_j = \frac{E[X(X-1)\cdots(X-j+1)]}{j!}$; thus the jth binomial moment is finite if and only if the jth factorial moment is finite, which is true if and only if the jth moment is finite.

We give an example.

Example 6.26 (Factorial Moments of Poisson). Let $X \sim Poi(\lambda)$. Let $n \geq 1$. Then,

$$\begin{aligned}
E[X(X-1)\cdots(X-n+1)] &= \sum_{x=0}^{\infty} x(x-1)\cdots(x-n+1)\frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=n}^{\infty} x(x-1)\cdots(x-n+1)\frac{e^{-\lambda}\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=n}^{\infty}\frac{\lambda^x}{(x-n)!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^{x+n}}{x!} = e^{-\lambda}\lambda^n\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\lambda^n \times e^{\lambda} = \lambda^n,
\end{aligned}$$

a remarkably pretty result.

In some problems, typically of a combinatorial nature, careful counting lets one use the binomial moments and establish the validity of a Poisson approximation.

Here is a theorem that shows how to do it; see Galambos and Simonelli (1996) for a proof of it.

Theorem 6.11 (Basic General Poisson Approximation Theorem). *Let $X_n = B_1 + B_2 + \cdots + B_n$, where B_i are Bernoulli random variables. Let $M_k = M_{k,n}$ be the kth binomial moment of X_n. If there exists $0 < \lambda < \infty$ such that, for every fixed k, $M_k \to \frac{\lambda^k}{k!}$ as $n \to \infty$, then $P(X_n = j) \to \frac{e^{-\lambda}\lambda^j}{j!}$ for any j as $n \to \infty$.*

Here is an application of this theorem.

Example 6.27 (The Committee Problem). From n people, $N = N(n)$ committees are formed, each committee of a fixed size m. We let $N, n \to \infty$, holding m fixed. The Bernoulli variable $B_i = B_{i,n}$ is the indicator of the event that the ith person is not included in any committee. The purpose of this example is to derive a Poisson approximation for X, the total number of people who are not included in any committee.

Under the usual assumptions of independence and also the assumption of random selection, the binomial moment M_k can be shown to be

$$M_k = \binom{n}{k}\left[\frac{\binom{n-k}{m}}{\binom{n}{m}}\right]^N.$$

Stirling's approximation now shows that $M_k \sim \frac{n^k}{k!}e^{-kN(\frac{m}{n}+O(n^{-2}))}$ as $n \to \infty$. One now sees, on inspection, that if N, n are related as $N = \frac{n\log n}{m} - n\log\lambda + o(n^{-1})$ for some $0 < \lambda < \infty$, then $M_k \to \frac{\lambda^{km}}{k!}$, so from the basic general Poisson approximation theorem above, the number of people who are left out of *all* committees converges to $Poi(\lambda^m)$.

6.9 Benford's Law

Benford's law asserts that if we pick numbers at random from a statistical data set or mathematical tables such as a table of logarithms or a table of physical constants, then the leading digit tends to be 1 with much greater frequency than the 11.1% one would expect if the distribution was just discrete uniform on $\{1, 2, \ldots, 9\}$. The *law* was first asserted by the astronomer Simin Newcomb (1881). However, the distribution has come to be known as the Benford distribution, attributable to a publication by Frank Benford (1938). The distribution was later found to give quite reasonable fits to various kinds of data, such as the first digit in randomly picked home addresses, daily returns of stocks, leading digits in geological variables, baseball statistics, half-lives of radioactive particles, etc.

The Benford distribution is a distribution on $\{1, 2, \ldots, 9\}$ with the pmf

$$p(x) = \frac{\log(x+1) - \log x}{\log 10}, x = 1, 2, \ldots, 9.$$

Note that clearly $p(x) \geq 0$ for any x and $\sum_{x=1}^{9} p(x) = \frac{1}{\log 10}[\log 2 - \log 1 + \log 3 - \log 2 + \cdots + \log 10 - \log 9] = 1$. Therefore, it is a valid pmf. The numerical values of the probabilities are as follows:

x	1	2	3	4	5	6	7	8	9
$p(x)$	.301	.176	.125	.097	.079	.067	.058	.051	.046

The moments of the distribution are easily calculated. In particular, the mean and the variance equal 3.44 and 6.0565. The distribution is right skewed, and the coefficient of skewness equals .796.

Contemporary literature on the Benford distribution includes Hill (1996), Diaconis (1977), and Diaconis and Freedman (1979). Benford's distribution is a very simple distribution from a purely mathematical point of view. Its appeal lies in its ability to give mysteriously good fits to the leading digit for diverse types of empirical data. A recent discovery is that the fits are better when data from apparently unrelated sources are combined. In other words, if a set of variables have their individual distributions and then those distributions are *mixed*, then the leading digit in the mixed distribution would often approximately follow the Benford law.

6.10 Distribution of Sums and Differences

Sums of random variables arise very naturally in practical applications. For example, the revenue over a year is the sum of the monthly revenues; the time taken to finish a test with ten problems is the sum of the times taken to finish the individual problems, etc. Likewise, the difference of two intrinsically similar random variables is also a natural quantity to study; e.g., the number of crimes of some specific kind committed last year and the number committed this year. It is also interesting to look at the absolute difference of similar random variables in addition to the difference itself.

Sometimes we can reasonably assume that the various random variables being added are independent. Thus, the following general question is an important one. Suppose $X_1, X_2, \ldots, X_k$ are k independent random variables and that we know the distributions of the individual X_i. What is the distribution of the sum $X_1 + X_2 + \cdots + X_k$?

In general, this is a very difficult question. Interestingly, if the individual X_i have one of the distinguished distributions we have discussed in this chapter, then their sum is also often a distribution of that same type. For example, sums of independent Poisson random variables would be Poisson also. This *loyalty to types* is a very useful fact, and we present a theorem in this regard below.

Theorem 6.12.

(a) Suppose $X_1, X_2, \ldots, X_k$ are k independent binomial random variables with $X_i \sim Bin(n_i, p)$. Then $X_1 + X_2 + \ldots + X_k \sim Bin(n_1 + n_2 + \cdots + n_k, p)$.

(b) Suppose $X_1, X_2, \ldots, X_k$ are k independent negative binomial random variables with $X_i \sim NB(r_i, p)$. Then $X_1 + X_2 + \cdots + X_k \sim NB(r_1 + r_2 + \cdots + r_k, p)$.

(c) Suppose $X_1, X_2, \ldots, X_k$ are k independent Poisson random variables with $X_i \sim Poi(\lambda_i)$. Then $X_1 + X_2 + \cdots + X_k \sim Poi(\lambda_1 + \lambda_2 + \cdots \lambda_k)$.

Proof. Each of the three parts can be proved by various means. One possibility is to attack the problem directly. Alternatively, the results can also be proved by using generating functions or mgfs. It is useful to see a proof using both methods, so we do this for the Poisson case. The proof for the other two cases is exactly the same and will be omitted.

First, note that it is enough to consider only the case $k = 2$ because then the general case follows by induction. We denote X_1, X_2 as X, Y for notational simplicity. Then,

$$\begin{aligned}
P(X+Y=z) &= \sum_{x=0}^{z} P(X=x, Y=z-x) = \sum_{x=0}^{z} P(X=x)P(Y=z-x) \\
&= \sum_{x=0}^{z} \frac{e^{-\lambda_1}\lambda_1^x}{x!}\frac{e^{-\lambda_2}\lambda_2^{z-x}}{(z-x)!} \\
&= e^{-(\lambda_1+\lambda_2)}\lambda_2^z \times \sum_{x=0}^{z} \frac{(\lambda_1/\lambda_2)^x}{x!(z-x)!} \\
&= e^{-(\lambda_1+\lambda_2)}\frac{\lambda_2^z}{z!}\sum_{x=0}^{z}\binom{z}{x}(\lambda_1/\lambda_2)^x \\
&= e^{-(\lambda_1+\lambda_2)}\frac{\lambda_2^z}{z!}(1+\lambda_1/\lambda_2)^z = e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^z}{z!},
\end{aligned}$$

as was required to prove.

The second method uses the formula for the mgf of a Poisson distribution. Since X and Y are both Poisson and they are independent, the mgf of $X + Y$ is

$$\begin{aligned}
\psi_{X+Y}(t) &= E[e^{t(X+Y)}] = E[e^{tX}]E[e^{tY}] = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} \\
&= e^{(\lambda_1+\lambda_2)(e^t-1)},
\end{aligned}$$

which agrees with the mgf of the $Poi(\lambda_1 + \lambda_2)$ distribution, and therefore, by the distribution-determining property of mgfs, the distribution of $X + Y$ must be $Poi(\lambda_1 + \lambda_2)$.

The calculation used in each of these two methods of proof is useful, and it is important to be familiar with each method.

Example 6.28. Suppose $X \sim Poi(1), Y \sim Poi(5)$, and $Z \sim Poi(10)$, and suppose X, Y, Z are independent. We want to find $P(X + Y + Z \geq 20)$.

By the previous theorem, $X + Y + Z \sim Poi(16)$, and therefore

$$P(X + Y + Z \geq 20) = 1 - P(X + Y + Z \leq 19) = 1 - \sum_{x=0}^{19} \frac{e^{-16}16^x}{x!}$$

$$= 1 - .8122 = .1878.$$

In the absence of the result that $X + Y + Z \sim Poi(16)$, computing this probability would call for enumeration of all the ways that $X + Y + Z$ could be 19 or smaller and adding up those probabilities. Clearly, it would be a much more laborious calculation.

6.10.1 ∗ Distribution of Differences

We now turn to differences of random variables of the same type; e.g., the difference of two independent Poisson random variables. Obviously, it cannot be Poisson, because it can take negative values. Similarly, the difference of two binomial random variables cannot be binomial because it will take negative values. Indeed, the distribution of differences is not nearly as nice or neat as the distribution of sums of variables of the same type. We present the Poisson case below; the binomial case is a chapter exercise.

Theorem 6.13 (Difference of Independent Poissons). *Let X and Y be independent random variables, $X \sim Poi(\lambda_1), Y \sim Poi(\lambda_2)$. Then,*

$$P(X - Y = z) = e^{-(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_z(2\sqrt{\lambda_1\lambda_2}) \text{ if } z \geq 0,$$

where $I_n(x)$ is the modified Bessel function of order n with the power series expansion

$$I_n(x) = \left(\frac{x}{2}\right)^n \sum_{k=0}^{\infty} \frac{x^{2k}}{4^k\, k!(n+k)!}.$$

Proof. The formula for $z \leq 0$ follows from the formula for $z \geq 0$ by switching the roles of X and Y. So we only consider the case $z \geq 0$.

By the independence of X and Y,

$$\begin{aligned}
P(X-Y=z) &= \sum_{y=0}^{\infty} P(X=y+z, Y=y) = \sum_{y=0}^{\infty} P(X=y+z)P(Y=y) \\
&= \sum_{y=0}^{\infty} \frac{e^{-\lambda_1}\lambda_1^{y+z}}{(y+z)!}\frac{e^{-\lambda_2}\lambda_2^{y}}{y!} \\
&= e^{-(\lambda_1+\lambda_2)}\lambda_1^z \sum_{y=0}^{\infty} \frac{(\lambda_1\lambda_2)^y}{(y+z)!y!} \\
&= e^{-(\lambda_1+\lambda_2)}\frac{\lambda_1^z}{(\lambda_1\lambda_2)^{z/2}} I_z(2\sqrt{\lambda_1\lambda_2}) \\
&= e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_z(2\sqrt{\lambda_1\lambda_2}),
\end{aligned}$$

as was claimed.

Of course, directly, $E(X-Y) = \lambda_1 - \lambda_2$ and $\mathrm{Var}(X-Y) = \lambda_1 + \lambda_2$.

6.11 ∗ Discrete Does Not Mean Integer-Valued

Although the named discrete distributions are all on integers, discrete random variables need not be integer-valued. Indeed, according to the definition we have provided in this text, a random variable taking values in any countable set is discrete. One naturally thinks of the rationals as a natural countable set after the set of integers. In particular, the rationals in the unit interval $[0, 1]$ also form a countably infinite set. There are many ways to write reasonable distributions on rationals in the unit interval. We give one example.

Example 6.29 (Distribution on Rationals in the Open Unit Interval). Let X and Y be independent random variables, each distributed as geometric with parameter p, and let $R = \frac{X}{X+Y}$. Then clearly R takes only rational values, and $P(0 < R < 1) = 1$ for any p.

The pmf can be investigated as follows. Let $r = \frac{m}{n}, 0 < m < n$ be a rational in its irreducible form; i.e., m and n have no common factors. Then,

$$\begin{aligned}
P(R=r) &= P(X=mk, X+Y=nk \text{ for some } k \geq 1) \\
&= P(X=mk, Y=(n-m)k \text{ for some } k \geq 1) \\
&= \sum_{k=1}^{\infty} P(X=mk)P(Y=(n-m)k)
\end{aligned}$$

$$= \sum_{k=1}^{\infty} (p(1-p)^{mk-1})(p(1-p)^{(n-m)k-1})$$

$$= \frac{p^2}{(1-p)^2} \sum_{k=1}^{\infty} (1-p)^{nk}$$

$$= \frac{p^2}{(1-p)^2} \frac{(1-p)^n}{1-(1-p)^n}$$

$$= \frac{p^2(1-p)^{n-2}}{1-(1-p)^n},$$

for all $n \geq 2$ and all $m < n$ such that m is *relatively prime to* n. Note that all such m's, for a given n, result in the same pmf value, as is seen in the formula above. In the special case $p = \frac{1}{2}$, the pmf becomes

$$P\left(R = \frac{m}{n}\right) = \frac{1}{2^n - 1}, n \geq 2, (m, n) = 1,$$

where the $(m, n) = 1$ notation means that they are relatively prime. These probabilities would have to add to one. The number of m's relatively prime to a given n is the so-called *Euler totient function* $\phi(n)$. It is interesting that this example therefore shows the number-theoretic identity

$$\sum_{n=2}^{\infty} \frac{\phi(n)}{2^n - 1} = 1.$$

6.12 Synopsis

(a) If $X \sim Bin(n, p)$ and $q = 1 - p$, then

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, 0 \leq x \leq n; E(X) = np; \text{Var}(X) = npq.$$

The mgf of X equals $\psi(t) = (pe^t + q)^n$ at any t. The integer part of $np + p$ is always a mode of X.

(b) If $X \sim Geo(p)$ and $q = 1 - p$, then

$$P(X = x) = pq^{x-1}, x \geq 1; E(X) = \frac{1}{p}; \text{Var}(X) = \frac{q}{p^2}.$$

More generally, if $X \sim NB(r, p), r \geq 1$, then

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, x \geq r; E(X) = \frac{r}{p}; \text{Var}(X) = \frac{rq}{p^2}.$$

The mgf and the generating function of X equal $\psi(t) = \left(\frac{pe^t}{1-qe^t}\right)^r, t < \log\left(\frac{1}{q}\right)$, and $G(s) = \left(\frac{ps}{1-qs}\right)^r, s < \frac{1}{q}$.

(c) If $X \sim Hypergeo(n, D, N)$, $p = \frac{D}{N}$, and $q = 1 - p$, then

$$P(X = x) = \frac{\binom{D}{x}}{\binom{N-D}{n-x}} \binom{N}{n}, n - N + D \leq x \leq D; E(X) = np;$$

$$\text{Var}(X) = npq\left(\frac{N-n}{N-1}\right).$$

(d) The geometric distribution satisfies the lack of memory property

$$P(X > m + n \,|\, X > n) = P(X > m)$$

for any $m, n \geq 1$.

(e) If $X \sim Poi(\lambda)$, then

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x \geq 0; E(X) = \text{Var}(X) = \lambda.$$

The integer part of λ is always a mode of X. The mgf of X equals $\psi(t) = e^{\lambda(e^t-1)}$.

(f) If $X = X_n \sim Bin(n, p), n \to \infty, p = p_n \to 0$, and $np \to \lambda$ for some $\lambda, 0 < \lambda < \infty$, then, for any fixed k, $P(X_n = k) \to \frac{e^{-\lambda}\lambda^k}{k!}$ as $n \to \infty$.

(g) Suppose $X_1, X_2, \ldots, X_k$ are k independent binomial random variables with $X_i \sim Bin(n_i, p)$. Then $X_1 + X_2 + \cdots + X_k \sim Bin(n_1 + n_2 + \cdots + n_k, p)$.

(h) Suppose $X_1, X_2, \ldots, X_k$ are k independent negative binomial random variables, with $X_i \sim NB(r_i, p)$. Then $X_1 + X_2 + \cdots + X_k \sim NB(r_1 + r_2 + \cdots + r_k, p)$.

(i) Suppose $X_1, X_2, \ldots, X_k$ are k independent Poisson random variables with $X_i \sim Poi(\lambda_i)$. Then $X_1 + X_2 + \cdots + X_k \sim Poi(\lambda_1 + \lambda_2 + \cdots \lambda_k)$.

6.13 Exercises

Exercise 6.1. Suppose a fair coin is tossed n times. Find the probability that exactly half of the tosses result in heads when $n = 10, 30, 50$. Where does the probability seem to converge as n becomes large?

Exercise 6.2. Suppose one coin with probability .4 for heads, one with probability .6 for heads, and one that is a fair coin are each tossed once. Find the pmf of the total number of heads obtained. Is it a binomial distribution?

Exercise 6.3. Suppose the IRS audits 5% of those having an annual income exceeding 200,000 dollars. What is the probability that at least one in a group of 15 such individuals will be audited? What is the expected number that will be audited? What is the most likely number of people who will be audited?

Exercise 6.4. Suppose that each day the price of a stock moves up 12.5 cents with probability 1/3 and moves down 12.5 cents with probability 2/3. If the movements of the stock from one day to another are independent, what is the probability that after ten days the stock has its original price?

Exercise 6.5. * **(Pepy's Problem).** Find the probability that at least n sixes are obtained when $6n$ fair dice are rolled. Write a formula for it, and compute it for $1 \leq n \leq 5$. Do you see a pattern in the values?

Exercise 6.6. In repeated rolling of a fair die, find the minimum number of rolls necessary in order for the probability of at least one six to be

(a) $\geq .5$.
(b) $\geq .9$.

Exercise 6.7. * In repeated rolling of a fair die, find the minimum number of rolls necessary in order for the probability of at least k sixes to be $\geq .9$ when $k = 2, 3$.

Exercise 6.8 (System Reliability). A communication system consists of n components, each of which will independently function with probability p. The total system will be able to operate effectively if at least half of its components function. For what values of p is a five-component system more likely to operate effectively than a three-component system?

Exercise 6.9. * **(A Waiting Time Problem).** Tim, Jack, and John are going to have coffee at the local coffee shop. They will each toss a fair coin, and if one comes out as the "odd man," then he pays for all three. They keep tossing until an odd man is found. What is the probability that a decision will be reached within two rounds of tosses?

Can you generalize this with n people, a general coin with probability p of heads, and the question being what the probability is that a decision will be reached within k rounds?

Exercise 6.10. A certain firm is looking for five qualified engineers to add to its staff. If from past experience it is known that only 20% of engineers applying for a position with this firm are judged to be qualified, what is the probability that the firm will interview exactly 40 applicants to fill the five positions? At least 40 applicants to fill the five positions?

Exercise 6.11. * **(Distribution of Maximum).** Suppose n numbers are drawn at random from $\{1, 2, \cdots, N\}$. What is the probability that the largest number drawn is a specified number k if sampling is (a) with replacement; (b) without replacement?

Exercise 6.12. * **(Poisson Approximation).** One hundred people will each toss a fair coin 200 times. Approximate the probability that at least 10 of the 100 people would have obtained exactly 100 heads and 100 tails.

Exercise 6.13. * **(A Design Problem).** You are allowed to choose a number n and then toss a fair coin n times. You will get a prize if you can get either seven or nine heads. What is your best choice of the number n?

Exercise 6.14. * **(A Novel Way to Give a Test).** A student takes a five-answer multiple-choice oral test. His grade is determined by the number of questions required in order for him to get five correct answers. A grade of A is given if he requires only five questions; a grade of B is given if he requires six or seven questions; a grade of C is given if he requires eight or nine questions; and he fails otherwise.

Suppose the student guesses independently at random on each question. What is his most likely grade?

Exercise 6.15. A binomial random variable has mean 14 and variance 4.2. Find the probability that it is strictly larger than 10.

Exercise 6.16 (Distribution of Sum). The demand for the daily newspaper in a vending stall is distributed as $Bin(20, .75)$ on weekdays and $Bin(50, .75)$ on the weekend. Assuming that all days are independent, what is the distribution of the weekly demand?

Exercise 6.17 (Distribution of Difference). The demand for the daily newspaper on a Monday in a vending stall is distributed as $Bin(20, .75)$ and that on a Sunday as $Bin(50, .75)$. Find the probability that at least 20 more newspapers are sold on a Sunday than on a Monday at this stall.

Exercise 6.18. * **(Distribution of Difference).** The number of earthquakes per year in Los Angeles of magnitude greater than 4 has a mean of .5 and that in Manila, Phillipines has a mean of 1. Find the pmf of the absolute difference between the number of earthquakes of magnitude greater than 4 in the two cities and approximately calculate the mean of the absolute difference.

Exercise 6.19. * Suppose $X \sim Poi(\lambda), Y \sim Bin(n, \frac{\lambda}{n})$, and that X and Y are independent. Derive a formula for $P(X = Y)$.

Exercise 6.20. * The most likely value of a binomial random variable is 50, and the probability that it takes the value n is .357. What is its variance?

Exercise 6.21. * **(An Interesting Property of Binomial Distributions).** Suppose $X \sim Bin(n, p)$ and $Y \sim Bin(n-1, p)$. Let $a_n = \max_k P(X = k), b_n = \max_k P(Y = k)$. Show that $b_n \leq a_n$, for any p.

Exercise 6.22. Suppose a fair coin is tossed repeatedly. Find the probability that three heads will be obtained before four tails.

Generalize to r heads and s tails.

Exercise 6.23. Twelve vegetable cans, all of the same size, have lost their labels. It is known that five contain tomatoes and the rest contain beets. What is the probability that in a random sample of four cans all contain beets?

Exercise 6.24. * **(Domination of the Minority).** In a small town in Alaska, there are 60 Republicans and 40 Democrats. Ten are selected at random for a council. What is the probability that there will be more Democrats on the council than Republicans?

Exercise 6.25 (Negative Hypergeometric Distribution). In a small town in Alaska, there are 60 Republicans and 40 Democrats. The mayor wants to form a council, selecting residents at random until five Democrats have been chosen. Find the distribution and the expected value of the number of Republicans in the council and hence the distribution and the expected value of the size of the council.

Exercise 6.26. The number of customers X that enter a store during a one-hour interval has a Poisson distribution. Experience has shown that $P(X = 0) = .1111$. Find the probability that during a randomly chosen one-hour interval, more than five customers enter the store.

Exercise 6.27 (A Skewness and Kurtosis Calculation). Suppose X and Y are independent Poisson, that X has skewness .5, and that $X + Y$ has skewness $\frac{1}{3}$. What are the skewness and kurtosis of Y?

Exercise 6.28. * **(A Pretty Question).** Suppose X is a Poisson-distributed random variable. Can three different values of X have equal probabilities?

Exercise 6.29. Suppose X has a Poisson distribution such that $P(X = k) = P(X = k + 1)$ for some fixed integer k. Find the mean of X.

Exercise 6.30 (A *P*-Value Calculation). It is estimated that the risk of going into a coma with surgical anesthesia is 6 in 100,000. In the movie *Coma*, two patients out of ten go into a coma during surgery. Calculate the p-value for these data.

Exercise 6.31. * **(Couples Wishing Large families).** Suppose a couple want to have children until they have two children of each sex. What are the mean and the variance of the total number of children they will have?

Exercise 6.32 (Capture-Recapture). Suppose there are 10,000 fish in a pond. One hundred were captured, marked, and released. Then 1000 were recaptured. What is the probability that the recapture will contain more than 15 marked fish? Also do a Poisson approximation.

Exercise 6.33. Suppose X has a hypergeometric distribution. Is it possible for $E(X)$ to be equal to $\text{Var}(X)$?

Exercise 6.34. * Suppose $X \sim Poi(\lambda)$. Find an expression for the probability that X takes an even value.

Exercise 6.35. * **(A Distribution on Rationals).** Suppose X and Y are independent Poisson random variables with means λ and μ, respectively. Let $R = \frac{X}{X+Y} I_{X>0}$. Find the distribution of R; i.e., $P(R = \frac{m}{n})$, $(m, n) = 1$, and also $P(R = 0)$. What can you say about the expected value of R if $\lambda = \mu$?

Exercise 6.36 (Poisson Approximation). Assume that each of 2000 individuals living near a nuclear power plant is exposed to particles of a certain kind of radiation at the rate of one per week. Suppose that each hit by a particle is harmless with probability $1 - 10^{-5}$ and produces a tumor with probability 10^{-5}. Find the approximate distribution of:

(a) the total number of tumors produced in the whole population over a one-year period by this kind of radiation;
(b) the total number of individuals acquiring at least one tumor over a year from this radiation.

Exercise 6.37. * **(Poisson Approximation).** Twenty couples are seated at a rectangular table, husbands on one side and wives on the other, in a random order. Using a Poisson approximation, find the probability that:

(a) exactly two husbands are seated directly across from their wives;
(b) at least three are;
(c) at most three are.

Exercise 6.38 (Poisson Approximation). There are five coins on a desk, with probabilities .05, .1, .05, .01, and .04 for heads. Using a Poisson approximation, find the probability of obtaining at least one head when the five coins are each tossed once. Is the number of heads obtained binomially distributed in this problem?

Exercise 6.39 (Poisson Approximation). Typically, about 6% of guests with a confirmed reservation at a hotel with 1100 rooms do not show up. During a convention, the hotel is already completely booked. How many additional reservations can the hotel grant and be 99% sure that the number of guests with a confirmed reservation who will be denied a room is at most two?

Exercise 6.40 (Use Your Computer). Simulate the birthday problem to find the first person with the same birthday as yours. Perform the simulation 500 times. How many people did it take to find the first match? Was it typically about the same as the theoretical expected value?

Exercise 6.41 (Use Your Computer). Simulate the capture-recapture experiment with $N = 5000$ fish, a first catch of size $D = 500$, and a second catch of size $n = 250$. Perform the simulation 500 times. About how many marked fish did you find in the second catch? Did you ever see a second catch without any marked fish?

Exercise 6.42. Let $X \sim Bin(n, p)$. Prove that $P(X \text{ is even}) = \frac{1}{2} + \frac{(1-2p)^n}{2}$. Hence, show that $P(X \text{ is even})$ is larger than $\frac{1}{2}$ for any n if $p < \frac{1}{2}$ but is larger than $\frac{1}{2}$ for only even values of n if $p > \frac{1}{2}$.

References

Barbour, A. and Hall, P. (1984). On the rate of Poisson convergences, Math. Proc. Cambridge Philos. Soc., 95, 473–480.

Barbour, A., Holst, L., and Janson, S. (1992). *Poisson Approximation*, Clarendon Press, New York.

Benford, F. (1938). The law of anomalous numbers, Proc. Am. Philos. Soc., 78, 551–572.

Chen, L. (1975). Poisson approximation for dependent trials, Ann. Prob., 3, 534–545.

Diaconis, P. (1977). The distribution of leading digits and uniform distribution mod 1, Ann. Prob., 5, 72–81.

Diaconis, P. and Freedman, D. (1979). On rounding percentages, J. AM. Statist. Assoc., 74, 359–364.

Diaconis, P. and Holmes, S. (2004). *Stein's Method: Expository Lectures and Applications*, Lecture Notes and Monographs Series, IMS, Beachwood, Ohio.

Diaconis, P. and Zabell, S. (1991). Closed form summation for classical distributions, Statist. Sci., 6(3), 284–302.

Feller, W. (1968). *Introduction to Probability Theory and its Applications, Vol. I*, Wiley, New York.

Galambos, J. and Simonelli, I. (1996). *Bonferroni-Type Inequalities with Applications*, Springer-Verlag, New York.

Hill, T. (1996). A statistical derivation of the significant digit law, Statist. Sci., 10, 354–363.

Le Cam, L. (1960). An approximation theorem for the Poisson binomial distribution, Pac. J. Math., 10, 1181–1197.

Newcomb, S. (1881). A note on the frequency of use of the different digits in natural numbers, Am. J. Math., 4, 39–40.

Poincaré, H. (1896). *Calcul de Probabilités*, Georges Carré, Paris.

Poisson, S. (1838). Recherches sur la probabilité des judgements en matieres criminelles of matiere civile, Bachalier, Paris.

Steele, J. M. (1994). *Le Cam's Inequality and Poisson Approximations*, Am. Math. Mon., 101, 48–54.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, in *Proceedings of the Sixth Berkeley Symposium*, volume 2, Le Cam, L., Neyman, J. and Scott. E. eds, 583–602, University of California Press, Berkeley.

Chapter 7
Continuous Random Variables

We mentioned in Chapter 4 that discrete random variables serve as good examples to develop probabilistic intuition, but they do not account for all the random variables that one studies in theory and applications. In this chapter, we introduce the so-called *continuous random variables*, which typically take all values in some nonempty interval; e.g., the unit interval, the entire real line, etc. The right probabilistic paradigm for continuous variables cannot be pmfs. Discrete probability, which is based on summing things, is replaced by integration when we deal with continuous random variables and, instead of pmfs, we operate with a density function for the variable The density function fully describes the distribution, and calculus occupies the place of discrete operations, such as sums, when we come to continuous random variables. The basic concepts and examples that illustrate how to do the basic calculations are discussed in this chapter. Distinguished continuous distributions that arise frequently in applications will be treated separately in later chapters.

7.1 The Density Function and the CDF

The main idea about calculation of probabilities concerning continuous random variables can be understood by thinking of probabilities of events as masses of parts of an object. Suppose the material of an object, such as a ball, had uniform density. Then the mass of a portion of the ball would simply be the constant density times the volume of that part of the ball. In the case of the material having nonuniform density, to get the mass of a portion of the ball, we just have to integrate the pointwise density over that part of the ball. Points at which the material has a heavy density will contribute more to the total mass. The mass of a single point is infinitely small, but the mass of a part that has nonzero volume will have a positive mass. All of this is physically intuitive.

Likewise, for a continuous random variable, any single value x is infinitely unlikely and has probability zero; $P(X = x) = 0$ for each specific number x. But intervals with nonzero length *usually* will not have zero probability. Just as we measure the mass of part of an object, we calculate the probability of an interval as the integral of a *density function* over that interval,

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_7,

$$P(a \leq X \leq b) = \int_a^b f(x)dx,$$

the function $f(x)$ being the density function and measuring the relative concentration of probability near x rather than being the probability *at* x. You can think of a point x where the density function $f(x)$ is large as being a relatively important point. A small interval, say $x \pm \epsilon$, will have probability about $2\epsilon f(x)$, and this will be large compared with $2\epsilon f(y)$ if y is some other point with a *low density*. The important things to remember are:

I. It is not interesting to talk about probabilities of single values, $P(X = x) = 0$ for any x.
II. A density function measures the relative importance of specific values, but the density function $f(x)$ is *not* $P(X = x)$.
III. The probability of any event, that is a subset of the real line is determined by integrating the density function over that event.

Here is a formal definition of a density function.

Definition 7.1. Let X be a real-valued random variable taking values in $\mathcal{R}$, the real line. A function $f(x)$ is called the *density function* or the *probability density function* (pdf) of X if

$$\text{for all } a, b, -\infty < a \leq b < \infty, P(a \leq X \leq b) = \int_a^b f(x)dx;$$

in particular, for a function $f(x)$ to be a density function of some random variable, it must satisfy

$$f(x) \geq 0 \, \forall x \in \mathcal{R}; \int_{-\infty}^{\infty} f(x)dx = 1.$$

The statement that $P(a \leq X \leq b) = \int_a^b f(x)dx$ is the same as saying that if we plot the density function $f(x)$, then the area under the graph between a and b will give the probability that X is between a and b, while the statement that $\int_{-\infty}^{\infty} f(x)dx = 1$ is the same as saying that the area under the entire graph must be one. This is a visually helpful way to think of probabilities for continuous random variables; larger areas under the graph of the density function correspond to larger probabilities.

The density function $f(x)$ can in principle be used to calculate the probability that the random variable X belongs to a general set A, not just an interval. Indeed, $P(X \in A) = \int_A f(x)dx$.

Caution. Integrals over completely general sets A in the real line are not defined. To make this completely rigorous, one has to use measure theory and concepts of a *Lebesgue integral*. We will, however, generally only want to calculate $P(X \in A)$ for sets A that are a countable union of intervals. For such sets, defining the integral $\int_A f(x)dx$ would not be a problem and we can proceed as if we are just calculating ordinary integrals.

The definition of the CDF (cumulative distribution function) remains the same as what was given in Chapter 4.

Definition 7.2. Let X be a continuous random variable with a pdf $f(x)$. Then the CDF of X is defined as

$$F(x) = P(X \le x) = P(X < x) = \int_{-\infty}^{x} f(t)dt.$$

Remark. At any point x_0 at which $f(x)$ is continuous, the CDF $F(x)$ is differentiable, and $F'(x_0) = f(x_0)$. In particular, if $f(x)$ is continuous everywhere, then $F'(x) = f(x)$ at all x.

Again, to be strictly rigorous, one really needs to say in the sentence above that $F'(x) = f(x)$ at *almost all* x, a concept in measure theory that we will not discuss or worry about in this text. The point is that continuous random variables give zero probabilities to specific values. So, one could play with a density function and change it at one or a *few values* and still not affect *anything* about the distribution of the variable. The *almost all* phrase guards against such manipulations at a few values, which are allowed for continuous random variables. Having been warned about it, we will not worry about this again and operate as if a pdf $f(x)$ has been defined, once and for all, *at all* x.

In the discrete case, we had defined the independence of several discrete variables $X_1, X_2, \ldots, X_n$ as $P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)\ \forall\, x_1, \ldots, x_n$. One definition that works for any type of variable, discrete or not, is the following. A specialized definition for a set of continuous variables $X_1, X_2, \ldots, X_n$ will be given in a later chapter. But we emphasize that the definition below always applies.

Definition 7.3. Let $X_1, X_2, \ldots, X_n$ be n random variables defined on some sample space Ω. We say that $X_1, X_2, \ldots, X_n$ are *independent* if

$$P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) = \prod_{i=1}^{n} P(X_i \le x_i)\ \forall\, x_1, \ldots, x_n,$$

$$\Leftrightarrow P(X_1 > x_1, X_2 > x_2, \ldots, X_n > x_n) = \prod_{i=1}^{n} P(X_i > x_i)\ \forall\, x_1, \cdots, x_n.$$

We start out with examples of pdfs and CDFs and illustrate some conceptual issues using these examples.

Example 7.1 (Density vs. CDF). Consider the functions

$$f(x) = 1, \text{ if } 0 \le x \le 1; 0 \text{ if } x \notin [0, 1];$$

$$f(x) = 3x^2, \text{ if } 0 \le x \le 1; 0 \text{ if } x \notin [0, 1];$$

$$f(x) = 6x(1 - x), \text{ if } 0 \le x \le 1; 0 \text{ if } x \notin [0, 1];$$

$$f(x) = \frac{1}{\sqrt{x(1-x)}}, \text{ if } 0 \le x \le 1; 0 \text{ if } x \notin [0,1];$$

$$f(x) = 4x^2 - \frac{2}{3}x, \text{ if } 0 \le x \le 1; 0 \text{ if } x \notin [0,1].$$

We want to verify which, if any, of these functions is a valid density function.

The first four functions are all clearly nonnegative; however, the last function in our list is negative if $4x^2 < \frac{2}{3}x$ (if $x < \frac{1}{6}$), and therefore it is not a valid pdf. Thus, we only need to verify if the first four functions integrate to one. Note that each function is zero when $x \notin [0,1]$, and so $\int_{-\infty}^{\infty} f(x)dx = \int_0^1 f(x)dx$. So we need to verify whether $\int_0^1 f(x)dx = 1$ for the first four functions.

For the first two functions, it is immediately verified that $\int_0^1 f(x)dx = 1$. For the third function,

$$\int_0^1 6x(1-x)dx = 6\int_0^1 x(1-x)dx = 6\left[\int_0^1 xdx - \int_0^1 x^2dx\right] = 6\left[\frac{1}{2} - \frac{1}{3}\right] = 1;$$

for the fourth function,

$$\int_0^1 \frac{1}{\sqrt{x(1-x)}}dx = \int_0^{\pi/2} \frac{1}{\sin t \cos t} 2\sin t \cos t dt = \int_0^{\pi/2} 2dt = \pi,$$

on making the substitution $x = \sin^2 t$. Since the function integrates to π rather than to one, it is not a valid pdf; however, if we consider instead the function

$$f(x) = \frac{1}{\pi\sqrt{x(1-x)}}, \text{ if } 0 \le x \le 1; 0 \text{ if } x \notin [0,1],$$

then it is both nonnegative and integrates to one, and so it will be a valid pdf. The constant $c = \frac{1}{\pi}$ is called a *normalizing constant*.

It is instructive to see a plot of these functions on $[0,1]$ to appreciate that density functions can take a variety of shapes (see Figure 7.1). There are no shape restrictions on a density function in general, the only restriction is that they should be nonnegative and should integrate to 1. For example, of our four functions, one is a constant, one is increasing, one is symmetric, first increasing and then decreasing, fourth one is shaped like a cereal bowl (see Figure 7.2), being unbounded as $x \to 0, 1$. The density function that is constantly equal to 1 in the interval $[0,1]$ is known as the *uniform density on* $[0,1]$. The word *uniform* suggests that we uniformly assign the same importance to every value in $[0,1]$. We can analogously define a uniform density on *any* bounded interval $[a,b]$; again, the density is constant throughout the interval $[a,b]$. If a random variable X is uniformly distributed on a bounded interval $[a,b]$, we write $X \sim U[a,b]$.

Side by side, for three of these density functions, we also plot the CDF. Note that the CDF is always a smooth, nondecreasing function, starting at zero when $x = 0$

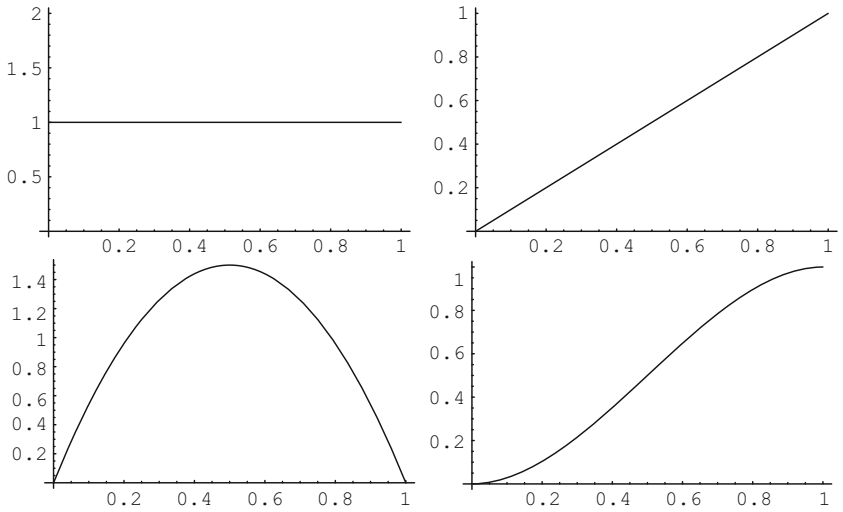


Fig. 7.1 Top: PDF (left) and CDF (right) for the first function. Bottom: PDF (left) and CDF (right) for the third function

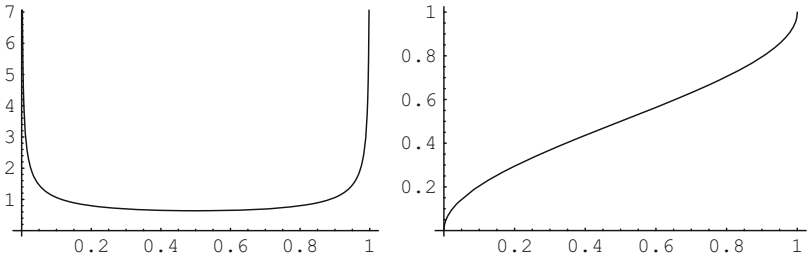


Fig. 7.2 PDF and CDF for the fourth function

and ending at one when $x = 1$ Unlike the density functions, the CDF has a certain uniformity in shape.

Example 7.2 (Density with Holes). A density function can be positive on some intervals and zero on some other intervals. Here is a very simple example. Consider the function

$$\begin{aligned} f(x) &= 48x(1-4x) \text{ if } 0 \le x \le \frac{1}{4} \\ &= 48(1-x)(4x-3) \text{ if } \frac{3}{4} \le x \le 1 \\ &= 0 \text{ if } x < 0 \text{ or } x > 1 \text{ or } \frac{1}{4} < x < \frac{3}{4}. \end{aligned}$$

This density function is plotted in Figure 7.3.

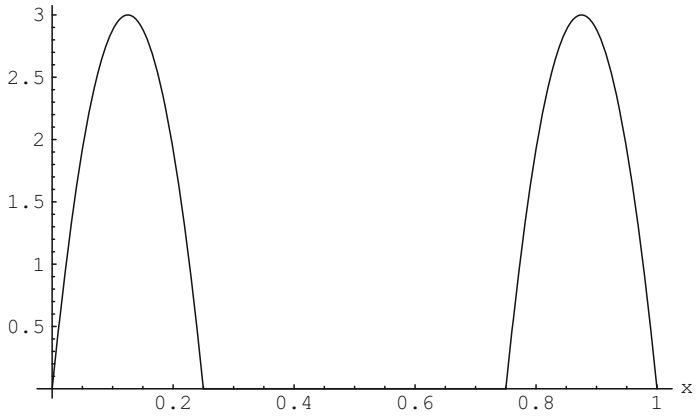


Fig. 7.3 Plot of a density with a hole

This function is always nonnegative, nonzero between 0 and .25 and between .75 and 1, and zero elsewhere; in particular, it is zero between .25 and .75. That is, it has a hole. Such a density function is trying to model a random variable that takes values near zero or near one but does not take moderate values. To model such random variables, one may have to use densities that are nonzero over various disjoint intervals but zero in between. In contrast, a CDF cannot take zero values at a point after having been positive at some previous point.

Example 7.3 (Using the Density to Calculate a Probability). Suppose X has the uniform density on $[0, 1]$, and we write $X \sim U[0, 1]$. Consider the events

$$\begin{aligned} A &= \{X \text{ is between .4 and .6}\}, \\ B &= \{X(1 - X) \leq .21\}, \\ C &= \left\{ \sin\left(\frac{\pi}{2} X\right) \geq \frac{1}{\sqrt{2}} \right\}, \\ D &= \{X \text{ is a rational number}\}. \end{aligned}$$

We will calculate each of $P(A)$, $P(B)$, $P(C)$, and $P(D)$. Recall that the probability of any event, say E, is calculated as $P(E) = \int_E f(x)dx$, where $f(x)$ is the density function, here $f(x) = 1$ on $[0, 1]$. Then,

$$P(A) = \int_{.4}^{.6} dx = .2.$$

Next, note that $x(1 - x) = .21$ has two roots in [0,1], namely $x = .3, .7$, and $x(1 - x) \leq .21$ if $x \leq .3$ or $\geq .7$. Therefore,

$$P(B) = P(X \leq .3) + P(X \geq .7) = \int_0^{.3} dx + \int_{.7}^{1} dx = .3 + .3 = .6.$$

For the event C, $\sin(\frac{\pi}{2}X) \geq \frac{1}{\sqrt{2}}$ if (and only if) $\frac{\pi}{2}X \geq \frac{\pi}{4} \Rightarrow X \geq \frac{1}{2}$. Thus,

$$P(C) = P\left(X \geq \frac{1}{2}\right) = \int_{\frac{1}{2}}^{1} dx = \frac{1}{2}.$$

Finally, the set of rationals in [0,1] is a countable set. Therefore,

$$P(D) = \sum_{x;x \text{ is rational}} P(X = x) = \sum_{x;x \text{ is rational}} 0 = 0.$$

7.1.1 Quantiles

We defined the median of a random variable in Chapter 4. That definition applies for continuous random variables also. We rephrase that definition for continuous random variables for convenience.

Definition 7.4. Let a continuous random variable X have the CDF $F(x)$. Any number m such that $F(m) = .5$ is a median of X, or equivalently a median of F.

Example 7.4 (From CDF to PDF and Median). Consider the function $F(x) = 0$ if $x < 0$; $F(x) = 1 - e^{-x}$ if $0 \leq x < \infty$. This is a nonnegative nondecreasing function that goes to one as $x \to \infty$, is continuous at any real number x, and is also differentiable at any x except $x = 0$. Thus, it is the CDF of a continuous random variable, and the PDF can be obtained using the relations $f(x) = F'(x) = e^{-x}$, $0 < x < \infty$, and $f(x) = F'(x) = 0, x < 0$. At $x = 0$, $F(x)$ is *not* differentiable. But we can define the PDF in any manner we like at one specific point, so, to be specific, we will write our PDF as

$$f(x) = e^{-x} \text{ if } 0 \leq x < \infty$$
$$= 0 \text{ if } x < 0.$$

This density is called *the standard exponential density* and is enormously important in practical applications.

From the formula for the CDF, we see that $F(m) = .5 \Rightarrow 1 - e^{-m} = .5 \Rightarrow e^{-m} = .5 \Rightarrow m = \log 2 = .693$. Thus, we have established that the standard exponential density has median $\log 2 = .693$.

Example 7.5 (CDF with a Flat Zone). Recall that, as a rule, the CDF $F(x)$ and the pdf $f(x)$ share the mutual relationship $f(x) = F'(x)$. Therefore, if the pdf $f(x)$ has a hole (that is, if $f(x)$ is zero in some interval $[a, b]$), then the CDF $F(x)$ will remain constant in that interval. If we plot the CDF, it will look flat in the interval $[a, b]$. As an example, consider our earlier example of a pdf with a hole. We obtain,

as always, the CDF from the density as $F(x) = \int_{-\infty}^{x} f(t)dt$. If we do the integral, then we find that the formula for the CDF is

$$\begin{aligned} F(x) &= 0,\ x < 0; \\ &= 24x^2 - 64x^3,\ 0 \le x \le \frac{1}{4}; \\ &= .5,\ \frac{1}{4} < x < \frac{3}{4}; \\ &= 41 - 144x + 168x^2 - 64x^3,\ \frac{3}{4} \le x \le 1; \\ &= 1,\ x > 1. \end{aligned}$$

The flat zone is the interval of values from .25 to .75 (see Figure 7.4).

This example shows that, given a number p, there can be infinitely many values x such that $F(x) = p$. Any such value splits the distribution into two parts, $100p\%$ of the probability below it and $100(1-p)\%$ above it. Such a value is called the pth quantile or percentile of F. However, in order to give a prescription for choosing a unique value when there is more than one x at which $F(x) = p$, the following definition is adopted.

Definition 7.5. Let X have the CDF $F(x)$. Let $0<p<1$. The pth *quantile* or pth *percentile* of X is defined to be the first x such that $F(x) \ge p$:

$$F^{-1}(p) = \inf\{x : F(x) \ge p\}.$$

The function $F^{-1}(p)$ is also sometimes denoted as $Q(p)$ and is called the *quantile function of F or X*.

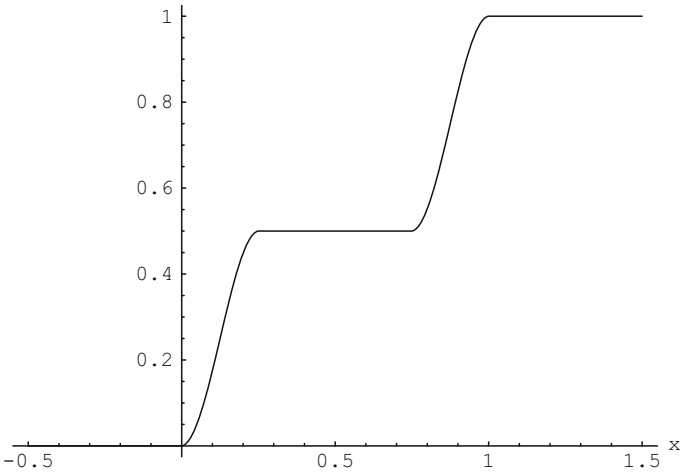


Fig. 7.4 CDF with a flat zone

Remark. Statisticians call $Q(.25)$ and $Q(.75)$ the first and third *quartiles* of F or X.

Example 7.6. Consider the CDF of Example 7.5, which had a flat zone. Then, $F^{-1}(.5)$ is .25, while $F^{-1}(.25)$ is the root of the equation $24x^2 - 64x^3 = .25$, which is $x = .125$. Likewise, $F^{-1}(.75)$ is $x = .875$.

7.2 Generating New Distributions from Old

The distribution of a continuous random variable is completely described if we describe either its density function or its CDF. For flexible modeling, it is useful to know how to create new densities or new CDFs out of densities or CDFs that we have already thought of. This is similar to generating new functions out of old functions in calculus. The following theorem describes some standard methods to make new densities or CDFs out of already available ones.

Theorem 7.1.

(a) *Let $f(x)$ be any density function. Then, for any real number μ and any $\sigma > 0$,*

$$g(x) = g_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

is also a valid density function.

(b) *Let $f_1, f_2, \ldots, f_k$ be k densities for some $k, 2 \le k < \infty$, and let $p_1, p_2, \ldots, p_k$ be k constants such that each $p_i \ge 0$ and $\sum_{i=1}^{k} p_i = 1$. Then,*

$$f(x) = \sum_{i=1}^{k} p_i f_i(x)$$

is also a valid density function.

(c) *Let $F_1, F_2, \ldots, F_k$ be k CDFs for some $k, 2 \le k < \infty$, and let $p_1, p_2, \ldots, p_k$ be k constants such that each $p_i \ge 0$ and $\sum_{i=1}^{k} p_i = 1$. Then,*

$$F(x) = \sum_{i=1}^{k} p_i F_i(x)$$

is also a valid CDF.

(d) *Let F be a CDF and α any positive real number. Then,*

$$G(x) = F^{\alpha}(x)$$

is also a valid CDF.

(e) *Let $F_1, F_2, \ldots, F_k$ be k CDFs, for some $k, 2 \le k < \infty$. Then,*

$$G(x) = F_1(x) F_2(x) \ldots F_k(x)$$

is also a valid CDF.

(f) *Let F be a CDF and $n \geq 2$ an integer. Then,*

$$G(x) = 1 - (1 - F(x))^n$$

is also a valid CDF.

Proof. The proof of each part follows from the defining properties of a density function and a CDF. For example, for part (a), since $f(x) \geq 0$ for all x, so is $g(x)$ because $\sigma > 0$. Also,

$$\int_{-\infty}^{\infty} g(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx = \int_{-\infty}^{\infty} f(z)dz = 1$$

on making the substitution $z = \frac{x-\mu}{\sigma}$.

For part (b),

$$\sum_{i=1}^{k} p_i f_i(x) \geq 0 \ \forall x$$

because each $f_i(x) \geq 0 \ \forall x$ and each $p_i \geq 0$. Also, $\int_{-\infty}^{\infty}[\sum_{i=1}^{k} p_i f_i(x)]dx = \sum_{i=1}^{k} p_i \int_{-\infty}^{\infty} f_i(x)dx = \sum_{i=1}^{k} p_i = 1$ using the fact that each f_i is a density function and integrates to one.

Part (c) is basically a restatement of part (b) but is also true for discrete random variables.

To prove part (d), observe that $G(x)$ goes to zero as $x \to -\infty$ because F does, and $G(x)$ goes to one as $x \to \infty$ because F does. Also, $G(x)$ is nondecreasing because F is, and finally, for the continuous case, G is continuous because F is. So $G(x)$ satisfies every defining property of a continuous CDF and is therefore a valid CDF.

The proofs of part (e) and part (f) use exactly the same argument.

Remark. A density of the form $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$ is called a *location scale parameter density*, the idea being that the base or the null density $f(x)$ has been shifted to some new location μ and the variable has been scaled by σ. Densities of the form $\sum_{i=1}^{k} p_i f_i(x)$ are called *mixture densities* because they are formed by *mixing* $f_1, f_2, \ldots, f_k$ according to the weights $p_1, p_2, \ldots, p_k$. Mixture densities are very useful in generating densities of various shapes and *tails*; i.e., for controlling the probabilities of values that are very large in magnitude. Densities that allow a random variable X to take large values with significant probabilities are often called *heavy-tailed densities*, and mixtures are very standard methods for generating heavy-tailed densities.

Example 7.7 (A Mixture Density). We have previously seen the following three densities on the unit interval [0, 1]:

$$f_1(x) = 1; f_2(x) = 6x(1-x); f_3(x) = \frac{1}{\pi\sqrt{x(1-x)}}.$$

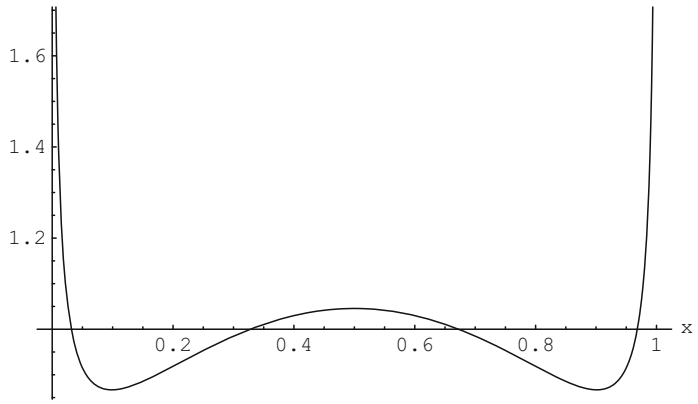


Fig. 7.5 Mixture of three densities on [0,1]

Now consider the mixture density

$$f(x) = \frac{1}{3}f_1(x) + \frac{1}{3}f_2(x) + \frac{1}{3}f_3(x).$$

A plot of this mixture density is given for illustration in Figure 7.5. Notice that the mixture density generates a new shape that is different from the shapes of each f_i.

7.3 Normal and Other Symmetric Unimodal Densities

Two very familiar concepts in probability and statistics are those of *symmetry* and *unimodality*. Symmetry of a density function means that around some point the density has two halves that are exact mirror images of each other. Unimodality means that the density has just one peak point at some value. We give the formal definitions.

Definition 7.6. A density function $f(x)$ is called *symmetric* around a number M if $f(M+u) = f(M-u)\ \forall\, u > 0$. In particular, $f(x)$ is symmetric around zero if $f(u) = f(-u)\ \forall\, u > 0$.

Definition 7.7. A density function $f(x)$ is called *strictly unimodal* at (or around) a number M if $f(x)$ is increasing for $x < M$ and decreasing for $x > M$.

Example 7.8 (The Triangular Density). Consider the density function

$$\begin{aligned} f(x) &= cx, 0 \le x \le \frac{1}{2} \\ &= c(1-x), \frac{1}{2} \le x \le 1, \end{aligned}$$

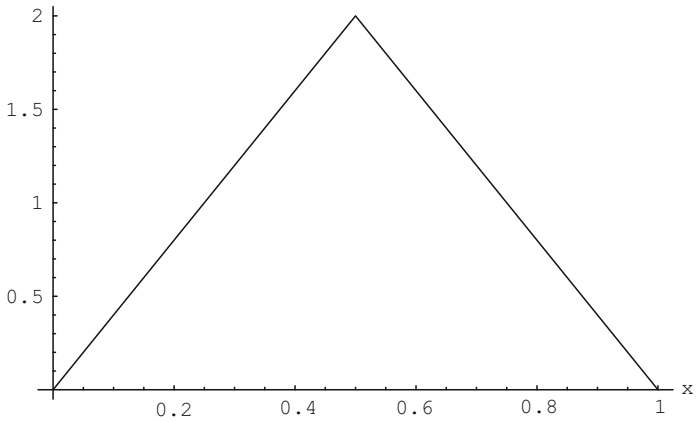


Fig. 7.6 Triangular density on [0, 1]

where c is a normalizing constant. It is easily verified that $c = 4$. This density consists of two different linear segments on $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. A plot of this density looks like a triangle (see Figure 7.6), and it is called the *triangular density* on [0, 1]. Note that it is symmetric and strictly unimodal.

Example 7.9 (The Double Exponential Density). We have previously seen the standard exponential density on $[0, \infty)$ defined as $e^{-x}, x \geq 0$. We can extend this to the negative real numbers by writing $-x$ for x in the formula above; i.e., simply define the density to be e^x for $x \leq 0$. Then, we have an overall function that equals

$$e^{-x} \text{ for } x \geq 0,$$
$$e^{x} \text{ for } x \leq 0.$$

This function integrates to

$$\int_0^\infty e^{-x} dx + \int_{-\infty}^0 e^x dx = 1 + 1 = 2.$$

So, if we use a *normalizing constant* of $\frac{1}{2}$, then we get a valid density on the entire real line:

$$f(x) = \frac{1}{2} e^{-x} \text{ for } x \geq 0,$$
$$f(x) = \frac{1}{2} e^{x} \text{ for } x \leq 0.$$

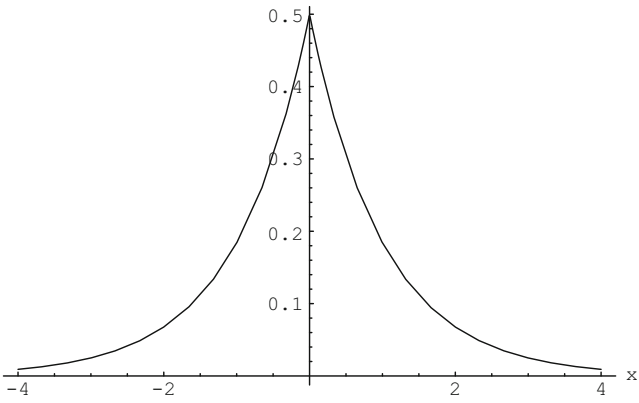


Fig. 7.7 Standard double exponential density

The two lines can be combined into one formula as

$$f(x) = \frac{1}{2}e^{-|x|}, -\infty < x < \infty.$$

This is the *standard double exponential density* and is symmetric, unimodal, and has a *cusp* at $x = 0$; see Figure 7.7.

Example 7.10 (The Normal Density). The double exponential density tapers off to zero at the linear exponential rate at both tails; i.e., as $x \to \pm\infty$. If we force the density to taper off at a quadratic exponential rate, then we will get a function like e^{-ax^2} for some chosen $a > 0$. While this is obviously nonnegative and also has a finite integral over the whole real line, it does not integrate to one. So we need a normalizing constant to make it a valid density function. Densities of this form are called *normal densities* and occupy the central place among all distributions in the theory and practice of probability and statistics. Gauss, while using the method of least squares for analyzing astronomical data, used the normal distribution to justify least squares methods; the normal distribution is also often called the *Gaussian distribution*, although de Moivre and Laplace both worked with it before Gauss. Physical data on many types of variables approximately fit a normal distribution. The theory of statistical methods is often best understood when the underlying distribution is normal. The normal distributions have many unique properties not shared by any other distribution. Because of all these reasons, the normal density, also called *the bell curve*, is the most used, most important, and most-studied distribution.

Let

$$f(x) = f(x|\mu, \sigma) = ce^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty,$$

where c is a normalizing constant. The normalizing constant can be proved to be equal to $\frac{1}{\sigma\sqrt{2\pi}}$. Thus, a normal density with parameters μ and σ is given by

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty.$$

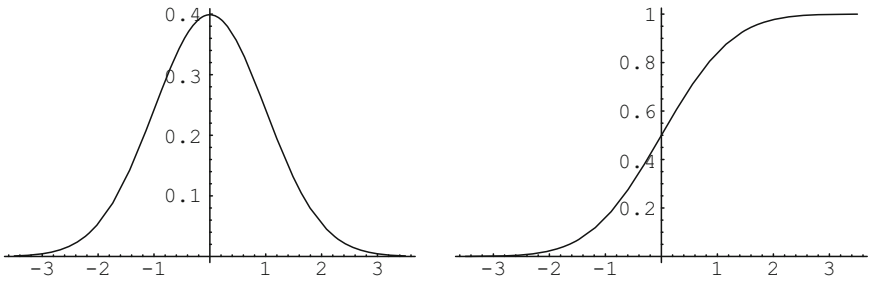


Fig. 7.8 The standard normal density and the CDF

We write $X \sim N(\mu, \sigma^2)$; we will see later that the two parameters μ and σ^2 are the mean and the variance of this distribution. Note that the $N(\mu, \sigma^2)$ density is a location-scale parameter density.

If $\mu = 0$ and $\sigma = 1$, this simplifies to the formula $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty$, and is universally denoted by the notation $\phi(x)$. It is called the *standard normal density*. The standard normal density, then, is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty.$$

Consequently, the CDF of the standard normal density is the function $\int_{-\infty}^{x} \phi(t)dt$. It is *not* possible to express the CDF in terms of the elementary functions. It is standard practice to denote it by using the notation $\Phi(x)$ and compute it using widely available tables or software for a given x, needed in an application.

A plot of the standard normal density and its CDF are given in Figure 7.8. Note the bell shape of the density function $\phi(x)$.

The normal distribution will be studied in greater detail in the next chapter.

7.4 Functions of a Continuous Random Variable

As with discrete random variables, we are often interested in the distribution of some function $g(X)$ of a continuous random variable X. For example, X could measure the input into some production process, and $g(X)$ could be a function that describes the output. Note that just because X is a continuous random variable, any function $g(X)$ need not also be a continuous random variable; $g(X)$ could be an indicator variable, for example. But, indeed, in some sense $g(X)$ often takes *as many values as X does*, in which case $g(X)$ will be a continuous random variable, too. The precise way to formulate that is the following result.

Theorem 7.2 (The Jacobian Formula). *Let X have a continuous pdf $f(x)$ and a CDF $F(x)$, and suppose $Y = g(X)$ is a strictly monotone function of X with a nonzero derivative. Then Y has the pdf*

$$f_Y(y) = \frac{f(g^{-1}(y))}{|g'(g^{-1}(y))|},$$

where y belongs to the range of g.

Proof. Since $g(X)$ is strictly monotone, it has an inverse function. Suppose $g(X)$ is strictly increasing. Then,

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(g(X) \le y) = P(X \le g^{-1}(y)) \\ &= F(g^{-1}(y)). \end{aligned}$$

On differentiating,

$$f_Y(y) = f(g^{-1}(y))\frac{d}{dy}g^{-1}(y) = \frac{f(g^{-1}(y))}{g'(g^{-1}(y))};$$

the proof for the strictly decreasing case is similar. We will often refer to this result as the *Jacobian formula*.

Remark. What kinds of functions $g(X)$ are we usually interested in? They tend to be familiar elementary functions, such as X^n for some power n, or an exponential function such as e^X, or a logarithmic function $\log X$ when X is a positive random variable, etc. Of course, for purposes of illustration, we could consider any function. We work out a few examples below.

Example 7.11 (Simple Linear Transformations). Suppose X is any continuous random variable with a pdf $f(x)$, and let $Y = g(X)$ be the linear function (a location and scale change on X) $g(X) = a + bX, b \neq 0$. This is obviously a strictly monotone function, as $b \neq 0$. Take $b > 0$. Then the inverse function of g is $g^{-1}(y) = \frac{y-a}{b}$, and of course $g'(x) \equiv b$. Putting it all together, from the theorem above,

$$f_Y(y) = \frac{f(g^{-1}(y))}{|g'(g^{-1}(y))|} = \frac{1}{b} f\left(\frac{y-a}{b}\right);$$

in general, whether b is positive or negative, the formula is

$$f_Y(y) = \frac{1}{|b|} f\left(\frac{y-a}{b}\right).$$

Example 7.12 (Uniform Functions Are Usually Not Uniform). Suppose $X \sim U[0,1]$, the uniform distribution on $[0,1]$. Let $g(X) = X^2$; note that $g(X)$ is a

strictly monotone function on [0, 1] (although it would not have been so on [−1, 1]). Furthermore, the inverse function is derived easily:

$$g(x) = x^2 = y \Rightarrow x = \sqrt{y} = g^{-1}(y).$$

Also, $g'(x) = 2x$. Putting it all together, the pdf of $Y = X^2$ is

$$f_Y(y) = \frac{f(g^{-1}(y))}{g'(g^{-1}(y))} = \frac{f(\sqrt{y})}{2\sqrt{y}} = \frac{1}{2\sqrt{y}}, 0 \le y \le 1.$$

Note, therefore, that although X is uniformly distributed, X^2 is not uniformly distributed. However, any linear function of X, say $a + bX$, will also be uniformly distributed, on the interval $[a, a + b]$.

Example 7.13 (Large Powers of Uniforms). This example is similar to the previous one, except we take a general power $g(X) = X^n$, where $X \sim U[0, 1]$. Again, $g(X)$ is a strictly monotone function on [0, 1], and the inverse function is $g^{-1}(y) = y^{1/n}$. Also, $g'(x) = nx^{n-1}$. Once again, putting it all together, the pdf of $Y = X^n$ is

$$f_Y(y) = \frac{f(g^{-1}(y))}{g'(g^{-1}(y))} = \frac{1}{ny^{(n-1)/n}} = \frac{1}{n} y^{\frac{1}{n}-1},$$

$0 < y < 1$. Let us see a plot of this pdf for a large n, say $n = 100$. The pdf, plotted in Figure 7.9, is very spiky, and *nearly all the probability* is concentrated near $y = 0$. Informally speaking, for large n, $X^n \approx 0$ with a very large probability. This is what statisticians and probabilists call *convergence in probability to zero*.

Example 7.14 (An Interesting Function that Is Not Strictly Monotone). Suppose X has the standard normal density $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ on $(-\infty, \infty)$. We want to find

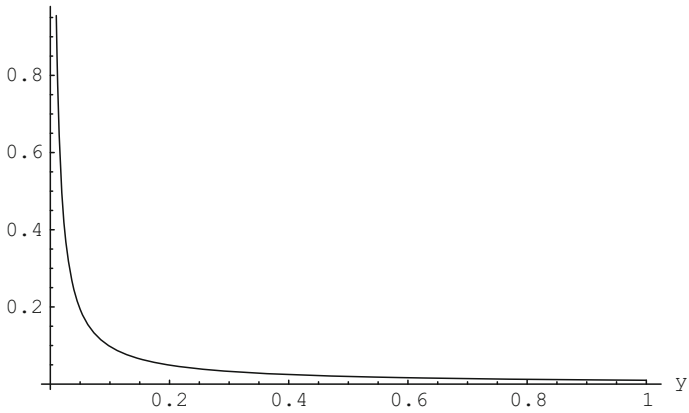


Fig. 7.9 PDF of the nth power of $U[0, 1]$; $n = 100$

the density of $Y = g(X) = X^2$. However, we immediately realize that X^2 is not a strictly monotone function on the whole real line (its graph is a parabola). Thus, the general formula given above for densities of strictly monotone functions cannot be applied in this problem. We attack the problem directly. Thus,

$$\begin{aligned} P(Y \le y) &= P(X^2 \le y) = P(X^2 \le y, X > 0) + P(X^2 \le y, X < 0) \\ &= P(0 < X \le \sqrt{y}) + P(-\sqrt{y} \le X < 0) \\ &= F(\sqrt{y}) - F(0) + [F(0) - F(-\sqrt{y})] = F(\sqrt{y}) - F(-\sqrt{y}), \end{aligned}$$

where F is the CDF of X; i.e., the standard normal CDF.

Since we have obtained the CDF of Y, we now differentiate to get the pdf of Y:

$$f_Y(y) = \frac{d}{dy}[F(\sqrt{y}) - F(-\sqrt{y})] = \frac{f(\sqrt{y})}{2\sqrt{y}} - \frac{f(-\sqrt{y})}{-2\sqrt{y}}$$

(by use of the chain rule)

$$= \frac{f(\sqrt{y})}{2\sqrt{y}} + \frac{f(\sqrt{y})}{2\sqrt{y}}$$

(since f is symmetric around zero, i.e., $f(-u) = f(u)$ for any u)

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

$y > 0$. This is a very special density in probability and statistics and is called the *chi-square density with one degree of freedom*. We have thus proved that the square of a standard normal random variable has a chi-square distribution with one degree of freedom.

There is an analogous Jacobian formula for transformations $g(X)$ that are not one-to-one. Basically, we need to break the problem up into disjoint intervals on each of which the function g is one-to-one, apply the usual Jacobian technique on each such subinterval, and then piece them together. Theorem 7.3 gives the formula.

Theorem 7.3 (Density of a Nonmonotone Transformation). *Let X have a continuous pdf $f(x)$ and let $Y = g(X)$ be a transformation of X such that, for a given y, the equation $g(x) = y$ has at most countably many roots, say $x_1, x_2, \ldots$, where the x_i depend on the given y. Assume also that g has a nonzero derivative at each x_i. Then, Y has the pdf*

$$f_Y(y) = \sum_i \frac{f(x_i)}{|g'(x_i)|}.$$

Here is an interesting example of an application of this formula.

Example 7.15 (Density of sin X*).* Suppose X has a density $f(x)$ on the real line. We want to find the density of $Y = \sin X$. Clearly, $\sin X$ is not a one-to-one transformation of X. Indeed, the equation $\sin x = y$ has the infinitely many roots

$x_i = \arcsin y + 2i\pi, i = 0, \pm 1, \pm 2, \ldots$, where $\arcsin y$ denotes the usual principal arcsine value of y (which is between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$). The derivative of the function $g(x) = \sin x$ is $g'(x) = \cos x$, and therefore

$$g'(x_i) = \cos(\arcsin y + 2i\pi) = \cos(\arcsin y)\cos(2i\pi) = \cos(\arcsin y) = \sqrt{1 - y^2}.$$

Therefore, the density of $Y = \sin X$, using an application of the formula above, is

$$f_Y(y) = \sum_{i=-\infty}^{\infty} \frac{f(\arcsin y + 2i\pi)}{\sqrt{1 - y^2}} = \frac{1}{\sqrt{1 - y^2}} \sum_{i=-\infty}^{\infty} f(\arcsin y + 2i\pi).$$

Of course, depending on how complex $f(x)$ is, there may or may not be a closed-form formula for the part $\sum_{i=-\infty}^{\infty} f(\arcsin y + 2i\pi)$.

Example 7.16 (From Exponential to Uniform). Suppose X has the standard exponential density $f(x) = e^{-x}, x \geq 0$. Let $Y = g(X) = e^{-X}$. Again, $g(X)$ is a strictly monotone function, and the inverse function is found as follows:

$$g(x) = e^{-x} = y \Rightarrow x = -\log y = g^{-1}(y).$$

Also, $g'(x) = -e^{-x}$,

$$\Rightarrow f_Y(y) = \frac{f(g^{-1}(y))}{g'(g^{-1}(y))} = \frac{e^{-(-\log y)}}{|e^{-(-\log y)}|} = \frac{y}{y} = 1, 0 \leq y \leq 1.$$

We have thus proved that if X has a standard exponential density, then $Y = e^{-X}$ is *uniformly distributed* on $[0, 1]$.

Remark. There is something special about the example above. The pdf of the standard exponential random variable is $e^{-x}, x > 0$. On the other hand, the CDF is $F(x) = 1 - e^{-x}, x > 0$ (and $F(x) = 0$ for $x \leq 0$). Our function $Y = g(X)$ in the example above is therefore $g(X) = 1 - F(X)$, and X was chosen to be standard exponential. There is actually nothing special about choosing X to be the standard exponential; the following important result says that what we saw in the example above is completely general for all continuous random variables.

7.4.1 Quantile Transformation

Theorem 7.4. *Let X have a continuous CDF $F(x)$. Consider the new random variables $Y = 1 - F(X)$ and $Z = F(X)$. Then both Y, and Z are distributed as $U[0, 1]$.*

Proof. First note that $Y = 1 - Z$. It is always the case that if $Z \sim U[0, 1]$, then so is Y. So we only prove that $Z \sim U[0, 1]$.

Recall the definition of the quantile function $Q(p) = F^{-1}(p)$ of a random variable X with CDF F. From its definition, we have

$$F^{-1}(p) \geq x \Rightarrow p \geq F(x).$$

Therefore, for any given p in $(0, 1)$,

$$P(Z \leq p) = P(F(X) \leq p) = P(X \leq F^{-1}(p)) = F(F^{-1}(p)) = p,$$

and therefore the CDF of Z matches the CDF of the $U[0, 1]$ distribution (and of course, on differentiating, the pdf of Z matches the uniform density, constantly equal to one on $(0, 1)$). Therefore, $Z \sim U[0, 1]$.

It is useful to remember this result in informal notation:

$$F(X) = U \text{ and } F^{-1}(U) = X.$$

The implication is a truly useful one. Suppose for purposes of computer experiments that we want to have computer-simulated values of *some* random variable X that has *some* CDF F and the quantile function $Q = F^{-1}$. Then, all we need to do is to have the computer generate $U[0, 1]$ values, say $u_1, u_2, \ldots, u_n$, and use $x_1 = F^{-1}(u_1), x_2 = F^{-1}(u_2), \ldots, x_n = F^{-1}(u_n)$ as the set of simulated values for our random variable of interest, namely X. Thus, the problem can be reduced to a simulation of uniform values, a simple task. The technique has so many uses that there is a name for this particular function $Z = F^{-1}(U)$ of a uniform random variable U: the quantite transformation.

Definition 7.8 (Quantile Transformation). Let U be a $U[0, 1]$ random variable and let $F(x)$ be a continuous CDF. Then the function of U defined as $X = F^{-1}(U)$ is called the *quantile transformation of* U, and it has exactly the CDF F.

What we have shown here is that we can simply start with a $U[0, 1]$ random variable and convert it to any other continuous random variable X we want simply by using a transformation of U, and that transformation is the quantile transformation.

7.4.2 Cauchy Density

Example 7.17 (The Cauchy Distribution). The Cauchy density, like the normal and the double exponential densities, is also symmetric and unimodal, but the properties are very different. It is such an atypical density that we often think of the Cauchy density first when we look for a counterexample to a conjecture. There is a very interesting way to obtain a Cauchy density from a uniform density by using the quantile transformation. We describe that derivation in this example.

Suppose a person holds a flashlight in his hand and, standing one foot away from an infinitely long wall, points the beam of light in a *random direction*. Here, by random direction we mean that the point where the light ray lands makes an angle X with the individual (considered to be a straight line one foot long), and this angle $X \sim U[-\pi/2, \pi/2]$. Let Y be the horizontal distance from the person of the point at which the light lands, with Y being considered negative if the light lands on the person's left and positive if it lands on the person's right.

Then, by elementary trigonometry,

$$\tan(X) = \frac{Y}{1} \Rightarrow Y = \tan(X).$$

Now $g(X) = \tan X$ is a strictly monotone function of X, and the inverse function is $g^{-1}(y) = \arctan(y), -\infty < y < \infty$. Also, $g'(x) = 1 + \tan^2 x$. Putting it all together,

$$f_Y(y) = \frac{\frac{1}{\pi}}{1 + [\tan(\arctan y)]^2} = \frac{1}{\pi(1 + y^2)}, -\infty < y < \infty.$$

This is the *standard Cauchy density*.

The Cauchy density is particularly notorious for its heavy tail. We will see certain consequences of this heavy tail in the next section. A plot of the standard Cauchy density stands out as much heavier tailed and at the same time more peaked than the standard normal density (see Figure 7.10).

Example 7.18 (An Oddity of the Cauchy Distribution). It is easy to show that essentially the only monotone function of a normal random variable that is also normally distributed is a linear function. However, this is not true of a Cauchy distribution; many nonlinear monotone functions of a Cauchy random variable also have Cauchy distributions. The simplest example is the function $\frac{1}{X}$.

Let X have the standard Cauchy density $f(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty$, and let $Y = g(X) = \frac{1}{X}$. This is a strictly monotone function, with the inverse function

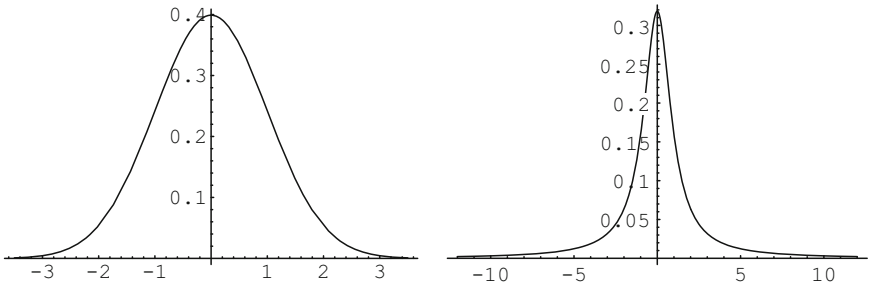


Fig. 7.10 Standard normal and standard Cauchy densities

$g^{-1}(y) = \frac{1}{y}$ and $g'(x) = -\frac{1}{x^2}$. Therefore, by the general formula for the density of a monotone function,

$$f_Y(y) = \frac{f\left(\frac{1}{y}\right)}{|g'\left(\frac{1}{y}\right)|} = \frac{1/\pi \times 1/(1+1/y^2)}{1/(1/y^2)}$$
$$= \frac{1}{\pi}\frac{y^2}{1+y^2}\frac{1}{y^2} = \frac{1}{\pi(1+y^2)},$$

which proves that $Y = \dfrac{1}{X}$ also has the standard Cauchy density.

7.5 Expectation of Functions and Moments

For discrete random variables, the expectation was seen to be equal to $\sum_x xP(X = x)$. Of course, for continuous random variables, the analogous sum $\sum_x xf(x)$ is not defined. We can think of approximating a continuous random variable by a very fine discrete random variable. To be specific, suppose X takes values in $[0, 1]$ and has the probability density $f(x)$. Now divide the interval $[0, 1]$ into a fine partition $[0, \frac{1}{n}], [\frac{1}{n}, \frac{2}{n}], [\frac{2}{n}, \frac{3}{n}], \ldots, [\frac{n-1}{n}, 1]$. On each subinterval $[(i-1)/n, i/n]$, replace X by the upper endpoint $\frac{i}{n}$ (you could replace it by the midpoint or the lower endpoint also). So, now we have devised a discrete random variable $Y = Y_n$ and, by construction, X and Y are never more than $\frac{1}{n}$ apart, which would be small if n is large. What is the expectation of Y? From our discrete case formula,

$$E(Y) = \sum_{i=1}^{n} \frac{i}{n} P\left(Y = \frac{i}{n}\right) = \sum_{i=1}^{n} \frac{i}{n} P((i-1)/n \le X \le i/n)$$
$$= \sum_{i=1}^{n} \frac{i}{n}\left[\int_{(i-1)/n}^{i/n} f(x)dx\right] \approx \sum_{i=1}^{n} \frac{i}{n}\frac{1}{n} f(i/n)$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{i}{n} f(i/n) \approx \int_0^1 xf(x)dx$$

by the definition of an integral as the limit of the (upper) Riemann sum.

This motivates the definition of expectation for continuous random variables.

Definition 7.9. Let X be a continuous random variable with a pdf $f(x)$. We say that the expectation of X exists if $\int_{-\infty}^{\infty} |x| f(x)dx < \infty$, in which case the expectation, or the expected value, or the mean of X is defined as

$$E(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx.$$

Suppose X is a continuous random variable with a pdf $f(x)$ and $Y = g(X)$ is a function of X. In the discrete case, we saw that we could calculate the expectation of Y in two equivalent ways: as $\sum_y yP(Y = y)$ or directly as $\sum_x g(x)P(X = x)$. A similar equivalence holds in the continuous case. We can compute the expectation as $\int yf_Y(y)dy$ or $\int g(x)f(x)dx$. Since Y need not always be a continuous random variable just because X is, it may not in general have a density $f_Y(y)$, but the second expression is always applicable and correct.

Theorem 7.5. *Let X be a continuous random variable with pdf $f(x)$. Let $g(X)$ be a function of X. The expectation of $g(X)$ exists if and only if $\int_{-\infty}^{\infty} |g(x)| f(x) dx < \infty$, in which case the expectation of $g(X)$ is*

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

The definitions of moments and variance remain the same as in the discrete case.

Definition 7.10. Let X be a continuous random variable with pdf $f(x)$. Then the kth moment of X is defined to be $E(X^k), k \geq 1$. We say that the kth moment does not exist if $E(|X|^k) = \infty$.

Definition 7.11. Let X be a continuous random variable with pdf $f(x)$. Suppose the expectation of X exists, and let $\mu = E(X)$. Then the variance of X is defined as $\text{Var}(X) = \sigma^2 = E[(X - \mu)^2]$.

We say that the variance of X does not exist if $E[(X - \mu)^2] = \infty$.

Proposition. Suppose X is a continuous random variable with pdf $f(x)$. Then its variance, provided it exists, is equal to

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2.$$

Proof. The first formula is simply a restatement of the definition of the variance. The second equation follows from simple algebra and is omitted.

One simple observation that saves calculations, but is sometimes overlooked, is that the proof merely uses the integration result that the integral of the product of an odd function and an even function on a symmetric interval is zero if the integral exists.

Proposition. Suppose X has a distribution symmetric around some number a; i.e., $X - a$ and $a - X$ have the same distribution. Then, $E[(X - a)^{2k+1}] = 0$, for every $k \geq 0$, provided the expectation $E[(X - a)^{2k+1}]$ exists.

For example, if X has a distribution symmetric about zero, then any odd moment (e.g., $E(X)$, $E(X^3)$, etc), provided it exists, must be zero. There is no need to calculate it; it is automatically zero.

We will now work out a number of examples.

Example 7.19 (Moments of the Uniform). Let $X \sim U[0, 1]$. We will evaluate the expectations of

$$X^k; \log X; e^{aX}; \left(X - \frac{1}{2}\right)^2.$$

First,

$$E(X^k) = \int_0^1 x^k f(x)dx = \int_0^1 x^k dx = \frac{1}{k+1}.$$

Next,

$$E(\log X) = \int_0^1 \log x dx = \int_{-\infty}^0 ye^y dy = -1,$$

on making the substitution $y = \log x$. Third,

$$E(e^{aX}) = \int_0^1 e^{ax} dx = \frac{1}{a} e^{ax}|_0^1 = \frac{e^a - 1}{a}.$$

Finally,

$$E\left[\left(X - \frac{1}{2}\right)^2\right] = E\left[X^2 - X + \frac{1}{4}\right] = \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{1}{12}.$$

Note that, since $E(X) = \frac{1}{2}$,

$$E\left[\left(X - \frac{1}{2}\right)^2\right]$$

is by definition the variance of X. Thus, we have proved that if $X \sim U[0, 1]$, then the mean and variance of X are $\frac{1}{2}$ and $\frac{1}{12}$.

Example 7.20. A lion sets a circular territory for itself by choosing a radius at random according to a standard exponential density (the unit being a mile). We want to compute the expected area of the lion's territory.

Denote by X the radius of the lion's territory; then X has the density e^{-x} for $x > 0$ and zero otherwise. Since the area of a circle with radius x is πx^2, we have

$$E(\text{area}) = \int_0^\infty \pi x^2 e^{-x} dx = \pi \int_0^\infty x^2 e^{-x} dx = 2\pi.$$

Example 7.21 (Area of a Random Triangle). Suppose an equilateral triangle is constructed by choosing the common side length X to be uniformly distributed on [0, 1]. We want to find the mean and the variance of the area of the triangle.

For a general triangle with sides a, b, c, the area equals

$$\text{area } = \sqrt{s(s-a)(s-b)(s-c)},$$

where $s = \frac{a+b+c}{2}$. When all the side lengths are equal, say, to a, this reduces to $\frac{\sqrt{3}}{4}a^2$. Therefore, in this example, we want the mean and variance of $Y = \frac{\sqrt{3}}{4}X^2$.

The mean is

$$E(Y) = \frac{\sqrt{3}}{4}E(X^2) = \frac{\sqrt{3}}{4}\frac{1}{3} = \frac{1}{4\sqrt{3}}.$$

The variance equals

$$\begin{aligned}\text{Var}(Y) &= E(Y^2) - [E(Y)]^2 = \frac{3}{16}E(X^4) - \frac{1}{48} = \frac{3}{16}\frac{1}{5} - \frac{1}{48} \\ &= \frac{3}{80} - \frac{1}{48} = \frac{1}{60}.\end{aligned}$$

Example 7.22 (Time to Response). An auto towing company services one 50 mile stretch of a highway and is situated 20 miles from one end. Breakdowns occur uniformly along the highway, and the towing company trucks travel at 50 mph. We want to find the mean and the variance of the time elapsed between the instant the company is called and that a truck arrives.

Call the left endpoint of the 50 mile stretch zero, and let X be the number of miles from the left endpoint that a breakdown occurs. Then $X \sim U[0, 50]$. Assume that the towing company is located 20 miles from the left endpoint, so that the distance Y of the breakdown from the location of the towing company is $Y = |X - 20|$. It will take the truck $Z = \frac{Y}{50} = \frac{|X-20|}{50}$ hours to reach the location of the breakdown. We want the mean and variance of Z.

First,

$$\begin{aligned}E(Z) &= E\left(\frac{|X-20|}{50}\right) = \frac{1}{50}\int_0^{50} |x-20| f(x)dx = \frac{1}{50}\frac{1}{50}\int_0^{50} |x-20|dx \\ &= \frac{1}{2500}\int_0^{20}(20-x)dx + \int_{20}^{50}(x-20)dx = \frac{1}{2500}[200+450] = .26 \text{ hours.}\end{aligned}$$

Next,

$$\begin{aligned}E(Z^2) &= \frac{1}{2500}E[(X-20)^2] = \frac{1}{2500}E[X^2 - 40X + 400] \\ &= \frac{1}{2500}\frac{1}{50}\int_0^{50}(x^2 - 40x + 400)dx = .0933,\end{aligned}$$

and therefore

$$\text{Var}(Z) = E(Z^2) - [E(Z)]^2 = .0933 - .26^2 = .0257.$$

The standard deviation would be $\sigma = \sqrt{.0257} = .16$ hours.

For the next example, we will need the definition of the *Gamma function*. It will repeatedly be necessary for us to work with the Gamma function in this text.

Definition 7.12. The Gamma function is defined as

$$\Gamma(\alpha) = \int_0^{\infty} e^{-x}x^{\alpha-1}dx, \alpha > 0.$$

In particular,

$$\Gamma(n) = (n-1)! \text{ for any positive integer } n;$$
$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \ \forall \alpha > 0;$$
$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Example 7.23 (Moments of Exponential). Let X have the standard exponential density. Then, all its moments exist and indeed

$$E(X^n) = \int_0^\infty x^n e^{-x} dx = \Gamma(n+1) = n!.$$

In particular,

$$E(X) = 1, E(X^2) = 2,$$

and therefore $\text{Var}(X) = E(X^2) - [E(X)]^2 = 2 - 1 = 1$. Thus, the standard exponential density has the same mean and variance.

Example 7.24 (A Nonsymmetric Density with All Odd Moments Zero). Consider the density function

$$f(x) = c(1 - \sin(|x|^{1/4}))e^{-x^{1/4}}, x > 0$$
$$= c(1 + \sin(|x|^{1/4}))e^{-|x|^{1/4}}, x < 0.$$

Note that $f(x) \neq f(-x)$, and therefore $f(x)$ is not symmetric. Also, it is clear that every moment of this density exists.

Consider, as an example, the first moment. The steps below use substitutions in simplifying the integration terms but are otherwise straightforward. The first moment is

$$\begin{aligned} E(X) &= c\left[\int_0^\infty xe^{-x^{1/4}}dx - \int_0^\infty x\sin(x^{1/4})e^{-x^{1/4}}dx + \int_{-\infty}^0 xe^{-(-x)^{1/4}}dx \right. \\ &\qquad \left. + \int_{-\infty}^0 x\sin((-x)^{1/4})e^{-(-x)^{1/4}}dx\right] \\ &= c\left[\int_0^\infty xe^{-x^{1/4}}dx - \int_0^\infty x\sin(x^{1/4})e^{-x^{1/4}}dx - \int_0^\infty xe^{-x^{1/4}}dx \right. \\ &\qquad \left. - \int_0^\infty x\sin(x^{1/4})e^{-x^{1/4}}dx\right] \\ &= -2c\int_0^\infty x\sin(x^{1/4})e^{-x^{1/4}}dx = 0, \end{aligned}$$

the last integral indeed being provably zero.

Similarly, any other odd moment $E(X^{2k+1})$ is also zero. This shows that although for a symmetric density any odd moment that is finite must be zero, the converse is not true. In other words, even if every odd moment of a random variable is zero, it does not mean that the distribution of the variable is necessarily symmetric about zero.

Example 7.25 (Absolute Value of a Standard Normal). This is often required in calculations in statistical theory. Let X have the standard normal distribution, and we want to find $E(|X|)$. By definition,

$$E(|X|) = \int_{-\infty}^{\infty} |x| f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |x| e^{-x^2/2} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x e^{-x^2/2} dx$$

(since $|x|e^{-x^2/2}$ is an even function of x on $(-\infty, \infty)$)

$$= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \left[\frac{d}{dx}(-e^{-x^2/2}) \right] dx = \frac{2}{\sqrt{2\pi}}(-e^{-x^2/2})|_0^{\infty}$$

$$= \frac{2}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}.$$

Example 7.26 (Discrete-Valued Function of a Continuous Random Variable). In real life, all measurements must be made on a suitable discrete scale because we cannot measure anything with true infinite precision. It is quite common to round certain measurements to their integer values; examples include temperature, age, income, etc.

As an example, suppose X has the standard exponential density, and let $Y = g(X) = \lfloor X \rfloor$ be the integer part of X. What is its mean value?

First, for purposes of examining loss of accuracy caused by rounding, note that the mean of X itself is

$$E(X) = \int_0^{\infty} x e^{-x} dx = 1.$$

But

$$E(Y) = \int_0^{\infty} \lfloor x \rfloor e^{-x} dx = \int_0^1 (0) e^{-x} dx + \int_1^2 (1) e^{-x} dx + \int_2^3 (2) e^{-x} dx + \cdots$$

$$= \sum_{i=1}^{\infty} i \int_i^{i+1} e^{-x} dx = \sum_{i=1}^{\infty} i [e^{-i} - e^{-(i+1)}]$$

$$= \sum_{i=1}^{\infty} i e^{-i} - \sum_{i=1}^{\infty} i e^{-(i+1)} = (1 - e^{-1}) \sum_{i=1}^{\infty} i e^{-i}$$

$$= (1 - e^{-1}) \frac{e}{(e-1)^2} = \frac{1}{e-1} = .582.$$

In this case, we have a nearly 42% loss of accuracy by rounding the true value of X to its integer part.

Example 7.27 (A Random Variable Whose Expectation Does Not Exist). Consider the standard Cauchy random variable with the density $f(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty$. Recall that, for $E(X)$ to exist, we must have $\int_{-\infty}^{\infty} |x| f(x)dx < \infty$. But,

$$\int_{-\infty}^{\infty} |x| f(x)dx = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{|x|}{1+x^2}dx \geq \frac{1}{\pi}\int_{0}^{\infty} \frac{x}{1+x^2}dx \geq \frac{1}{\pi}\int_{0}^{M} \frac{x}{1+x^2}dx$$

(for any $M < \infty$)

$$= \frac{1}{2\pi}\log(1+M^2),$$

and on letting $M \to \infty$, we see that

$$\int_{-\infty}^{\infty} |x| f(x)dx = \infty.$$

Therefore, the expectation of a standard Cauchy random variable, or synonymously the expectation of a standard Cauchy distribution, does not exist.

Example 7.28 (Moments of the Standard Normal). In contrast to the standard Cauchy variable, every moment of a standard normal variable exists. The basic reason is that the tail of the standard normal density is too thin. A formal proof follows.

Fix $k \geq 1$. Then,

$$|x|^k e^{-x^2/2} = |x|^k e^{-x^2/4} e^{-x^2/4} \leq C e^{-x^2/4},$$

where C is a finite constant such that $|x|^k e^{-x^2/4} \leq C$ for any real number x (such a constant C does exist). Therefore,

$$\int_{-\infty}^{\infty} |x|^k e^{-x^2/2}dx \leq C\int_{-\infty}^{\infty} e^{-x^2/4}dx < \infty.$$

Hence, by definition, for any $k \geq 1$, $E(X^k)$ exists.

Now, take k to be an odd integer, say $k = 2n+1, n \geq 0$. Then,

$$E(X^k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^{2n+1} e^{-x^2/2}dx = 0$$

because x^{2n+1} is an *odd function* and $e^{-x^2/2}$ is an *even function*. Thus, every odd moment of the standard normal distribution is zero.

Next, take k to be an even integer, say $k = 2n, n \geq 1$. Then,

$$E(X^k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2n} e^{-x^2/2} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x^{2n} e^{-x^2/2} dx$$
$$= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} z^n e^{-z/2} \frac{1}{2\sqrt{z}} dz = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} z^{n-1/2} e^{-z/2} dz$$

on making the substitution $z = x^2$.

Now make a further substitution, $u = \frac{z}{2}$. Then, we get

$$E(X^{2n}) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} (2u)^{n-1/2} e^{-u} 2du = \frac{2^n}{\sqrt{\pi}} \int_0^{\infty} u^{n-1/2} e^{-u} du.$$

Now, we recognize $\int_0^{\infty} u^{n-1/2} e^{-u} du$ to be $\Gamma(n + \frac{1}{2})$, and so we get the formula

$$E(X^{2n}) = \frac{2^n \Gamma\left(n + \frac{1}{2}\right)}{\sqrt{\pi}}, n \geq 1.$$

By using the *Gamma duplication formula*

$$\Gamma\left(n + \frac{1}{2}\right) = \sqrt{\pi} 2^{1-2n} \frac{(2n-1)!}{(n-1)!},$$

this reduces to

$$E(X^{2n}) = \frac{(2n)!}{2^n n!}, n \geq 1.$$

Example 7.29 (A Simple Two-Layered Example). Suppose each week Jack calls his mother twice and the length of each call is uniformly distributed in [5, 10] (minutes). What is the expected number of times next month that Jack's call will be over eight minutes?

Let $X_1, X_2, \ldots, X_8$ be the lengths of the eight phone calls next month, and let A_i be the event that $X_i > 8$. We assume $X_1, X_2, \ldots, X_8$ to be independent random variables. Then, $A_1, A_2, \ldots, A_8$ are independent events, and

$$T = \text{ number of calls that go over eight minutes}$$
$$= \sum_{i=1}^{8} I_{A_i} \sim Bin(8, p),$$

where $p = P(X_1 > 8) = \frac{1}{10-5} \int_8^{10} dx = .4$.

Therefore, $E(T) = 8p = 3.2$. Of course, we could have calculated the expectation directly as $\sum_{i=1}^{8} P(A_i) = 8p = 3.2$ without using the binomial distribution property.

7.6 The Tail Probability Method for Calculating Expectations

For nonnegative integer-valued random variables, we saw the formula $E(X) = \sum_{x=0}^{\infty} P(X > x)$. A similar formula exists for general random variables, in particular for continuous random variables. Even higher moments can be found by applying this technique. The tail probability $P(X > x)$ is often referred to as the *survival probability* or the *survival function*. The idea is that if X is the time from diagnosis until death of a person afflicted with a disease, then $P(X > x)$ simply measures the probability that the patient survives beyond a time period x (say x years). It could also refer to the probability that a machine failure does not occur at least until time x, etc.

7.6.1 ∗ Survival and Hazard Rate

Definition 7.13. Let X be a random variable with CDF $F(x)$. Then $\overline{F}(x)=1-F(x)$ is called the *survival function* of X. If X is continuous with a continuous pdf $f(x)$, then $h(x) = \frac{d}{dx}[-\log \overline{F}(x)] = \frac{f(x)}{\overline{F}(x)}$ is called the *instantaneous hazard rate* or simply the *hazard rate* of X.

Remark. Note that, for any random variable X, the survival function $\overline{F}(x) \to 0$ as $x \to \infty$, but it can go to zero very slowly, or also rapidly, depending on the specific random variable X. The hazard rate has the interpretation that $\delta h(x)$ is approximately the probability that a patient who has survived until time x will die within a very short time δ after time x; in this sense, $h(x)$ measures the instantaneous or the immediate risk of death.

The hazard rate of a random variable can be a *constant*, monotonically decreasing, monotonically increasing, or not monotone at all. Exponentially distributed random variables have a constant hazard rate. The hazard rate decreases for systems or devices that have an increasingly smaller chance of immediate failure as the device ages; for increasing hazard rates, the device behaves in the opposite fashion. The mortality of humans typically shows a nonmonotone hazard rate. Initially, the child has a relatively high chance of death. But the risk of immediate death decreases if the child has survived the initial period after birth. Then, it ultimately increases as the person becomes old or very old. Thus, human mortality typically leads to a *bathtub-shaped* hazard curve, at first decreasing, then becoming more or less flat, and then increasing.

7.6.2 ∗ Moments and the Tail

We now describe methods to calculate moments of a random variable from its survival function and the relationship of the rapidity with which the survival function goes to zero with various moments.

For proving the next theorem, we need a well-known result in real analysis about interchanging the order of integration in an iterated double integral; we state this below.

Theorem 7.6 (Fubini's Theorem). *Suppose $g(x, y)$ is a real-valued function on $\mathcal{R}^2$ such that the double integral $I = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |g(x, y)|dxdy < \infty$. Then,*

$$I = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(x, y)dx\right]dy = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(x, y)dy\right]dx.$$

Theorem 7.7.

(a) *Let X be a nonnegative random variable, and suppose $E(X)$ exists. Then*

$$x\overline{F}(x) = x[1 - F(x)] \to 0 \text{ as } x \to \infty.$$

(b) *Let X be a nonnegative random variable, and suppose $E(X)$ exists. Then $E(X) = \int_0^{\infty} \overline{F}(x)dx$.*

(c) *Let X be a nonnegative random variable, and suppose $E(X^k)$ exists, where $k \geq 1$ is a given positive integer. Then*

$$x^k\overline{F}(x) = x^k[1 - F(x)] \to 0 \text{ as } x \to \infty.$$

(d) *Let X be a nonnegative random variable, and suppose $E(X^k)$ exists. Then*

$$E(X^k) = \int_0^{\infty} (kx^{k-1})[1 - F(x)]dx.$$

(e) *Let X be a general real-valued random variable, and suppose $E(X)$ exists. Then*

$$x[1 - F(x) + F(-x)] \to 0 \text{ as } x \to \infty.$$

(f) *Let X be a general real-valued random variable, and suppose $E(X)$ exists. Then*

$$E(X) = \int_0^{\infty} [1 - F(x)]dx - \int_{-\infty}^{0} F(x)dx.$$

Proof. We only consider the case where X is a continuous random variable with a density $f(x)$. For part (a), note that, by hypothesis, $E(|X|) = E(X) < \infty$, and therefore $\int_x^{\infty} uf(u)du \to 0$ as $x \to \infty$. But, for any $x > 0$,

$$0 \leq x[1 - F(x)] \leq \int_x^{\infty} uf(u)du,$$

and therefore $x[1 - F(x)] \to 0$ as $x \to \infty$.

For part (b), first observe that, for any $y > 0, x = \int_0^x dy$. Therefore,

$$E(X) = E\left[\int_0^X dy\right] = \int_0^{\infty}\left[\int_0^x dy\right]f(x)dx = \int_0^{\infty}\left[\int_y^{\infty} f(x)dx\right]dy$$

(by choosing $g(x, y) = I_{y \le x}$ in Fubini's theorem)

$$= \int_0^\infty [1 - F(y)]dy.$$

The proofs of the remaining parts use the same line of argument and will be omitted.

Caution. The conditions in parts (a), (c), and (e) are only necessary and not sufficient.

Let us see two examples.

Example 7.30. Consider the density function $f(x) = \frac{1}{2x^2}, |x| \ge 1$ (and zero otherwise). Since $f(x) = f(-x)$, the distribution of X is symmetric about zero, and therefore, for all $x > 0$, $F(-x) = 1 - F(x)$. Hence,

$$x[1 - F(x) + F(-x)] = 2x[1 - F(x)] = 2x \int_x^\infty f(y)dy = 2x \times \frac{1}{2x} = 1,$$

which does not go to zero as $x \to \infty$. Therefore, $E(X)$ does not exist for this density.

Example 7.31 (Expected Value of the Minimum of Several Uniform Variables). Suppose $X_1, X_2, \ldots, X_n$ are independent $U[0, 1]$ random variables, and let $m_n = \min\{X_1, X_2, \ldots, X_n\}$ be their minimum. By virtue of the independence of $X_1, X_2, \ldots, X_n$,

$$\begin{aligned} P(m_n > x) &= P(X_1 > x, X_2 > x, \cdots, X_n > x) \\ &= \prod_{i=1}^n P(X_i > x) = (1 - x)^n, 0 < x < 1, \end{aligned}$$

and $P(m_n > x) = 0$ if $x \ge 1$. Therefore, by the theorem above,

$$\begin{aligned} E(m_n) &= \int_0^\infty P(m_n > x)dx = \int_0^1 P(m_n > x)dx = \int_0^1 (1 - x)^n dx \\ &= \int_0^1 x^n dx = \frac{1}{n + 1}. \end{aligned}$$

7.7 ∗ Moment Generating Function and Fundamental Tail Inequalities

The mgf of a random variable was defined in Chapter 5, and that definition is completely general. We recall that definition and use it to derive other useful results.

Definition 7.14. Let X be a continuous random variable with pdf $f(x)$. The moment generating function of X is defined as $\psi(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$, provided the integral is not equal to $+\infty$.

Remark. We also recall the property that if $\psi(t)$ is finite in some nonempty open interval containing zero, then it is infinitely differentiable in that open interval, and for any $n \geq 1$, $\psi^{(n)}(0) = E(X^n)$.

Example 7.32 (Moment Generating Function of Standard Exponential). Let X have the standard exponential density. Then,

$$E(e^{tX}) = \int_0^{\infty} e^{tx} e^{-x} dx = \int_0^{\infty} e^{-(1-t)x} dx = \frac{1}{1-t}$$

if $t < 1$, and it equals $+\infty$ if $t \geq 1$. Thus, the mgf of the standard exponential distribution is finite if and only if $t < 1$. So, the moments can be found by differentiating the mgf, namely $E(X^n) = \psi^{(n)}(0)$. Now, at any $t < 1$, by direct differentiation, $\psi^{(n)}(t) = \frac{n!}{(1-t)^{n+1}} \Rightarrow E(X^n) = \psi^{(n)}(0) = n!$, a result we have derived before directly.

Example 7.33 (Moment Generating Function of Standard Normal). Let X have the standard normal density. Then,

$$\begin{aligned} E(e^{tX}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \times e^{t^2/2} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz \times e^{t^2/2} = 1 \times e^{t^2/2} = e^{t^2/2} \end{aligned}$$

because $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz$ is the integral of the standard normal density, and so must be equal to one.

We have therefore proved that the mgf of the standard normal distribution exists at any real t and equals $\psi(t) = e^{t^2/2}$.

7.7.1 ∗ Chernoff-Bernstein Inequality

The mgf is useful in deriving inequalities on probabilities of tail values of a random variable that have proved to be extremely useful in many problems in statistics and probability. In particular, these inequalities typically give much sharper bounds on the probability that a random variable would be far from its mean value than Chebyshev's inequality can give. Such probabilities are called *large-deviation probabilities*. We present a particular large-deviation inequality below and then present some neat applications.

Theorem 7.8. *Let X have the mgf $\psi(t)$, and assume that $\psi(t) < \infty$ for $t < t_0$ for some $t_0, 0 < t_0 \le \infty$. Let $\kappa(t) = \log \psi(t)$, and for a real number x, define*

$$I(x) = \sup_{0<t<t_0}[tx - \kappa(t)].$$

Then,

$$P(X \ge x) \le e^{-I(x)}.$$

See Bernstein (1947) and Chernoff (1952) for this inequality and other refinements of it.

Proof. By Markov's inequality (see Chapter 4), for any real x and any $t > 0$,

$$P(X \ge x) = P(tX \ge tx) = P(e^{tX} \ge e^{tx}) \le \frac{E(e^{tX})}{e^{tx}} = e^{-tx}\psi(t) = e^{\kappa(t)-tx}.$$

Therefore,

$$\begin{aligned} P(X \ge x) &\le \inf_{0<t<t_0} e^{\kappa(t)-tx} = e^{inf_{0<t<t_0}[\kappa(t)-tx]} \\ &= e^{-\sup_{0<t<t_0}[tx-\kappa(t)]} = e^{-I(x)}. \end{aligned}$$

The function $\kappa(t)$ is called the *cumulant generating function* of X because the rth cumulant κ_r equals $\kappa^{(r)}(0)$; see Chapter 5.

To apply the Chernoff-Bernstein inequality, it is necessary to be able to find the mgf $\psi(t)$ and then be able to find the function $I(x)$, which is called the *rate function* of X. There is a huge amount of literature on the topic of large-deviation probabilities; see Bucklew (2004), Varadhan (2003), den Hollander (2000), Dembo and Zeitouni (1998), and DasGupta (2008) for detailed expositions and overviews.

We will now see an example.

Example 7.34 (Testing the Bound in the Standard Normal Case). Suppose X is a standard normal variable. Then, the exact value of the probability $P(X > x) = 1 - P(X \le x) = 1 - \Phi(x)$ is easily computable, although no formula can be written for it. The Chebyshev inequality will give, for $x > 0$,

$$P(X > x) = \frac{1}{2}P(|X| > x) \le \frac{1}{2x^2}.$$

To apply the Chernoff-Bernstein bound, use the formula $\psi(t) = e^{t^2/2} \Rightarrow \kappa(t) = t^2/2 \Rightarrow I(x) = \sup_{t>0}[tx - t^2/2] = x^2/2$. Therefore,

$$P(X > x) \le e^{-I(x)} = e^{-x^2/2}.$$

Obviously, the Chernoff-Bernstein bound is much smaller than the Chebyshev bound for large x. We have plotted in Figure 7.11 the exact value of $P(X > x)$

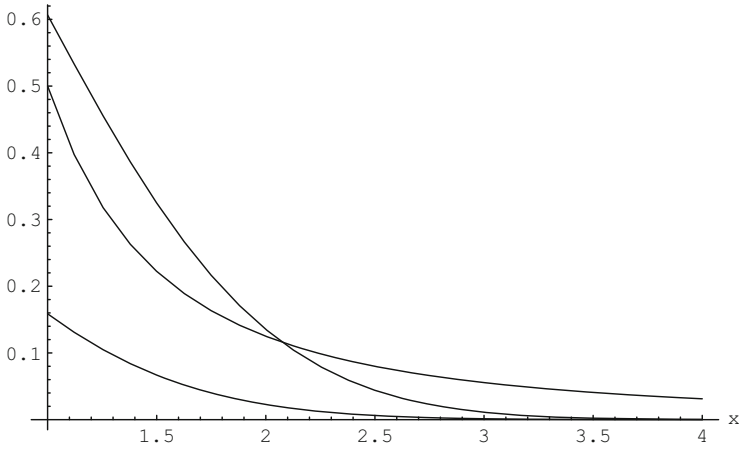


Fig. 7.11 From top: Chernoff bound, Chebyshev bound, and exact value of $P(N(0,1) > x)$

and the Chebyshev and the Chernoff-Bernstein bounds, and interestingly we see that the Chebyshev bound is better (comes closer to the exact value) if $x \leq 2.1$ (approximately), the Chernoff-Bernstein bound is better if $x > 2.1$, and for $x > 2.8$ or so, the Chernoff-Bernstein bound is much better. It turns out, however, that the Chernoff-Bernstein bound still has a large *relative error* in comparison with the exact value of $P(X > x)$; i.e., although $P(X > x) - e^{-x^2/2}$ obviously goes to zero as $x \to \infty$, the ratio $\frac{e^{-x^2/2}}{P(X>x)}$ *does not go to one*. On the contrary, the ratio goes to ∞! This follows from a result we will present in Chapter 9 that says that $\frac{e^{-x^2/2}}{P(X>x)}$ is of the exact order of $\sqrt{2\pi}x$ as $x \to \infty$.

7.7.2 ∗ Lugosi's Improved Inequality

Lugosi (2006) gives an inequality that improves on the Chernoff-Bernstein inequality for nonnegative random variables. The improved inequality is based on the moments themselves rather than the moment generating function.

Theorem 7.9. *Let X be a positive random variable. Then, for any $x > 0$,*

$$P(X > x) \leq \min_{k \geq 1} x^{-k} E(X^k) \leq e^{-I(x)},$$

where $I(x)$ is as in the Chernoff-Bernstein inequality.

The proof of the first inequality in this theorem is a trivial consequence of Markov's inequality; the proof of the second inequality uses the power series expansion of the exponential function e^x and then inequalities obtained by truncation of the power series expansion. We omit the proof.

Here is an example that illustrates the greater effectiveness of the Lugosi bound in comparison with the bound obtained from the Chernoff-Bernstein inequality.

Example 7.35. Suppose X has the standard normal density and that we want an upper bound on $P(|X| > x), x > 0$. We have previously used the Chernoff-Bernstein inequality to bound this probability. By the inequality in the theorem above,

$$P(|X| > x) = P(X^2 > x^2) \le \min_{k \ge 1} x^{-2k} E(X^{2k}) = \min_{k \ge 1} x^{-2k} \frac{(2k)!}{2^k k!};$$

here, we have used our previously derived formula for $E(X^{2k})$. One can show that $x^{-2k} E(X^{2k})$ is minimized at

$$k_0 = \lfloor \frac{x^2 + 1}{2} \rfloor.$$

Putting it back into the inequality, we get

$$P(|X| > x) \le x^{-2k_0} \frac{(2k_0)!}{2^{k_0} k_0!}.$$

As an example, take $x = 3$. Then, the bound above will on calculation give that $P(|X| \ge 3) \le .016$. In comparison, the Chernoff-Bernstein bound, on calculation, gives $P(|X| \ge 3) \le .022$; clearly, the bound due to Lugosi is quite a bit better.

7.8 ∗ Jensen and Other Moment Inequalities and a Paradox

Since the variance of the absolute value of any random variable X equals $E(X^2) - [E(|X|)]^2 \ge 0$, we have perhaps the most basic moment inequality, that $E(X^2) \ge [E(|X|)]^2$. There are numerous moment inequalities on positive and general real-valued random variables. They have a variety of uses in theoretical calculations. We present a few fundamental moment inequalities in this section.

Theorem 7.10 (Jensen's Inequality). *Let X be a random variable with a finite mean and $g(x) : \mathcal{R} \to \mathcal{R}$ a convex function. Then $g(E(X)) \le E(g(X))$.*

Proof. Denote the mean of X by μ, and suppose that g has a finite derivative $g'(\mu)$ at μ. Now consider any $x > \mu$. By the convexity of g, $\frac{g(x)-g(\mu)}{x-\mu} \ge g'(\mu) \Rightarrow g(x) - g(\mu) \ge (x - \mu)g'(\mu)$. For $x < \mu$, $\frac{g(x)-g(\mu)}{x-\mu} \le g'(\mu) \Rightarrow g(x) - g(\mu) \ge (x - \mu)g'(\mu)$. For $x = \mu, g(x) - g(\mu) = (x - \mu)g'(\mu)$. Since we have $g(x) - g(\mu) \ge (x - \mu)g'(\mu)\ \forall x$, by taking an expectation,

$$E[g(X) - g(\mu)] \ge E[(X - \mu)g'(\mu)] = 0 \Rightarrow g(\mu) \le E(g(X)).$$

When g does not have a finite derivative at μ, the proof uses the geometric property of a convex function that the chord line joining two points is always above the graph of the convex function between the two points. We will leave that case as an exercise.

Example 7.36. Let X be any positive random variable with a finite mean μ. Consider the function $g(x) = \frac{1}{x}, x > 0$. Since g is a convex function (because, for example, its second derivative $\frac{2}{x^3} > 0$ for any positive x) for $x > 0$, by Jensen's inequality

$$E\left(\frac{1}{X}\right) \geq \frac{1}{E(X)} \Leftrightarrow E\left(\frac{1}{X}\right) E(X) \geq 1,$$

and an equality holds only if $P(X = \mu) = 1$.

Example 7.37. Let X be any random variable with a finite mean μ. Consider the function $g(x) = e^{ax}$, where a is a real number. Then, by the second derivative test, g is a convex function on the entire real line and therefore, by Jensen's inequality,

$$E(e^{aX}) \geq e^{a\mu}.$$

We now state a number of other important moment inequalities.

Theorem 7.11.

(a) ***(Lyapounov Inequality).*** *Given a nonnegative random variable X and $0 < \alpha < \beta$,*

$$(EX^{\alpha})^{\frac{1}{\alpha}} \leq (EX^{\beta})^{\frac{1}{\beta}}.$$

(b) *Given $r \geq s \geq t \geq 0$ and any random variable X such that $E|X|^r < \infty$,*

$$(E|X|^r)^{s-t}(E|X|^t)^{r-s} \geq (E|X|^s)^{r-t},$$

and, in particular for any variable X with a finite fourth moment,

$$E|X| \geq \frac{(EX^2)^{\frac{3}{2}}}{\sqrt{EX^4}}.$$

(c) ***(Log Convexity Inequality of Lyapounov).*** *Given a nonnegative random variable X and $0 \leq \alpha_1 < \alpha_2 \leq \frac{\beta}{2}$,*

$$EX^{\alpha_1} EX^{\beta-\alpha_1} \geq EX^{\alpha_2} EX^{\beta-\alpha_2}.$$

(d) *Given an integer-valued random variable X with $\mu = E(X), \nu_k = E|X-\mu|^k$,*

$$\nu_k \leq 2\nu_{k+1}.$$

(e) *Given independent random variables with an identical distribution, $X_1, \ldots, X_n$, each with mean zero,*

$$E|\sum X_i| \geq \sqrt{\frac{n}{8}} E|X_1|.$$

(f) *Given independent random variables $X_1, \ldots, X_n$ with mean zero,*

$$E|\sum X_i| \leq \left(2 - \frac{1}{n}\right) \sum_{k=1}^{n} E(|X_k|).$$

References to these inequalities can be seen in DasGupta (2008, Chapter 35).

We finish with an example of a paradox of expectations.

Example 7.38 (An Expectation Paradox). Suppose X, and Y are two positive non-constant independent random variables, with the same distribution; for example, X, and Y could be independent variables with a uniform distribution on $[5, 10]$. We need the assumption that the common distribution of X and Y is such that $E(\frac{1}{X}) = E(\frac{1}{Y}) < \infty$.

Let $R = \frac{X}{Y}$. Then, by Jensen's inequality,

$$E(R) = E\left(\frac{X}{Y}\right) = E(X)E\left(\frac{1}{Y}\right) > E(X)\frac{1}{E(Y)} = 1.$$

So, we have proved that $E\left(\frac{X}{Y}\right) > 1$. But we can repeat exactly the same argument to conclude that $E\left(\frac{Y}{X}\right) > 1$. So, we seem to have the paradoxical conclusion that we expect X to be somewhat larger than Y, and we also expect Y to be somewhat larger than X.

There are many other such examples of paradoxes of expectations.

7.9 Synopsis

(a) The density function (pdf) of a continuous random variable X is a function $f(x)$ such that $f(x) \geq 0$ for all real x, and $\int_{-\infty}^{\infty} f(x)dx = 1$. Furthermore, for any $a, b, -\infty < a \leq b < \infty$, $P(a \leq X \leq b) = \int_a^b f(x)dx$.

(b) More generally, for any event A, $P(X \in A) = \int_A f(x)dx$. In particular, the CDF $F(x) = \int_{-\infty}^{x} f(t)dt$ for all real x. Conversely, $F'(x) = f(x)$ for *almost all* x.

(c) The quantile function of X is defined as $Q(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}, 0 < p < 1$. $F^{-1}(p)$ is called the pth quantile ($100p$th percentile) of X. The 50th percentile is also called the median of X.

(d) If X has the continuous CDF $F(x)$, then $Y = F(X) \sim U[0, 1]$. Conversely, if $U \sim U[0, 1]$ and F is a continuous CDF, then $F^{-1}(U)$ has F as its CDF, and this is known as the quantile transformation of U.

(e) A pdf $f(x)$ is symmetric around a number M if $f(M+u) = f(M-u)$ for all positive u. The pdf f is unimodal around M if it is increasing for $x < M$ and decreasing for $x > M$. The normal, double exponential, Cauchy, and triangular densities are all symmetric and unimodal.

(f) The standard normal density is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty.$$

The standard double exponential density is given by

$$f(x) = \frac{1}{2} e^{-|x|}, -\infty < x < \infty.$$

The standard Cauchy density is given by

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}, -\infty < x < \infty.$$

(g) If X has the pdf $f(x)$ and $Y = g(X)$ is a one-to-one function (transformation) of X, then Y has the density

$$f_Y(y) = \frac{f(g^{-1}(y))}{|g'(g^{-1}(y))|}.$$

This is the *Jacobian formula*.

(h) If X has the pdf $f(x)$ and $Y = g(X)$ is not a one-to-one function of X, then Y has the density

$$f_Y(y) = \sum_i \frac{f(x_i)}{|g'(x_i)|},$$

where $x_i = x_i(y)$ are the roots of the equation $g(x) = y$. There are additional conditions assumed for this formula to be valid.

(i) If X has the pdf $f(x)$ and $Y = g(X)$ is a function of X, then $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$. In particular,

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx; E(X^k) = \int_{-\infty}^{\infty} x^k f(x) dx;$$
$$\text{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x) dx - \left[\int_{-\infty}^{\infty} x f(x) dx \right]^2.$$

(j) If X is a nonnegative random variable, and $E(X^k)$ exists, then

$$E(X^k) = \int_0^{\infty} (k x^{k-1}) \bar{F}(x) dx,$$

where $\bar{F}(x)$ is the survival function of X. For a general real-valued random variable, if $E(X)$ exists, then

$$E(X) = \int_0^\infty \bar{F}(x)dx - \int_{-\infty}^0 F(x)dx.$$

(k) If $X \sim U[0,1]$, then

$$E(X^k) = \frac{1}{k+1}; \text{Var}(X) = \frac{1}{12}.$$

(l) If $X \sim Exp(1)$, then

$$E(X) = \text{Var}(X) = 1; E(X^n) = n!.$$

(m) If $X \sim N(0,1)$, then

$$E(X^{2k+1}) = 0 \text{ for all } k \geq 0; E(X^{2k}) = \frac{(2k)!}{2^k k!}, k \geq 1.$$

Also, $E(|X|) = \sqrt{\frac{2}{\pi}}$. The Cauchy density has the notorious property that none of its moments exist; even the mean does not exist. This is a consequence of its extremely heavy tails.

(n) Three special inequalities valid for continuous or discrete random variables are:

(1) For positive random variables, $E(X)E\left(\frac{1}{X}\right) > 1$, unless X is a constant, in which case $E(X)E\left(\frac{1}{X}\right) = 1$.

(2) Jensen's inequality: If X has mean μ, and $g(x)$ is a convex function, then $E[g(X)] \geq g(\mu)$.

(3) Chernoff-Bernstein inequality: If X has a finite mgf $\psi(t)$ for $0 < t \leq t_0$, then,

$$P(X \geq x) \leq e^{-I(x)}, \text{ where } I(x) = \sup_{0<t<t_0} [tx - \log \psi(t)].$$

7.10 Exercises

Exercise 7.1. Let $f(x) = \frac{1}{x^k}, x \geq 1$. For what values of k, if any, is $f(x)$ a density function?

Exercise 7.2. Let

$$\begin{aligned} f(x) &= c(1+x), \text{ if } -1 \leq x \leq 0; \\ &= c(1-x), \text{ if } 0 \leq x \leq 1. \end{aligned}$$

Is there any possible value of c that makes $f(x)$ a density function?

Exercise 7.3. Let $f(x) = c|x|(1+x)(1-x), -1 \leq x \leq 1$.

(a) Find the normalizing constant c that makes $f(x)$ a density function.
(b) Find the CDF corresponding to this density function. Plot it.
(c) Use the CDF to find

$$P(X < -.5); P(X > .5); P(-.5 < X < .5).$$

Exercise 7.4. Show that, for every $p, 0 \leq p \leq 1$, the function $f(x) = p \sin x + (1-p)\cos x, 0 \leq x \leq \pi/2$ (and $f(x) = 0$ otherwise), is a density function. Find its CDF and use it to find all the medians.

Exercise 7.5. * Give an example of a density function on $[0, 1]$ by giving a formula such that the density is finite at zero, unbounded at one, has a unique minimum in the open interval $(0, 1)$, and such that the median is .5.

Exercise 7.6. The time taken by a group of students to finish a quiz has the density function $f(x) = cx^2(15-x), 0 \leq x \leq 15$.

(a) Find the normalizing constant c that makes $f(x)$ a density function.
(b) What percentage of the students take more than ten minutes to finish the quiz? Less than five minutes?
(c) What is the mean time taken to finish the quiz?

Exercise 7.7. * **(A Mixed Distribution).** Suppose the damage claims on a particular type of insurance policy are uniformly distributed on $[0, 5]$ (in thousands of dollars), but the maximum payout by the insurance company is 2500 dollars. Find the CDF and the expected value of the payout, and plot the CDF. What is unusual about this CDF?

Exercise 7.8. Let $f(x)$ be the uniform density on $[0, 1]$ and let $g(x)$ be the density on $[0, 1]$ given by $g(x) = ce^x$, where c is the normalizing constant.

Let F, and G be the corresponding CDFs. Find the density function corresponding to the new CDF $H(x) = F(x)G(x)$.

Exercise 7.9. * **(Random Division).** Jen's dog broke off her six-inch-long pencil at a random point on the pencil. Find the density function and the expected value of the ratio of the lengths of the shorter piece and the longer piece of the pencil.

Exercise 7.10 (Square of a PDF Need Not Be a PDF). Give an example of a density function $f(x)$ on $[0, 1]$ such that $cf^2(x)$ cannot be a density function for any c.

Exercise 7.11 (Square Root of a PDF Need Not Be a PDF). Give an example of a density function $f(x)$ such that $c\sqrt{f(x)}$ cannot be a density function for any c.

Exercise 7.12. Give an example of two different density functions $f(x)$, and $g(x)$ on $[0, 1]$ such that $cf(x)g(x)$ cannot be a density function for any c.

Exercise 7.13. Consider the density function $f(x) = \frac{1}{x^2}, x \geq 1$. Find the quantile function corresponding to $f(x)$, and find the pth percentile for $p = .5, .9$.

Exercise 7.14. * Give an example of a continuous random variable X on $[0, 1]$ such that X has mean .5, $F(.4) = F(.6)$, and $F(.9) = .9$.

Exercise 7.15. * **(A Density with Infinitely Many Modes).** Show that the density $f(x) = c2^{-\lfloor x \rfloor} \sin(\frac{\pi}{2}[x - \lfloor x \rfloor]), x \geq 0$ has infinitely many modes.

Exercise 7.16. Suppose θ is uniformly distributed on $(-\pi/2, \pi/2)$. Find the mean and variance of

(a) $\sin\theta$;
(b) $\sin 2\theta$.

Exercise 7.17. The diameter of onion rings cut for hamburgers in a fast food restaurant has the density $f(x) = c(x - x^2 + x^3), 0 < x < 1$, where c is the normalizing constant. Find the mean and the variance of the area of a ring. You should assume the rings to be solid circles, not hollow rings.

Exercise 7.18 (Percentiles of the Standard Cauchy Density). Find the pth percentile of the standard Cauchy density for a general p and compute it for $p = .75$.

Exercise 7.19. * **(Functional Similarity).** Suppose X has the standard Cauchy density. Show that $X - \frac{1}{X}$ also has a Cauchy density. Can you find another function with this property on your own? Hint: Think of simple rational functions.

Exercise 7.20. * **(An Intriguing Identity).** Suppose X has the standard Cauchy density. Give a rigorous proof that $P(X > 1) = P(X > 2) + P(X > 3)$.

Exercise 7.21 (Inverse Chi-Square Density). Suppose X has the standard normal density. Find the density of $\frac{1}{X^2}$.

Exercise 7.22 (The Density Function of the Density Function). Suppose X has a density function $f(x)$. Find the density function of $f(X)$ when $f(x)$ is the standard normal density.

Exercise 7.23. * **(The Average Density).** Let $f(x)$ be a density that has a finite upper bound, $f(x) \leq M < \infty$. Suppose that $f(x)$ is a continuous function. Show that there is at least one number x_0 such that $E[f(X)] = f(x_0)$. Find all values of x_0 when $f(x)$ is the standard normal density.

Exercise 7.24. * **(Integer Part).** Suppose X has a uniform distribution on $[0, 10.5]$. Find the expected value of the integer part of X.

Exercise 7.25 (Random Triangle). The lengths of the three sides of a triangle are $X, 2X, 2.5X$, where X is uniformly distributed on $[0, 1]$. Find the mean and the variance of the area of the triangle.

Exercise 7.26 (An Optimization Problem). Suppose the location of an archaeological treasure is distributed along a 50 mile stretch according to the density $f(x) = cx^2(50 - x), 0 < x < 50$, where c is the normalizing constant. A company is planning to dig along the fifty mile stretch for the treasure, and they need to select a location for their headquarters. The cost of transportation of the treasure from its spot of discovery to the headquarters is a function $g(d)$, where d is the distance between those two points. Find the optimum location for the headquarters if

(a) $g(d) = d^2$;
(b) $g(d) = d$;
(c) $g(d) = \log(1 + d)$.

Exercise 7.27 (Fractional Normal Moments). Suppose X has a standard normal distribution. Find a general formula for $E(|X|^\alpha), \alpha > 0$. Does $E(|X|^\alpha)$ exist for any $\alpha < 0$?

Exercise 7.28 (Hazard Rate). Find and plot the hazard rate for the *folded Cauchy density* $f(x) = \frac{2}{\pi(1+x^2)}, x > 0$.

Exercise 7.29. Find and plot the hazard rate for the density $f(x) = ce^{-x^\alpha}, x, \alpha > 0$, where c is the normalizing constant.

Exercise 7.30 (Expectation and Hazard Rate). For a general nonnegative random variable, write a formula for the expectation in terms of the hazard rate, assuming that the expectation exists.

Exercise 7.31. Let X be a positive random variable with the CDF $F(x)$. Show that $\int_0^\infty [1 - F(\sqrt{x})]dx \geq (\int_0^\infty [1 - F(x)]dx)^2$. When are they equal?

Exercise 7.32 (Maximum of Uniforms). Let $X_1, X_2, \ldots, X_n$ be n independent $U[0, 1]$ random variables. Find an expression for $E[\max\{X_1, \ldots, X_n\}]$.

Exercise 7.33 (Minimum of Exponentials). Let $X_1, X_2, \ldots, X_n$ be n independent standard exponential random variables. Find an expression for $E[\min\{X_1, \ldots, X_n\}]$.

Exercise 7.34. Suppose X is a positive random variable with mean one. Show that $E(\log X) \leq 0$.

Exercise 7.35. Suppose X is a positive random variable with four finite moments. Show that $E(X)E(X^3) \geq [E(X^2)]^2$.

Exercise 7.36. Suppose X has a geometric distribution with parameter $p = \frac{1}{2}$. Show that $E(X \log X) \geq \log 4$.

Exercise 7.37 (Rate Function for the Exponential). Derive the rate function $I(x)$ for the standard exponential density and hence derive a bound for $P(X > x)$.

Exercise 7.38. * (Rate Function for the Double Exponential). Derive the rate function $I(x)$ for the double exponential density and hence derive a bound for $P(X > x)$.

Exercise 7.39 (Use Your Computer). Simulate a set of 500 values from a standard exponential density by using the quantile transformation method. Find the mean of your 500 simulated values. Is it close to the theoretical mean value?

Exercise 7.40 (Use Your Computer). Simulate a set of 500 values from a standard Cauchy density by using the quantile transformation method. What are the most striking features you see in your simulated values?

Exercise 7.41 (Use Your Computer). Simulate a set of 500 values from the density $f(x) = c\cos^2 x$ on $[0, \pi]$ by using the quantile transformation method. Find the mean of your simulated values. Is it close to the theoretical mean value?

References

Bernstein, S. (1947). *Theory of Probability* (Russian), Moscow Leninghad.
Bucklew, J. (2004). *Introduction to Rare Event Simulation*, Springer, New York.
Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann. Math. Statist., 23, 493–507.
DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer, New York.
Dembo, A. and Zeitouni, O. (1998). *Large Deviations, Techniques and Applications*, Springer, New York.
den Hollander, F. (2000). *Large Deviations*, Fields Institute Monograph, AMS, Providence, RI.
Lugosi, G. (2006). *Concentration of Measure Inequalities*, Lecture Notes, Dept. of Economics, Pompeu Fabra University., Barcelona.
Varadhan, S.R.S. (2003). *Large Deviations and Entropy*, Princeton University Press, Princeton, NJ.

Chapter 8
Some Special Continuous Distributions

A number of densities, by virtue of their popularity in modeling or because of their special theoretical properties, are considered to be special. In this chapter, we present a collection of these densities with their basic properties. We discuss, when suitable, their moments, the form of the CDF, the mgf, shape and modal properties, and interesting inequalities. Classic references to standard continuous distributions are Johnson et al. (1994) and Kendall and Stuart (1976); Everitt (1998) contains many unusual facts. The normal distribution is treated separately in the next chapter because of its unique importance in statistics and probability.

8.1 Uniform Distribution

The uniform distribution is the continuous analog of random selection from a finite population. A typical example is that of choosing *a random fraction*. Small measurement errors sometimes are approximately uniformly distributed. In Bayesian statistics, a uniform distribution is often used to reflect lack of knowledge about an unknown parameter. Uniform distributions can only be defined on bounded sets; for instance, there is no such thing as a uniform distribution on the real line.

Definition 8.1. Let X have the pdf

$$f(x) = \frac{1}{b-a}, a \le x \le b,$$
$$= 0 \text{ otherwise},$$

where $-\infty < a < b < \infty$ are given real numbers.

Then we say that X has the uniform distribution on $[a, b]$ and write $X \sim U[a, b]$. We derive the basic properties of the $U[a, b]$ density next.

Theorem 8.1.

(a) *If* $X \sim U[0, 1]$, *then* $a+(b-a)X \sim U[a, b]$, *and if* $X \sim U[a, b]$, *then* $\frac{X-a}{b-a} \sim U[0, 1]$.

(b) *The CDF of the* $U[a, b]$ *distribution equals*

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_8,

$$F(x) = 0, x < a;$$
$$= \frac{x-a}{b-a}, a \le x \le b;$$
$$= 1, x > b.$$

(c) *The mgf of the* $U[a,b]$ *distribution equals* $\psi(t) = \frac{e^{tb}-e^{ta}}{(b-a)t}$.
(d) *The nth moment of the* $U[a,b]$ *distribution equals*

$$E(X^n) = \frac{b^{n+1}-a^{n+1}}{(b-a)(n+1)}.$$

(e) *The mean and the variance of the* $U[a,b]$ *distribution equal*

$$\mu = \frac{a+b}{2}; \sigma^2 = \frac{(b-a)^2}{12}.$$

Proof. Part (a) follows from the general result (see Chapter 7) that if X has density $f(x)$, then a linear function, say $Y = c + dX$, has density $\frac{1}{|d|} f(\frac{y-c}{d})$.

For part (b), it is clear that $F(x) = 0$ for $x < a$ and 1 for $x > b$. For $a \le x \le b$, $F(x) = \int_{-\infty}^{x} f(t)dt = \int_a^x f(t)dt = \int_a^x 1/(b-a)dt = \frac{x-a}{b-a}$. Parts (c) and (d) follow by direct integration. For part (e), the mean is simply the first moment, and so, by part (d), $E(X) = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$, and the variance formula follows by using $\sigma^2 = E(X^2) - \mu^2$, where $E(X^2) = \frac{b^3-a^3}{3(b-a)} = (a^2+ab+b^2)/3$.

Example 8.1. A point is selected at random on the unit interval, dividing it into two pieces with total length 1. Find the probability that the ratio of the length of the shorter piece to the length of the longer piece is less than $1/4$.

Let $X \sim U[0,1]$; we want $P\left(\frac{\min\{X,1-X\}}{\max\{X,1-X\}} < 1/4\right)$. This happens only if $X < 1/5$ or $> 4/5$. Therefore, the required probability is $P(X < 1/5) + P(X > 4/5) = 1/5 + 1/5 = 2/5$.

Example 8.2. Suppose $X \sim U[-1,1]$. We want to find the conditional probability $P(|X| < 1/3 \mid |X| < 1/2)$. By definition,

$$P(|X| < 1/3 \mid |X| < 1/2) = \frac{P(|X| < 1/3 \cap |X| < 1/2)}{P(|X| < 1/2)} = \frac{P(|X| < 1/3)}{P(|X| < 1/2)}$$
$$= \frac{1/3}{1/2} = 2/3.$$

Example 8.3. The diameters (measured in centimeters) of circular strips made by a machine are uniform in the interval $[0,2]$. Strips with an area larger than 3.1 cm^2 cannot be used. If 200 strips are made in one shift, what is the expected number that have to be discarded?

The area of a circular strip of radius r is πr^2. Therefore, the radius $r \sim U[0, 1]$. We have

$$p = P(\pi r^2 > 3.1) = P(r^2 > 3.1/\pi) = P(r^2 > .9868)$$
$$= P(r > .9934) = .0066.$$

Therefore, the number of strips among 200 (independent ones) that cannot be used has the $Bin(200, .0066)$ distribution and its expected value is $200 \times .0066 = 1.32$.

8.2 Exponential and Weibull Distributions

We defined the standard exponential density in Chapter 7 and now introduce the general exponential density. Exponential densities are used to model waiting times (e.g., waiting times for an elevator or at a supermarket checkout), failure times, (e.g., the time until the first failure of some piece of equipment), or renewal times (e.g., time elapsed between successive earthquakes at a location), etc. The exponential density also has some very interesting theoretical properties.

Definition 8.2. A nonnegative random variable X has the exponential distribution with parameter $\lambda > 0$ if it has the pdf $f(x) = \frac{1}{\lambda} e^{-x/\lambda}, x > 0$. We write $X \sim Exp(\lambda)$.

A plot of the exponential density with $\lambda = 1$, called the standard exponential density, shows that it is a decreasing and bounded density on $(0, \infty)$ (see Figure 8.1).

Here are the basic properties of an exponential density.

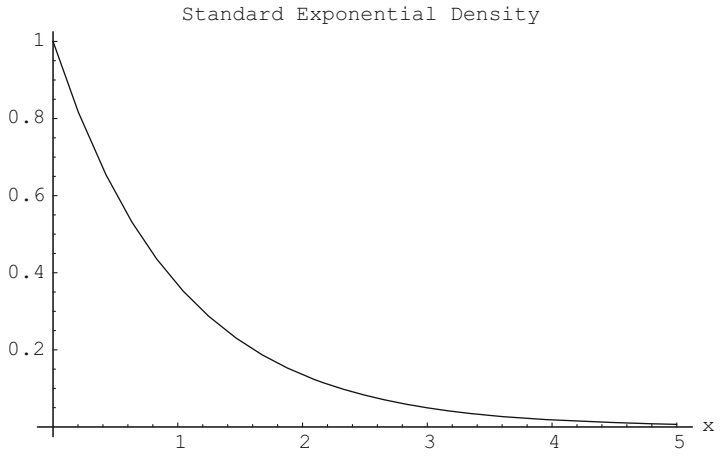


Fig. 8.1 Standard exponential density

Theorem 8.2. *Let $X \sim Exp(\lambda)$. Then,*

(a) $\frac{X}{\lambda} \sim Exp(1)$,
(b) *The CDF* $F(x) = 1 - e^{-x/\lambda}, x > 0$, *(and zero for* $x \leq 0$.*)*
(c) $E(X^n) = \lambda^n n!, n \geq 1$,
(d) *The mgf* $\psi(t) = \frac{1}{1-\lambda t}, t < 1/\lambda$.

Proof. Part (a) follows from the general result that if X has density $f(x)$, then bX has density $\frac{1}{|b|} f(x/b)$; we identify b with $1/\lambda$ and use the formula of the $Exp(\lambda)$ density.

Part (b) follows by direct integration, and part (c) is proved as

$$E(X^n) = E(\lambda \times X/\lambda)^n = \lambda^n E(X/\lambda)^n = \lambda^n n!,$$

as the standard exponential density has nth moment equal to $n!$ (see Chapter 7).

Part (d) also follows on direct integration.

Example 8.4 (Mean Is Larger than Median for Exponential). Suppose $X \sim Exp(4)$. What is the probability that $X > 4$?

Since $X/4 \sim Exp(1)$,

$$P(X > 4) = P(X/4 > 1) = \int_1^\infty e^{-x} dx = e^{-1} = .3679,$$

quite a bit smaller than 50%. This implies that the median of the distribution has to be smaller than 4, where 4 is the mean. Indeed, the median is a number m such that $F(m) = \frac{1}{2}$ (the median is unique in this example) $\Rightarrow 1 - e^{-m/4} = \frac{1}{2} \Rightarrow m = 4 \log 2 = 2.77$.

This phenomenon that the mean is larger than the median is quite typical of distributions that have a long right tail, such as the exponential.

In general, if $X \sim Exp(\lambda)$, the median of X is $\lambda \log 2$.

Example 8.5 (The Spares Problem). The general version of this problem is very interesting. We will work out only a specific example for ease of illustration.

Tires of a certain make have an exponentially distributed lifetime with a mean of 10,000 miles. How many spare tires should one keep on a 15,000 mile trip to be 60% sure that it would not be necessary to procure more tires during the trip?

The probability that at least one of the four tires in the car will fail during the trip is

$$\begin{aligned} 1 - P(\text{None will fail}) &= 1 - P(\text{Each tire works for 15,000 miles or more}) \\ &= 1 - \left[\int_{15000}^\infty 1/10000\, e^{-x/10000} dx\right]^4 = 1 - \left[\int_{1.5}^\infty e^{-x} dx\right]^4 \\ &= 1 - (.2231)^4 = .9975, \end{aligned}$$

and so certainly carrying no spares at all will get us into trouble.

How about the probability that all four tires fail during the trip? This is

$$[P(\text{One tire works for less than 15,000 miles})]^4$$
$$= \left[\int_0^{1.5} e^{-x}dx\right]^4 = (.7769)^4 = .3643.$$

Therefore, the probability that at most three will fail is $1 - .3643 = .6357$, which exceeds our 60% threshold. In fact, three spare tires just suffice, as can be verified by a similar calculation that the chances that at most two tires will fail is $< .6$.

Example 8.6 (Lack of Memory of the Exponential Distribution). The exponential densities have a lack of memory property similar to the one we established for the geometric distribution. Let $X \sim Exp(\lambda)$, and let s and t be positive numbers. The lack of memory property is that $P(X > s + t|X > s) = P(X > t)$. So, suppose that X is the waiting time for an elevator, and suppose that you have already waited $s = 3$ minutes. Then the probability that you have to wait another two minutes is the same as the probability that you would have to wait two minutes if you just arrived at the elevator. This is not true if the waiting time distribution is something other than exponential.

The proof of the property is simple:

$$P(X > s + t|X > s) = \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-(s+t)/\lambda}}{e^{-s/\lambda}}$$
$$= e^{-t/\lambda} = P(X > t).$$

Example 8.7 (Fractional Part of an Exponential). Suppose $X \sim Exp(\lambda)$. We had previously found the expected value of the integer part when $\lambda = 1$ (the standard exponential case). We will now find the expected value of the *fractional part* of X for a general λ.

First note that the fractional part $\{X\}$ equals $X - \lfloor X \rfloor$. Therefore, $E(\{X\}) = E(X) - E(\lfloor X \rfloor) = \lambda - E(\lfloor X \rfloor)$. Now,

$$E(\lfloor X \rfloor) = \sum_{n=0}^{\infty} nP(n \le X < n + 1) = \sum_{n=0}^{\infty} n \int_n^{n+1} 1/\lambda\, e^{-x/\lambda}dx$$
$$= \sum_{n=1}^{\infty} n(e^{-n/\lambda} - e^{-n/\lambda-1/\lambda}) = \sum_{n=1}^{\infty} n(1 - e^{-1/\lambda})e^{-n/\lambda}$$
$$= (1 - e^{-1/\lambda}) \sum_{n=1}^{\infty} ne^{-n/\lambda}$$
$$= (1 - e^{-1/\lambda}) \frac{e^{-1/\lambda}}{(1 - e^{-1/\lambda})^2} = \frac{1}{e^{1/\lambda} - 1}.$$

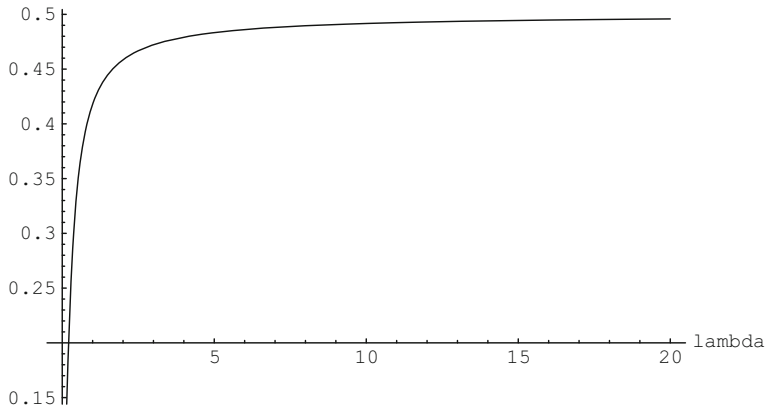


Fig. 8.2 Expected value of the fractional part of an exponential

Therefore,

$$E(\{X\}) = \lambda - \frac{1}{e^{1/\lambda} - 1}.$$

We plot the expected value of the fractional part in Figure 8.2, and we notice that the expected value is always $< .5$ and converges monotonically to .5 when $\lambda \to \infty$.

Example 8.8 (Geyser Eruption). The number of eruptions per t minutes at the Old Faithful Geyser at Calistoga, California, is Poisson with mean $.02t$. If you arrived at the geyser at 12:00 noon, what is the density of the waiting time until you see an eruption?

Denote the waiting time by X. Then the event $X > t$ is the same as saying that no eruptions occurred in the first t minutes. We are told that Y, the number of eruptions in a t minute interval, is Poisson with mean $.02t$. Therefore,

$$P(X > t) = P(Y = 0) = e^{-.02t},$$

and hence the density of X is

$$f(t) = .02e^{-.02t}.$$

Therefore, the waiting time has an exponential distribution with mean $\frac{1}{.02} = 50$ minutes. This is a well-known link between the exponential density and events that occur according to a so-called *Poisson process*.

Example 8.9 (The Weibull Distribution). Suppose $X \sim Exp(1)$, and let $Y = X^\alpha$, where $\alpha > 0$ is a constant. Since this is a strictly monotone function with the inverse function $y^{1/\alpha}$, the density of Y is

$$\begin{aligned} f_Y(y) &= \frac{f(y^{1/\alpha})}{|g'(y^{1/\alpha})|} = e^{-y^{1/\alpha}} \times \frac{1}{\alpha y^{(\alpha-1)/\alpha}} \\ &= \frac{1}{\alpha} y^{(1-\alpha)/\alpha} e^{-y^{1/\alpha}}, \, y > 0. \end{aligned}$$

This final answer can be made to look a little simpler by writing $\beta = \frac{1}{\alpha}$. If we do so, the density becomes

$$\beta y^{\beta-1} e^{-y^\beta}, y > 0.$$

We can introduce an extra scale parameter akin to what we do for the exponential case itself. In that case, we have the general two-parameter Weibull density

$$f(y|\beta, \lambda) = \frac{\beta}{\lambda}\left(\frac{x}{\lambda}\right)^{\beta-1} e^{-(\frac{x}{\lambda})^\beta}, y > 0.$$

This is the *Weibull density* with parameters β, λ.

8.3 Gamma and Inverse Gamma Distributions

The exponential density is decreasing on $[0, \infty)$. A generalization of the exponential density with a mode usually at some strictly positive number m is the Gamma distribution. It includes the exponential as a special case and can be very skewed, to being almost a bell-shaped density. We will later see that it also arises naturally as the density of the sum of a number of independent exponential random variables.

Definition 8.3. A positive random variable X is said to have a Gamma distribution with shape parameter α and scale parameter λ if it has the pdf

$$f(x|\alpha, \lambda) = \frac{e^{-x/\lambda} x^{\alpha-1}}{\lambda^\alpha \Gamma(\alpha)}, x > 0, \alpha, \lambda > 0;$$

we write $X \sim G(\alpha, \lambda)$. The Gamma density reduces to the exponential density with mean λ when $\alpha = 1$; for $\alpha < 1$, the Gamma density is decreasing and unbounded, while for large α it becomes nearly a bell-shaped curve. A plot of some Gamma densities reveals these features (see Figure 8.3).

The basic facts about a Gamma distribution are given in the following theorem.

Theorem 8.3.

(a) *The CDF of the $G(\alpha, \lambda)$ density is the* normalized incomplete Gamma function

$$F(x) = \frac{\gamma(\alpha, x/\lambda)}{\Gamma(\alpha)},$$

where $\gamma(\alpha, x) = \int_0^x e^{-t} t^{\alpha-1} dt$.

(b) *The nth moment equals*

$$E(X^n) = \lambda^n \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)}, n \geq 1.$$

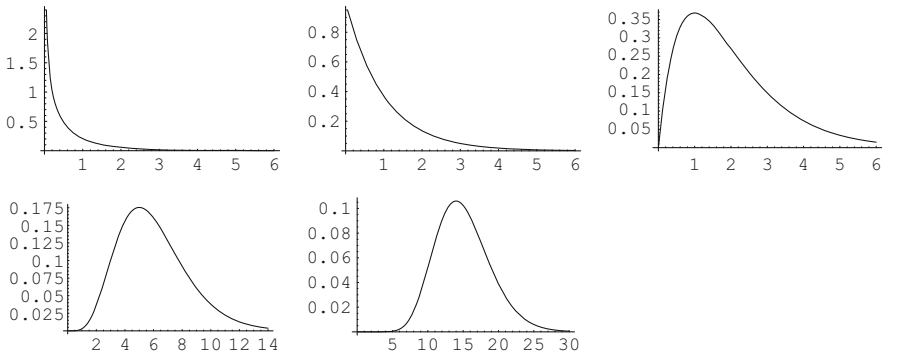


Fig. 8.3 Plot of gamma density with lambda = 1, alpha = .5, 1, 2, 6, 15

(c) *The mgf equals*

$$\psi(t) = (1 - \lambda t)^{-\alpha}, t < \frac{1}{\lambda}.$$

(d) *The mean and the variance equal*

$$\mu = \alpha\lambda; \sigma^2 = \alpha\lambda^2.$$

Proof. The CDF formula is simply a restatement of the definition of the function $\gamma(\alpha, x)$. For part (b),

$$\begin{aligned} E(X^n) &= \frac{\int_0^\infty e^{-x/\lambda} x^{\alpha+n-1} dx}{\lambda^\alpha \Gamma(\alpha)} = \frac{\lambda^{\alpha+n}\Gamma(\alpha+n)}{\lambda^\alpha\Gamma(\alpha)} \\ &= \lambda^n \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}. \end{aligned}$$

For part (c), if $t < \frac{1}{\lambda}$,

$$\begin{aligned} E(e^{tX}) &= \frac{\int_0^\infty e^{tx} e^{-x/\lambda} x^{\alpha-1} dx}{\lambda^\alpha\Gamma(\alpha)} = \frac{\int_0^\infty e^{-(1/\lambda - t)x} x^{\alpha-1} dx}{\lambda^\alpha\Gamma(\alpha)} \\ &= \frac{\Gamma(\alpha)}{(1/\lambda - t)^\alpha \lambda^\alpha \Gamma(\alpha)} = \frac{1}{(1/\lambda - t)^\alpha \lambda^\alpha} \\ &= (1 - \lambda t)^{-\alpha}. \end{aligned}$$

The mgf does not exist if $t \geq \frac{1}{\lambda}$.

The mean and the variance formulas follow from part (b). An important consequence of the mgf formula is the following result.

Corollary. Suppose $X_1, X_2, \ldots, X_n$ are independent $Exp(\lambda)$ variables. Then $X_1 + X_2 + \cdots + X_n \sim G(n, \lambda)$.

Proof. Since $X_1, X_2, \ldots, X_n$ are independent, for $t < 1/\lambda$,

$$\begin{aligned} E(e^{t(X_1+X_2+\cdots+X_n)}) &= E(e^{tX_1}e^{tX_2}\cdots e^{tX_n}) \\ &= E(e^{tX_1})E(e^{tX_2})\cdots E(e^{tX_n}) \\ &= (1-\lambda t)^{-1}(1-\lambda t)^{-1}\cdots(1-\lambda t)^{-1} = (1-\lambda t)^{-n}, \end{aligned}$$

which agrees with the mgf of a $G(n, \lambda)$ distribution and therefore, by the distribution determining the property of mgfs, it follows that $X_1 + X_2 + \cdots + X_n \sim G(n, \lambda)$.

Example 8.10 (Total Demand of a Commodity). Suppose 40 people have been invited to a party and the amount of diet soda that each guest will consume is distributed as $Exp(8)$ (in ounces); that is, it is exponentially distributed with a mean of 8 oz. If two bottles of soda, each containing 200 oz, are available, what is the probability that the supply will fall short of the demand?

Let $n = 40$ and $X_1, X_2, \ldots, X_n$ be the amount of soda consumed by the n guests. We assume that $X_1, X_2, \ldots, X_n$ are independent $Exp(8)$ variables. Then, the total demand $X_1 + X_2 + \cdots + X_n \sim G(n, 8)$, so the probability that the supply will fall short of the demand is

$$\begin{aligned} P(X_1 + X_2 + \cdots + X_n > 400) &= 1 - P(X_1 + X_2 + \cdots + X_n \le 400) \\ &= 1 - \frac{\gamma(40,400/8)}{\Gamma(40)} = 1 - .93543 = .065, \end{aligned}$$

where $\gamma(40, 50)$ was computed on a computer and $\Gamma(40) = 39!$.

Example 8.11 (The Skewness of a Gamma Distribution). We saw in our Gamma density plots that the density appears to become nearly bell-shaped when the shape parameter α becomes large. Since the *coefficient of skewness* is an index of asymmetry in a distribution, we may expect to see that it becomes small when α becomes large. Indeed, by definition, the coefficient of skewness is

$$\begin{aligned} \beta &= \frac{E(X-\mu)^3}{\sigma^3} = \frac{E(X^3) - 3\mu E(X^2) + 2\mu^3}{\sigma^3} \\ &= \frac{\alpha(\alpha+1)(\alpha+2)\lambda^3 - 3\alpha^2(\alpha+1)\lambda^3 + 2\alpha^3\lambda^3}{\alpha^{3/2}\lambda^3} \\ &= \frac{2\alpha}{\alpha^{3/2}} = 2/\sqrt{\alpha} \end{aligned}$$

$\to 0$ as $\alpha \to \infty$, as we had anticipated.

Example 8.12 (The General Chi-Square Distribution). We saw in the previous chapter that the distribution of the square of a standard normal variable is the chi-square distribution with one degree of freedom. A natural question is what the distribution of the sum of squares of several independent standard normal variables

is. Although we do not yet have the technical tools necessary to derive this distribution, it turns out that it is in fact a Gamma distribution. Precisely, if $X_1, X_2, \ldots, X_m$ are m independent standard normal variables, then $T = \sum_{i=1}^{m} X_i^2$ has a $G\left(\frac{m}{2}, 2\right)$ distribution and therefore has the density

$$f_m(t) = \frac{e^{-t/2} t^{m/2-1}}{2^{m/2}\Gamma\left(\frac{m}{2}\right)}, t > 0.$$

This is called the *chi-square density with m degrees of freedom* and arises in numerous contexts in statistics and probability. We write $T \sim \chi_m^2$. From the general formulas for the mean and variance of a Gamma distribution, we get that

$$\text{mean of a } \chi_m^2 \text{ distribution } = m;$$
$$\text{variance of a } \chi_m^2 \text{ distribution } = 2m.$$

The chi-square density is rather skewed for small m but becomes approximately bell-shaped when m gets large; we have seen this for general Gamma densities.

One especially important context in which the chi-square distribution arises is when considering of the distribution of the *sample variance* for iid normal observations. The sample variance of a set of n random variables $X_1, X_2, \ldots, X_n$ is defined as $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, where $\bar{X} = \frac{X_1+\cdots+X_n}{n}$ is the mean of $X_1, \ldots, X_n$. The name sample variance derives from the following property.

Theorem 8.4. *Suppose $X_1, \ldots, X_n$ are independent with a common distribution F having a finite variance σ^2. Then, for any n, $E(s^2) = \sigma^2$.*

Proof. First note the algebraic identity

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.$$

Therefore,

$$E(s^2) = \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] = \frac{1}{n-1}\left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] = \sigma^2.$$

If, in particular, $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then $\frac{X_i - \bar{X}}{\sigma}$ are also normally distributed, each with mean zero. However, they are no longer independent. If we sum their squares, then the sum of the squares will still be distributed as a chi square, but there will be a loss of one degree of freedom due to the fact that $X_i - \bar{X}$ are not independent even though the X_i are independent.

We state this important fact formally in the following theorem.

Theorem 8.5. *Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. Then* $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$.

Example 8.13 (Inverse Gamma Distribution). Suppose $X \sim G(\alpha, \lambda)$. The distribution of $\frac{1}{X}$ is called the inverse Gamma distribution. We will derive its density.

Since $Y = g(X) = \frac{1}{X}$ is a strictly monotone function with the inverse function $g^{-1}(y) = \frac{1}{y}$, and since the derivative of g is $g'(x) = -\frac{1}{x^2}$, the density of Y is

$$f_Y(y) = \frac{f\left(\frac{1}{y}\right)}{|g'\left(\frac{1}{y}\right)|} = \frac{e^{-1/(\lambda y)} y^{1-\alpha}}{\lambda^\alpha \Gamma(\alpha)} \frac{1}{y^2}$$

$$= \frac{e^{-1/(\lambda y)} y^{-1-\alpha}}{\lambda^\alpha \Gamma(\alpha)}, y > 0.$$

The inverse Gamma density (see Figure 8.4) is extremely skewed for small values of α; furthermore, the right tail is so heavy for small α that the mean does not exist if $\alpha \leq 1$. Inverse Gamma distributions are quite popular in studies of economic inequality, reliability problems, and as prior distributions in Bayesian statistics.

Example 8.14 (Simulating a Gamma Variable). If the shape parameter α is an integer, say $\alpha = n$, then it is simple to simulate values from a Gamma distribution. Here is why. Consider an $Exp(1)$ random variable X. Its CDF is $F(x) = 1 - e^{-x}$. Setting it equal to p, we get the quantile function

$$F(x) = p \Leftrightarrow 1 - e^{-x} = p \Leftrightarrow x = -\log(1 - p).$$

So the quantile function is $F^{-1}(p) = -\log(1 - p)$. Therefore, by the general quantile transform method, if we take $U \sim U[0, 1]$, then $-\log(1 - U)$ will have an

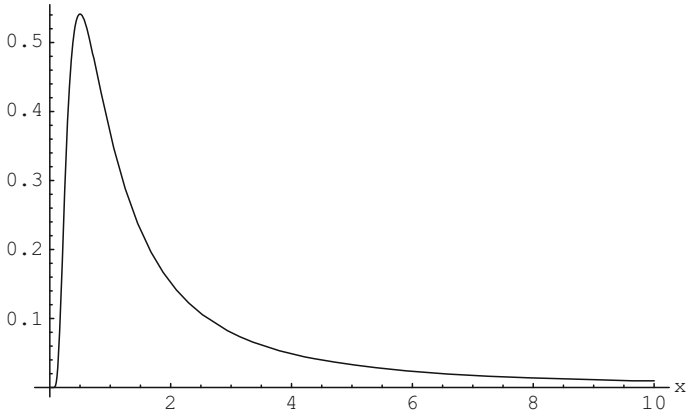


Fig. 8.4 Inverse gamma density when alpha = lambda =1

$Exp(1)$ distribution. But if $U \sim U[0,1]$, then $1-U$ is also distributed as $U[0,1]$. So, we can just take $-\log U$ as an $Exp(1)$ random variable. To get a simulated value for $G(n,1)$, we need to add n independent standard exponentials; i.e., if we want $Y \sim G(n,1)$, we can use $Y = -\log U_1 - \log U_2 - \cdots - \log U_n = -\log(U_1 U_2 \ldots U_n)$, where $U_1, U_2, \ldots, U_n$ are n independent $U[0,1]$ values. To get a simulated value for $G(n,\lambda)$, we simply multiply this Y value obtained by λ.

8.4 Beta Distribution

Beta densities are the most commonly used densities for random variables that take values between 0 and 1. Their popularity is due to their analytic tractability and the large variety of shapes that Beta densities can take when the parameter values change. The Beta density is a generalization of the $U[0,1]$ density.

Definition 8.4. X is said to have a Beta density with parameters α and β if it has the density

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, 0 \le x \le 1, \alpha, \beta > 0,$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. We write $X \sim Be(\alpha,\beta)$. An important point is that, by its very notation, $\frac{1}{B(\alpha,\beta)}$ must be the normalizing constant of the function $x^{\alpha-1}(1-x)^{\beta-1}$; thus, another way to think of $B(\alpha,\beta)$ is that, for any $\alpha, \beta > 0$,

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

This fact will be useful repeatedly in the following.

Theorem 8.6. *Let $X \sim Be(\alpha,\beta)$.*

(a) *The CDF equals*

$$F(x) = \frac{B_x(\alpha,\beta)}{B(\alpha,\beta)},$$

where $B_x(\alpha,\beta)$ is the incomplete Beta function $\int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt$.

(b) *The nth moment equals*

$$E(X^n) = \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)}.$$

(c) *The mean and the variance equal*

$$\mu = \frac{\alpha}{\alpha+\beta}; \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

(d) *The mgf equals*

$$\psi(t) =_1 F_1(\alpha, \alpha+\beta, t),$$

where ${}_1F_1(a,b,z)$ *denotes the confluent hypergeometric function.*

Proof. The formula for the CDF is a restatement of the definition of the incomplete Beta function. Regarding the moment formula,

$$E(X^n) = \frac{\int_0^1 x^{\alpha+n-1}(1-x)^{\beta-1}dx}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx}$$

$$= \frac{B(\alpha+n,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)}$$

on using the definition of the function $B(a,b)$ for $a,b>0$. The mean is just the first moment, and the variance formula follows on using the formulas for $E(X^2)$ and $E(X)$ and using $\sigma^2 = E(X^2) - [E(X)]^2$. Finally, the mgf formula follows from the integral representation of the confluent hypergeometric function

$${}_1F_1(a,a+b,z) = \frac{1}{B(a,b)} z^{1-a-b} \int_0^z e^x x^{a-1}(1-x)^{b-1}dx.$$

This integral representation is a fact in advanced calculus, and we just use it in order to derive our mgf formula here.

A major appeal of the family of Beta densities is that it produces densities of many shapes. A Beta density can be increasing, decreasing, symmetric and unimodal, unimodal but asymmetric, or U-shaped. Its only shortcoming is that it cannot be *bimodal*, i.e., it cannot have two local maxima in the interval [0, 1]. A few Beta densities are plotted in Figure 8.5 to show the various shapes that they can take.

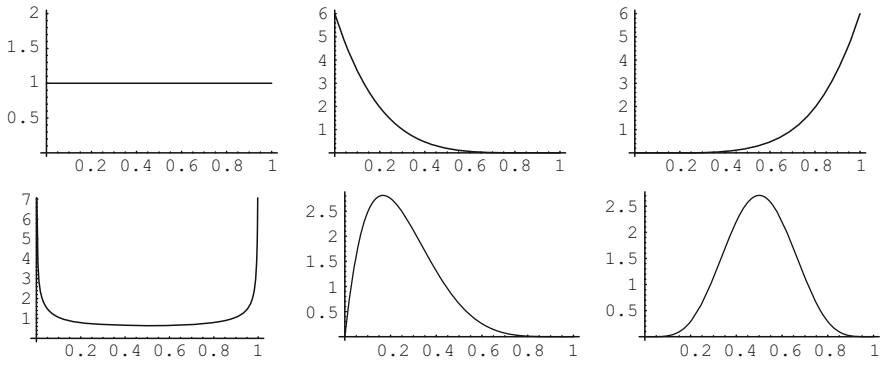


Fig. 8.5 Six beta densities $Be(1,1)$; $Be(1,6)$; $Be(6,1)$; $Be(.5,.5)$; $Be(2,6)$; $Be(6,6)$

Example 8.15 (Fitting a Beta Density). Suppose a standardized one hour exam takes 45 minutes on average to finish and the standard deviation of the finishing times is ten minutes. We want to know what percentage of examinees finish in less than 40 minutes.

We cannot answer this question if we know only the mean and the standard deviation of the distribution of finishing times. But if we use a Beta density as the density of the finishing time, then we can answer the question because we can uniquely determine the parameters of a Beta distribution from its mean and variance. Converting from minutes to hours, we want to solve for α, β from the two equations

$$\frac{\alpha}{\alpha+\beta} = \frac{3}{4},$$
$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{36}.$$

From the first equation, we get $\alpha = 3\beta$. Substituting this into the second equation, we get $3/(16(4\beta+1)) = \frac{1}{36}$. Solving, we get $\beta = 1.44$ and $\alpha = 4.32$. Therefore,

$$P\left(X < \frac{2}{3}\right) = \frac{\int_0^{2/3} x^{3.32}(1-x)^{\cdot 44}dx}{B(4.32, 1.44)} = \frac{.0283}{.1006} = .281.$$

So if we fit a Beta distribution to the information that was given to us, we will conclude that 28.1% of the examinees can finish the test in less that 40 minutes.

Example 8.16 (Square of a Beta). Suppose X has a Beta density. Then, X^2 also takes values in $[0, 1]$, but it does not have a Beta density. To give a specific example, suppose $X \sim Be(7, 7)$. Then, the density of $Y = X^2$ is

$$f_Y(y) = \frac{f(\sqrt{y})}{2\sqrt{y}} = \frac{y^3(1-\sqrt{y})^6}{B(7,7)2\sqrt{y}} = 6006y^{5/2}(1-\sqrt{y})^6, 0 \le y \le 1.$$

Clearly, this is not a Beta density.

Example 8.17 (Mixture of Two Beta Densities). It was remarked before that a Beta density cannot have two modes in $(0, 1)$. This can be a deficiency in modeling some random variables that have two modes for some inherent physical reason. To circumvent this deficiency, we can use a suitable mixture of Beta densities. Consider for example the mixture density

$$f(x) = .5f_1(x) + .5f_2(x),$$

where f_1 and f_2 are densities of $Be(6, 2)$ and $Be(2, 6)$, respectively. Thus,

$$f(x) = \frac{1}{2}[42x^5(1-x)] + \frac{1}{2}[42x(1-x)^5] = 21x(1-x)[x^4 + (1-x)^4], 0 \le x \le 1.$$

A plot of this mixture density shows the two modes (see Figure 8.6).

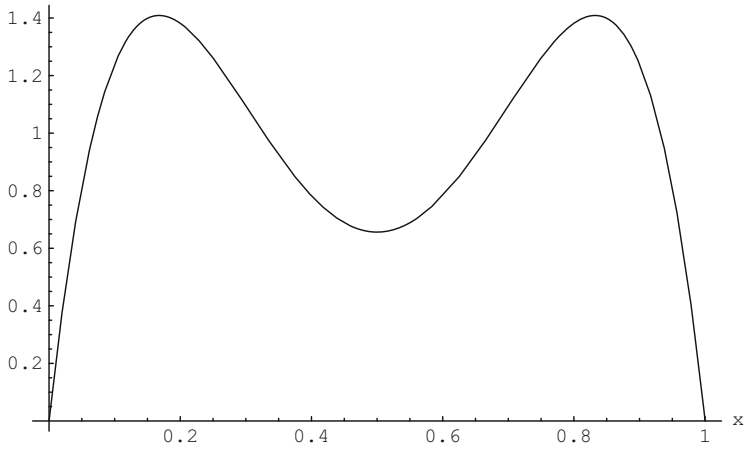


Fig. 8.6 A mixture of two betas can be bimodal

8.5 Extreme-Value Distributions

In practical applications, certain types of random variables consistently exhibit a long right tail in the sense that a lot of small values are mixed with a few large or excessively large values in the distributions of these random variables. Economic variables such as wealth typically manifest such heavy-tail phenomena. Other examples include sizes of oil fields, insurance claims, stock market returns, river height in a flood, etc. The tails are sometimes so heavy that the random variable may not even have a finite mean. Extreme value distributions are common and increasingly useful models for such applications. A brief introduction to two specific extreme-value distributions is provided in this section. These two distributions are the *Pareto* distribution and the *Gumbel* distribution. One peculiarity of semantics is that the Gumbel distribution is often called the *Gumbel law*.

A random variable X is said to have the Pareto density with parameters θ and α if it has the density

$$f(x) = \frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}, x \geq \theta > 0, \alpha > 0.$$

We write $X \sim Pa(\alpha, \theta)$. The density is monotonically decreasing. It may or may not have a finite expectation, depending on the value of α. It never has a finite mgf in any nonempty interval containing zero. The basic facts about a Pareto density are given in the next result.

Theorem 8.7. *Let* $X \sim Pa(\alpha, \theta)$.

(a) *The CDF of* X *equals*

$$F(x) = 1 - \left(\frac{\theta}{x}\right)^{\alpha}, x \geq \theta,$$

and zero for $x < \theta$.

(b) *The nth moment exists if and only if $n < \alpha$, in which case*

$$E(X^n) = \frac{\alpha\theta^n}{\alpha - n}.$$

(c) *For $\alpha > 1$, the mean exists; for $\alpha > 2$, the variance exists. Furthermore, they equal*

$$E(X) = \frac{\alpha\theta}{\alpha - 1}; \text{Var}(X) = \frac{\alpha\theta^2}{(\alpha - 1)^2(\alpha - 2)}.$$

Proof. Each part follows from elementary calculus and integration. For example,

$$E(X^n) = \int_\theta^\infty x^n \frac{\alpha\theta^\alpha}{x^{\alpha+1}} dx = \alpha\theta^\alpha \int_\theta^\infty \frac{1}{x^{\alpha-n+1}} dx,$$

which converges if and only if $\alpha - n > 0 \Leftrightarrow n < \alpha$, in which case the formula for $E(X^n)$ follows by simply evaluating the integral. The formula for the mean is just the special case $n = 1$, and that for the variance is found by using the fact that the variance equals $E(X^2) - [E(X)]^2$.

A particular Pareto density is plotted in Figure 8.7; the heavy right tail is clear.

We next define the Gumbel law. A random variable X is said to have the Gumbel density with parameters μ and σ if it has the density

$$f(x) = \frac{1}{\sigma} e^{\left(-e^{-\frac{x-\mu}{\sigma}}\right)} e^{-\frac{x-\mu}{\sigma}}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0.$$

If $\mu = 0$ and $\sigma = 1$, the density is called the *standard Gumbel density*. Thus, the standard Gumbel density has the formula $f(x) = e^{-e^{-x}} e^{-x}, -\infty < x < \infty$. The density converges extremely fast (at a superexponential rate) at the left tail, but only at a regular exponential rate at the right tail. Its relation to the density of the

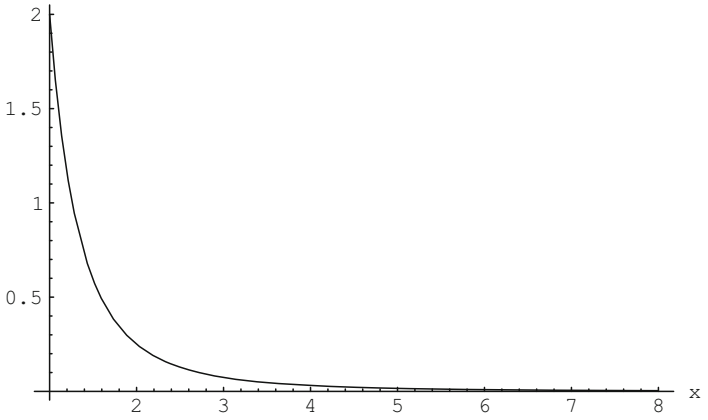


Fig. 8.7 Pareto density with theta = 1, alpha = 2

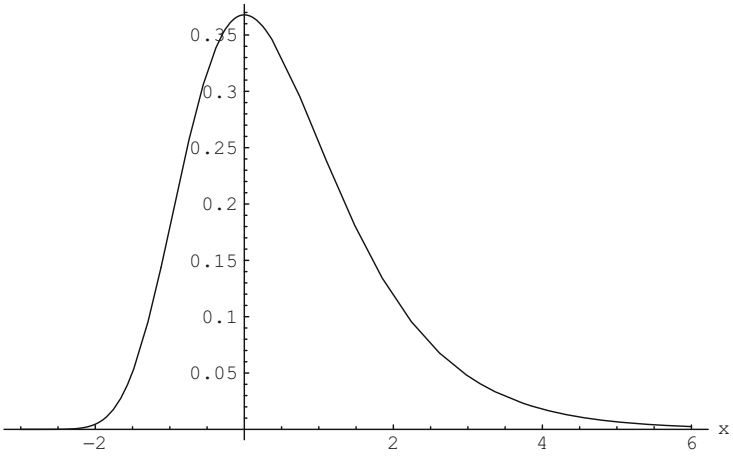


Fig. 8.8 Standard Gumbel density

maximum of a large number of independent normal variables makes it a special density in statistics and probability. The basic facts about a Gumbel density are collected together in the result below. All Gumbel distributions have a finite mgf $\psi(t)$ at any t. But no simple formula for it is possible.

Theorem 8.8. *Let X have the Gumbel density with parameters μ, σ. Then,*

(a) *The CDF equals*

$$F(x) = e^{\left(-e^{-\frac{x-\mu}{\sigma}}\right)}, -\infty < x < \infty.$$

(b) $E(X) = \mu - \gamma\sigma$, *where* $\gamma \approx .577216$ *is the* Euler constant.
(c) $\text{Var}(X) = \frac{\pi^2}{6}\sigma^2$.
(d) *The mgf of X exists everywhere.*

We will not prove this result, except by making the comment that by differentiating the given formula for $F(x)$ we indeed get the formula for $f(x)$. This is a proof by inspection of part (a) of this theorem. The other parts require integration tricks and are omitted.

The standard Gumbel density is plotted in Figure 8.8. The right tail is clearly much heavier than the left tail.

8.6 ∗ Exponential Density and the Poisson Process

A single theme that binds together a number of important probabilistic concepts and distributions and is at the same time a major tool for the applied probabilist and the applied statistician is the *Poisson process*. The Poisson process is a probabilistic model of situations where *events* occur completely at random at intermittent

times and we wish to study the number of times the particular event has occurred up to a specific time instant, or perhaps the waiting time until the next event, etc. Some simple examples are receiving phone calls at a telephone call center, receiving an e-mail from someone, arrival of a customer at a pharmacy or some other store, catching a cold, occurrence of earthquakes, mechanical breakdown in a computer or some other machine, and so on. There is no end to how many examples we can think of where an event happens, then nothing happens for a while, and then it happens again, and it keeps going like this, apparently at random. It is therefore not surprising that the Poisson process is such a valuable tool in the probabilist's toolbox. It is also a fascinating feature of the Poisson process that it is connected in various interesting ways to a number of special distributions, including the exponential and the Gamma in particular. These embracing connections and wide applications make the Poisson process a very special topic in probability. A detailed treatment of the Poisson process will be made in the companion volume of this book; below, we only give an elementary introduction.

The Poisson process is a special family of an *uncountably infinite number* of nonnegative random variables, indexed by a running label t. We call t the time parameter and, for the purpose of our discussion here, it belongs to the infinite interval $[0, \infty)$. For each $t \geq 0$, there is a nonnegative random variable $X(t)$, that counts how many events have occurred up to and including time t. As we vary t, we can think of $X(t)$ as a function. It is a *random function* because each $X(t)$ is a random variable. Like all functions, $X(t)$ has a graph. The graph of $X(t)$ is called a *path* of $X(t)$. It is helpful to look at a typical path of a Poisson process (see Figure 8.9).

We notice that the path is a nondecreasing function of the time parameter t and that it increases by jumps of size one. The time instants at which these jumps occur are called the *renewal or arrival times of the process*. Thus, we have an infinite sequence of *arrival times*, say $Y_1, Y_2, Y_3, \ldots$; the first arrival occurs exactly at time Y_1, the second arrival occurs at time Y_2, and so on. We define Y_0 to be zero. The gaps

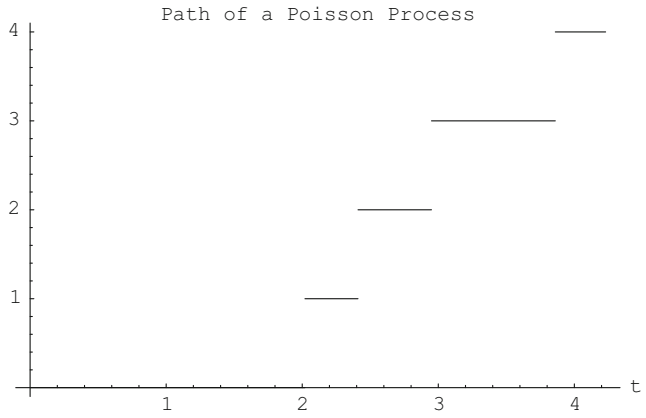


Fig. 8.9 Path of a Poisson process

between the arrival times, $Y_1 - Y_0, Y_2 - Y_1, Y_3 - Y_2, \ldots$, are called the *interarrival times*. Writing $Y_n - Y_{n-1} = T_n$, we see that the interarrival times and the arrival times are related by the simple identity

$$Y_n = (Y_n - Y_{n-1}) + (Y_{n-1} - Y_{n-2}) + \cdots + (Y_2 - Y_1) + (Y_1 - Y_0) = T_1 + T_2 + \cdots + T_n.$$

A special property of a Poisson process is that these interarrival times are iid exponential. So, for instance, if T_3, the time that you had to wait between the second and the third events, was large, then you have no right to believe that T_4 should be small because T_3 and T_4 are actually independent for a Poisson process.

Definition 8.5. Let $T_1, T_2, \cdots$ be an infinite sequence of iid exponential random variables with a common mean λ. For $t \geq 0$, define $X(t)$ by the relation

$$X(t) = k \Leftrightarrow T_1 + \cdots + T_k \leq t < T_1 + \cdots + T_{k+1}; X(0) = 0.$$

Then the family of random variables $\{X(t), t \geq 0\}$ is called a *stationary or homogeneous Poisson process* with constant arrival rate λ.

We will state, without proof, the most important property of a homogeneous Poisson process.

Theorem 8.9. *Let $\{X(t), t \geq 0\}$ be a homogeneous Poisson process with constant arrival rate λ. Then,*

(a) *Given any $0 \leq t_1 \leq t_2 < \infty$, $X(t_2) - X(t_1) \sim Poi(\lambda(t_2 - t_1))$.*
(b) *Given any $n \geq 2$ and disjoint time intervals $[a_i, b_i], i = 1, 2, \ldots, n$, the random variables $X(b_i) - X(a_i), i = 1, 2, \ldots, n$ are mutually independent.*

Property (b) in the theorem is called the independent increments property. *Independent increments simply mean that the number of events over nonoverlapping time intervals are mutually independent.*

Example 8.18 (A Medical Example). Suppose between the months of May and October you catch allergic rhinitis at the constant average rate of once in six weeks. Assuming that the incidences follow a Poisson process, let us answer some simple questions.

First, what is the expected total number of times that you will catch allergic rhinitis between May and October in one year? Take the start date of May 1 as $t = 0$ and $X(t)$ as the number of fresh incidences up to (and including) time t. Note that time is being measured in some implicit unit, say weeks. Then, the arrival rate of the Poisson process for $X(t)$ is $\lambda = \frac{1}{6}$. There are 24 weeks between May and October, and $X(24)$ is distributed as Poisson with mean $24\lambda = 4$, which is the expected total number of times that you will catch allergic rhinitis between May and October. Next, what is the probability that you will catch allergic rhinitis at least once before the start of August and at least once after the start of August and before the end of October? This is the same as asking what $P(X(12) \geq 1, X(24) - X(12) \geq 1)$ is. By the property of independence of $X(12)$ and $X(24) - X(12)$, this probability equals

$$P(X(12) \geq 1)P(X(24) - X(12) \geq 1) = [P(X(12) \geq 1)]^2$$
$$= [1 - P(X(12) = 0)]^2 = \left[1 - e^{-\frac{12}{6}}\right]^2 = .7476.$$

8.7 Synopsis

(a) If $X \sim U[a, b]$, then

$$f(x) = \frac{1}{b-a} I_{\{a \leq x \leq b\}}; E(X) = \frac{a+b}{2}; \text{Var}(X) = \frac{(b-a)^2}{12}.$$

(b) If $X \sim Be(\alpha, \beta)$, then

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{\{0 \leq x \leq 1\}}; E(X) = \frac{\alpha}{\alpha+\beta};$$

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

(c) If $X \sim Exp(\lambda)$, then

$$f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, x \geq 0; E(X) = \lambda; \text{Var}(X) = \lambda^2.$$

The median of X equals $\lambda \log 2$.

(d) If $X \sim Gamma(\alpha, \lambda)$, then

$$f(x) = \frac{e^{-x/\lambda} x^{\alpha-1}}{\lambda^\alpha \Gamma(\alpha)}, x > 0; E(X) = \alpha\lambda; \text{Var}(X) = \alpha\lambda^2.$$

(e) If $X_1, \ldots, X_n$ are iid $Exp(\lambda)$, then $X_1 + \cdots + X_n \sim Gamma(n, \lambda)$.

(f) The mgf of a Gamma distribution equals $\psi(t) = (1 - \lambda t)^{-\alpha}, t < \frac{1}{\lambda}$.

(g) Any exponential density satisfies the lack of memory property

$$P(X > s + t \,|\, X > s) = P(X > t)$$

for all $s, t > 0$.

(h) The Gamma density with parameters $\alpha = \frac{m}{2}, \lambda = 2$ is called the chi-square density with m degrees of freedom. The mean and the variance of a chi-square distribution with m degrees of freedom are m and $2m$.

(i) If $X \sim Pa(\alpha, \theta)$, then

$$f(x) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}}, x \geq \theta > 0; E(X) = \frac{\alpha\theta}{\alpha-1}, \text{ if } \alpha > 1;$$

$$\text{Var}(X) = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)}, \text{ if } \alpha > 2.$$

(j) If X has the Gumbel density with parameters μ, σ, then

$$f(x) = \frac{1}{\sigma} e^{\left(-e^{-\frac{x-\mu}{\sigma}}\right)} e^{-\frac{x-\mu}{\sigma}}, -\infty < x < \infty;$$

$$E(X) = \mu - \gamma\sigma; \text{Var}(X) = \frac{\pi^2}{6}\sigma^2;$$

here, γ is the Euler constant (see the text).

(k) If $\{X(t), t \geq 0\}$ is a homogeneous Poisson process with constant arrival rate λ, then, for any $0 \leq t_1 \leq t_2 < \infty$, $X(t_2) - X(t_1) \sim Poi(\lambda(t_2 - t_1))$. Additionally, given any n disjoint time intervals $[a_i, b_i], 1 \leq i \leq n$, the random variables $X(b_i) - X(a_i), 1 \leq i \leq n$ are mutually independent.

(l) If $\{X(t), t \geq 0\}$ is a homogeneous Poisson process with constant arrival rate λ, then the interarrival times $T_1, T_2, \cdots$ are iid exponential with mean λ.

8.8 Exercises

Exercise 8.1. Suppose $X \sim U[-2, 2]$. For what a, b is $a + bX \sim U[0, 1]$?

Exercise 8.2. X is uniformly distributed on some interval $[a, b]$. If its mean is 2 and variance is 3, what are the values of a and b?

Exercise 8.3. A city bus is supposed to arrive at a fixed stop at 12:00 noon, but its arrival time is uniformly distributed between 11:57 AM and 12:04 PM. If it has not yet arrived at 12:01 PM, what is the probability that it will arrive by 12:02 PM?

Exercise 8.4. Let $f(x) = ax^2 + bx + c, 0 < x < 1$, and zero otherwise. For what values of a, b, c is $f(x)$ a density function? For what values of a, b, c is $f(x)$ a density function with mean 0.5?

Exercise 8.5. Let $X \sim U[0, 1]$. Find the density of each of the following:

(a) $X^3 - 3X$;

(b) $\left(X - \frac{1}{2}\right)^2$;

(c) $\left(\sin\left(\frac{\pi}{2}X\right)\right)^4$.

Exercise 8.6. *(**Using the Quantile Transformation to Simulate a Variable).** Let $U \sim U[0, 1]$. Describe how you will use U to simulate X if X has the following densities:

(a) $f(x) = 5x^4, 0 < x < 1$;

(b) $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}, 0 < x < 1$;

(c) $f(x) = \frac{1}{2}e^{-|x-2|}, -\infty < x < \infty$.

Exercise 8.7. It is known that, for a Beta random variable X with parameters α and β, $P(X < .2) = .22$. If $\alpha + \beta = 26$, find α and β.

Exercise 8.8 (Triangular Density). Suppose X has the triangular density on $[0, 1]$. For what values of a and b does $a + bX$ have a triangular density on $[-1, 1]$? Here, a triangular density on $[-1, 1]$ is the symmetric triangular function centered at zero.

Exercise 8.9 (Moments of a Triangular Density). Suppose X has a triangular density on $[0, 1]$. Find a formula for the nth moment of X.

Exercise 8.10. *(**Mode of a Beta Density).** Show that if a Beta density has a mode in the open interval $(0, 1)$, then we must have $\alpha > 1, \alpha + \beta > 2$, in which case the mode is unique and equals $\frac{\alpha-1}{\alpha+\beta-2}$.

Exercise 8.11. *(**Mean Absolute Deviation of Beta).** Suppose $X \sim Be(m, n)$, where m, n are positive integers. Derive a formula for the mean absolute deviation of X.

Exercise 8.12. The concentration of acetic acid in table vinegar has a Beta distribution with mean 0.083 and standard deviation .077. In what percentage of bottles of vinegar does the acetic acid concentration exceed 20%?

Exercise 8.13 (Subexponential Density). Find the constant c such that $f(x) = ce^{-\sqrt{x}}$ is a pdf on $(0, \infty)$.

Exercise 8.14. An exponential random variable with mean 4 is known to be larger than 6. What is the probability that it is larger than 8?

Exercise 8.15. * **(A Two-Layered Problem).** The time that you have to wait to speak to a customer service representative when you call a bank is exponentially distributed with mean 1.5 minutes. If you make ten calls in one month (and never hang up), what is the probability that at least twice you will have to wait more than three minutes?

Exercise 8.16 (Truncated Exponential). Suppose $X \sim Exp(1)$. What is the density of $2X + 1$?

Exercise 8.17. * Suppose $X_1, X_2, \cdots, X_n$ are independent $Exp(1)$ variables and $a, b, b > 0$ are constants. What is the density of $a + b \sum_{i=1}^{n} X_i$?

Exercise 8.18 (The Jovial Professor). The number of jokes your professor tells in class per t minutes has a Poisson distribution with mean $0.1t$. If the class started at 12:00 noon, what is the probability that the first joke will be told before 12:20 PM?

Exercise 8.19. *(**Sum of Gammas).** Suppose X and Y are independent random variables and $X \sim G(\alpha, \lambda), Y \sim G(\beta, \lambda)$. Find the distribution of $X + Y$ using moment generating functions.

Exercise 8.20 (Inverse Gamma Moments). Suppose $X \sim G(\alpha, \lambda)$. Find a formula for $E[(\frac{1}{X})^n]$ when this expectation exists.

Exercise 8.21 (Product of Chi-Squares). Suppose $X_1, X_2, \ldots, X_n$ are independent chi-square variables with $X_i \sim \chi^2_{m_i}$. Find the mean and variance of $\prod_{i=1}^n X_i$.

Exercise 8.22. ***(Chi-Square Skewness).** Let $X \sim \chi^2_m$. Find the coefficient of skewness of X and prove that it converges to zero as $m \to \infty$.

Exercise 8.23. ***(A Half-Life Problem).** A piece of rock contains 10^{25} atoms. Each atom has an exponentially distributed lifetime with a half-life of one century; here, half-life means the distribution's median. How many centuries must pass before there is just about a 50% chance that at least one atom still remains?

Exercise 8.24. Let $X \sim Exp(\lambda)$. Find a formula for $P(X > 2\lambda)$. What is special about the formula?

Exercise 8.25. ***(An Optimization Problem).** Suppose that a battery has a lifetime with a general density $f(x), x > 0$. A generator using this battery costs $\$c_1$ per hour to run, and while it runs, a profit of $\$c_2$ is made. Suppose also that the labor charge per hour to operate the generator is $\$c_3$.

(a) Find the expected profit if labor is hired for t hours.
(b) Is there an optimum value of t? How will you characterize it?
(c) What is such an optimum value if $f(x)$ is an exponential density with mean λ?

Exercise 8.26. ***(A Relation Between Poisson and Gamma).** Suppose $X \sim Poi(\lambda)$. Prove by repeated integration by parts that

$$P(X \leq n) = P(G(n+1, 1) > \lambda),$$

where $G(n+1, 1)$ means a Gamma random variable with parameters $n+1$ and 1.

Exercise 8.27. ***(A Relation Between Binomial and Beta).** Suppose $X \sim Bin(n, p)$. Prove that

$$P(X \leq k-1) = P(B(k, n-k+1) > p),$$

where $B(k, n-k+1)$ means a Beta random variable with parameters $k, n-k+1$.

Exercise 8.28. Suppose X has the standard Gumbel density. Find the density of e^{-X}.

Exercise 8.29. Suppose X is uniformly distributed on $[0, 1]$. Find the density of $\log\log\frac{1}{X}$.

Exercise 8.30. Suppose X has a Pareto distribution with parameters α and θ. Find the distribution of $\frac{\theta}{X}$.

Exercise 8.31 (Poisson Process for Catching a Cold). Suppose that you catch a cold according to a Poisson process once every three months.

(a) Find the probability that between the months of July and October, you will catch at least four colds.
(b) Find the probability that between the months of May and July, and also between the months of July and October, you will catch at least four colds.
(c) * Find the probability that you will catch more colds between the months of July and October than between the months of May and July.

Exercise 8.32 (Correlation in a Poisson Process). Suppose $X(t)$ is a Poisson process with average constant arrival rate λ. Let $0 < s < t < \infty$. Find the correlation between $X(s)$ and $X(t)$.

Exercise 8.33 (Two Poisson Processes). Suppose $X(t)$ and $Y(t)$, with $t \geq 0$ are two Poisson processes with rates λ_1 and λ_2. Assume that the processes run independently.

(a) Prove or disprove: $X(t) + Y(t)$ is also a Poisson process.
(b) * Prove or disprove: $|X(t) - Y(t)|$ is also a Poisson process.

Exercise 8.34 (Connection of a Poisson Process to Binomial Distribution). Suppose $X(t)$ with $t \geq 0$ is a Poisson process with constant average rate λ. Given that $X(t) = n$, show that the number of events up to the time u, where $u < t$, has a binomial distribution. Identify the parameters of this binomial distribution.

Exercise 8.35 (Use Your Computer). Use the quantile transformation method to simulate 100 values from a distribution with density $4x^3$ on $[0, 1]$. Repeat the simulation 500 times. For each such set of 100 values, compute the mean. Do these 500 means cluster around some number? Would you expect that?

Exercise 8.36 (Use Your Computer). Use the quantile transformation method to simulate 100 values from a Gamma distribution with parameters 20 and 1. Repeat the simulation 500 times. How can you use these simulated sets to approximate the median of a Gamma distribution with parameters 20 and 1?

Exercise 8.37 (Use Your Computer). Design a simulation exercise to approximate the value of $E(X^X)$ when X has a $U[0, 1]$ distribution.

References

Everitt, B. (1998). *Cambridge Dictionary of Statistics*, Cambridge University Press, New York.
Johnson, N., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. I, Wiley, New York.
Kendall, M. and Stuart, A. (1976). *Advanced Theory of Statistics, Vol. I*, Fourth ed., Macmillan, New York.

Chapter 9
Normal Distribution

Empirical data on many types of variables across disciplines tend to exhibit unimodality and only a small amount of skewness. It is quite common to use a normal distribution as a model for such data. The normal distribution occupies the central place among all distributions in probability and statistics. When a new methodology is presented, it is usually first tested on the normal distribution. The most well-known procedures in the toolbox of a statistician have their exact inferential optimality properties when sample values come from a normal distribution. There is also *the central limit theorem*, which says that the sum of many small independent quantities approximately follows a normal distribution. Theoreticians sometimes think that empirical data are often approximately normal, while empiricists think that theory shows that many types of variables are approximately normally distributed. By a combination of reputation, convenience, mathematical justification, empirical experience, and habit, the normal distribution has become the most ubiquitous of all distributions. It is also the most studied; we know more theoretical properties of the normal distribution than of others. It satisfies intriguing and elegant characterizing properties not satisfied by any other distribution. Because of its clearly unique position and its continuing importance in every emerging problem, we discuss the normal distribution exclusively in this chapter.

Stigler (1975, 1986) gives authoritative accounts of the history of the normal distribution. Galton, de Moivre, Gauss, Quetelet, Laplace, Karl Pearson, Edgeworth, and of course Ronald Fisher all contributed to the popularization of the normal distribution. Detailed algebraic properties can be seen in Johnson et al. (1994), Rao (1973), Kendall and Stuart (1976), and Feller (1971). Patel and Read (1996) is a good source for other references. Petrov (1975), Tong (1990), Bryc (1995), and Freedman (2005) are important recent references; of these, Petrov (1975) is a masterly account of the role of the normal distribution in the limit theorems of probability.

9.1 Definition and Basic Properties

We have actually already defined a normal density in Chapter 7. We recall the definition here.

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_9,

Definition 9.1. A random variable X is said to have a normal distribution with parameters μ and σ^2 if it has the density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty,$$

where μ can be any real number, and $\sigma > 0$. We write $X \sim N(\mu, \sigma^2)$. If $X \sim N(0, 1)$, we call it a *standard normal variable*.

The density of a standard normal variable is denoted as $\phi(x)$ and equals the function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty,$$

and the CDF is denoted as $\Phi(x)$. Note that the standard normal density is symmetric and unimodal about zero. The general $N(\mu, \sigma^2)$ density is symmetric and unimodal about μ.

By the definition of a CDF,

$$\Phi(x) = \int_{-\infty}^{x} \phi(z) dz.$$

The CDF $\Phi(x)$ cannot be written in terms of the elementary functions but can be computed at a given value x, and tables of the values of $\Phi(x)$ are widely available. For example, here are some selected values.

Example 9.1 (Standard Normal CDF at Selected Values).

x	$\Phi(x)$
-4	.00003
-3	.00135
-2	.02275
-1	.15866
0	.5
1	.84134
2	.97725
3	.99865
4	.99997

By inspection, we find in this table that $\Phi(-x) + \Phi(x)$ is always one. This is a mathematical fact and a consequence of the symmetry of the standard normal distribution around zero:

$$\Phi(-x) = 1 - \Phi(x) \; \forall \; x.$$

If we keep σ^2 fixed and change μ, a normal distribution only gets shifted to a new center. If we keep μ fixed and increase σ^2, the distribution becomes more spread out. In fact, we will shortly see that σ^2 is the variance of an $N(\mu, \sigma^2)$ distribution. Figure 9.1 helps visualize these facts about normal distributions.

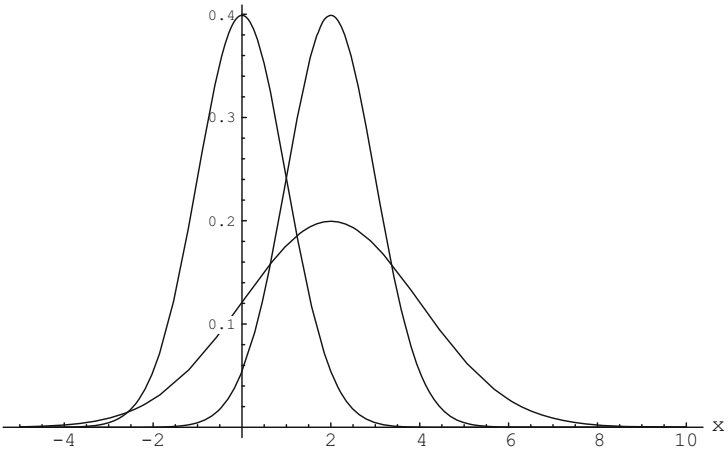


Fig. 9.1 $N(0,1)$, $N(2,1)$, and $N(2,4)$ Densities

Theorem 9.1 states the most basic properties of a normal distribution.

Theorem 9.1.

(a) *If* $X \sim N(\mu, \sigma^2)$, *then* $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$, *and if* $Z \sim N(0,1)$, *then* $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.
In words, if X *is any normal random variable, then its standardized version is always a standard normal variable.*

(b) *If* $X \sim N(\mu, \sigma^2)$, *then*

$$P(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \ \forall\, x.$$

In particular, $P(X \le \mu) = P(Z \le 0) = .5$*; i.e., the median of* X *is* μ.

(c) *Every moment of any normal distribution exists, and the odd central moments* $E[(X-\mu)^{2k+1}]$ *are all zero.*

(d) *If* $Z \sim N(0,1)$, *then*

$$E(Z^{2k}) = \frac{(2k)!}{2^k k!}, k \ge 1.$$

(e) *The mgf of the* $N(\mu, \sigma^2)$ *distribution exists at all real* t *and equals*

$$\psi(t) = e^{t\mu + \frac{t^2\sigma^2}{2}}.$$

(f) *If* $X \sim N(\mu, \sigma^2)$,

$$E(X) = \mu; \mathrm{Var}(X) = \sigma^2; E(X^3) = \mu^3 + 3\mu\sigma^2; E(X^4) = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4.$$

(g) *If* $X \sim N(\mu, \sigma^2)$, *then* $\kappa_1 = \mu, \kappa_2 = \sigma^2$, *and* $\kappa_r = 0\ \forall\, r > 2$, *where* κ_j *is the* j*th cumulant of* X.

Proof. Part (a) follows from the general fact that if Z has density $f(z)$, then $X = a + bZ$ has density $\frac{1}{|b|} f(\frac{x-a}{b})$; we simply identify a with μ and b with σ.

For part (b), denoting a standard normal variable by Z, observe that

$$P(X \le x) = P(X - \mu \le x - \mu) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right)$$
$$= P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Part (c) follows from the fact that every moment of a standard normal distribution exists and that all its odd moments $E(Z^{2k+1}) = 0$, which we have already proved in Chapter 7. Likewise, part (d) also has already been proved in Chapter 7.

For part (e), if $X \sim N(\mu, \sigma^2)$, represent X as $X = \mu + \sigma Z$, where $Z \sim N(0, 1)$, and observe that

$$E\left(e^{tX}\right) = E\left(e^{t(\mu+\sigma Z)}\right) = e^{t\mu} E\left(e^{t\sigma Z}\right)$$
$$= e^{t\mu} e^{t^2\sigma^2/2} = e^{t\mu + t^2\sigma^2/2},$$

where we have used the formula $E(e^{sZ}) = e^{s^2/2}$, derived in Chapter 7.

For part (f),

$$E(X) = E(\mu + \sigma Z) = \mu + \sigma E(Z) = \mu$$

since $E(Z) = 0$. Likewise, $\mathrm{Var}(X) = \mathrm{Var}(\mu + \sigma Z) = \mathrm{Var}(\sigma Z) = \sigma^2 \mathrm{Var}(Z) = \sigma^2$. For the third moment, $E(X^3) = E[(\mu + \sigma Z)^3] = E[\mu^3 + 3\mu^2\sigma Z + 3\mu\sigma^2 Z^2 + \sigma^3 Z^3] = \mu^3 + 3\mu\sigma^2$ since $E(Z)$, and $E(Z^3)$ are both zero. The fourth-moment formula follows similarly on using $E(Z^4) = 3$. Finally, for part (g), the first two cumulants are always the mean and the variance of the distribution, and the higher-order cumulants all vanish because $\log \psi(t) = t\mu + t^2\sigma^2/2$, which is a quadratic in t, so its third and higher derivatives are identically equal to zero.

An important consequence of part (b) of this theorem is the following result.

Corollary. Let $X \sim N(\mu, \sigma^2)$ and let $0 < \alpha < 1$. Let $Z \sim N(0, 1)$. Suppose x_α is the $(1 - \alpha)$th quantile (also called percentile) of X and z_α is the $(1 - \alpha)$th quantile of Z. Then

$$x_\alpha = \mu + \sigma z_\alpha.$$

Remark. Part (b) of the theorem and this corollary together say that to compute the value of the CDF of any normal distribution at a point, or to compute any percentile of a general normal distribution, we only need to know how to compute the CDF of a standard normal distribution and percentiles of a standard normal distribution. In other words, only one table of CDF values is needed to compute the CDF and percentiles of arbitrary normal distributions; we can reduce the arbitrary normal case problem to a standard normal problem. This is a very important practical point.

9.2 Working with a Normal Table

It is essential that we learn how to correctly use a standard normal table. We may want to consult a standard normal table for evaluating probabilities or finding the value of some percentile. A detailed standard normal table is provided in the appendix. Here are some examples that reinforce the results we have described above.

Example 9.2 (Selected Standard Normal Percentiles). By using a *standard normal CDF table*, one can see that the 75th, 90th, 95th, 97.5th, 99th, and 99.5th percentiles of a standard normal distribution are the following:

α	$1-\alpha$	z_α
.25	.75	.675
.1	.9	1.282
.05	.95	1.645
.025	.975	1.960
.01	.99	2.326
.005	.995	2.576

Example 9.3. The age of the subscribers to a newspaper has a normal distribution with mean 50 years and standard deviation 5 years. We want to compute the percentage of subscribers who are less than 40 years old and the percentage who are between 40 and 60 years old.

Let X denote the age of a subscriber; we have $X \sim N(\mu, \sigma^2), \mu = 50, \sigma = 5$. Therefore,

$$P(X < 40) = \Phi\left(\frac{40-50}{5}\right) = \Phi(-2) = .02275$$

and

$$\begin{aligned} P(40 \le X \le 60) &= P(X \le 60) - P(X \le 40) = \Phi(2) - \Phi(-2) \\ &= (1 - .02275) - .02275 = 1 - 2 \times .02275 = .9545. \end{aligned}$$

Example 9.4 (Using a Standard Normal Table). Let $Z \sim N(0,1)$; we will find the values of $P(|Z-1| < 2)$, $P(Z^2 \le 9)$, and $P(\frac{Z}{1+Z^2} < \frac{1}{2})$.

$$P(|Z-1| < 2) = P(-1 < Z < 3) = \Phi(3) - \Phi(-1) = .99865 - .15866 = .84.$$

$$P(Z^2 \le 9) = P(-3 \le Z \le 3) = \Phi(3) - \Phi(-3) = .99865 - .00135 = .9973.$$

$$P\left(\frac{Z}{1+Z^2} < \frac{1}{2}\right) = P(1 + Z^2 > 2Z) = P((Z-1)^2 > 0) = 1.$$

Example 9.5 (Using a Standard Normal Table). Suppose $X \sim N(5, 16)$; we want to know which number x has the property that $P(X \le x) = .95$.

This amounts to asking what the 95th percentile of X is. By the general formula for percentiles of a normal distribution, the 95th percentile of X equals

$$\sigma \times \text{ 95th percentile of standard normal } + \mu = 4 \times 1.645 + 5 = 11.58.$$

Now change the question: Which number x has the property that $P(x \leq X \leq 9) = .68$? This means, by standardizing X to a standard normal,

$$\Phi(1) - \Phi\left(\frac{x-5}{4}\right) = .68 \Rightarrow \Phi\left(\frac{x-5}{4}\right) = \Phi(1) - .68 = .8413 - .68 = .1613.$$

By reading a standard normal table, $\Phi(-.99) = .1613$, so

$$\frac{x-5}{4} = -.99 \Rightarrow x = 1.04.$$

9.3 Additional Examples and the Lognormal Density

Example 9.6 (A Reliability Problem). Let X denote the length of time (in minutes) an automobile battery will continue to crank an engine. Assume that $X \sim N(10, 4)$. What is the probability that the battery will crank the engine longer than $10 + x$ minutes given that it is still cranking at 10 minutes?

We want to find

$$P(X > 10 + x \,|\, X > 10) = \frac{P(X > 10 + x)}{P(X > 10)} = \frac{P(Z > x/2)}{1/2} = 2\left[1 - \Phi\left(\frac{x}{2}\right)\right].$$

Note that this is decreasing in x. If X had been exponentially distributed, then by the lack of memory property, this probability would have been $P(X > x) = e^{-x/10}$, assuming that the mean was still 10 minutes. But if the distribution is normal, we can no longer get an analytic expression for the probability; we only get an expression involving the standard normal CDF.

As a specific choice, if $x = 2$, then we get

$$P(X > 10 + x \,|\, X > 10) = 2\left[1 - \Phi\left(\frac{x}{2}\right)\right] = 2[1 - \Phi(1)] = .3174.$$

Example 9.7 (Setting a Thermostat). Suppose that when the thermostat is set at d degrees Celsius, the actual temperature of a certain room is a normal random variable with parameters $\mu = d$ and $\sigma = .5$.

If the thermostat is set at 75° C, what is the probability that the actual temperature of the room will be below 74° C?

By standardizing to an $N(0, 1)$ random variable,

$$P(X < 74) = P(Z < (74 - 75)/.5) = P(Z < -2) = .02275.$$

Next, what is the lowest setting of the thermostat that will maintain a temperature of at least 72° C with a probability of .99?

We want to find the value of d that makes $P(X \geq 72) = .99 \Rightarrow P(X < 72) = .01$. Now, from a standard normal table, $P(Z < -2.326) = .01$. Therefore, we want to find d, which makes $d + \sigma \times (-2.326) = 72 \Rightarrow d - .5 \times 2.326 = 72 \Rightarrow d = 72 + .5 \times 2.326 = 73.16°\ C$.

Example 9.8 (A Two-Layered Example). Suppose the distribution of heights in a population is approximately normal. Ten percent of individuals are over 6 feet tall, and the average height is 5 ft. 10 in. What is the approximate probability that in a group of 50 people picked at random there will be two or more people who are over 6 ft. 1 in. tall?

Denoting height as X, we have $X \sim N(\mu, \sigma^2), \mu = 70$, and $P(X > 72) = .1$; i.e., 72 is the 90th percentile of $X \Rightarrow 72 = 70 + 1.282\sigma \Rightarrow \sigma = 1.56$.

So, the probability that one individual is taller than 6 ft. 1 in. is $p = P(X > 73) = P(Z > (73 - 70)/1.56) = P(Z > 1.92) = .0274$.

Therefore, T, the number of people among 50 who are taller than 6 ft. 1 in. is distributed as $Bin(50, p)$ and we want

$$\begin{aligned} P(T \geq 2) &= 1 - P(T = 0) - P(T = 1) = 1 - (1 - .0274)^{50} \\ &\quad -50 \times .0274 \times (1 - .0274)^{49} = 1 - .6004 = .3996. \end{aligned}$$

Example 9.9 (Rounding a Normal Variable). Suppose $X \sim N(0, \sigma^2)$ and that the absolute value of X is rounded to the nearest integer. We have seen in Chapter 7 that the expected value of $|X|$ itself is $\sigma\sqrt{2/\pi}$. How does rounding affect the expected value?

Denote the rounded value of $|X|$ by Y. Then, $Y = 0 \Leftrightarrow |X| < .5; Y = 1 \Leftrightarrow .5 < |X| < 1.5; \cdots$, etc. Therefore,

$$\begin{aligned} E(Y) &= \sum_{i=1}^{\infty} iP(i - 1/2 < |X| < i + 1/2) = \sum_{i=1}^{\infty} iP(i - 1/2 < X < i + 1/2) \\ &\quad + \sum_{i=1}^{\infty} iP(-i - 1/2 < X < -i + 1/2) \\ &= 2\sum_{i=1}^{\infty} i\,[\Phi((i+1/2)/\sigma) - \Phi((i - 1/2)/\sigma)] = 2\sum_{i=1}^{\infty}[1 - \Phi((i + 1/2)/\sigma)] \end{aligned}$$

on some manipulation.

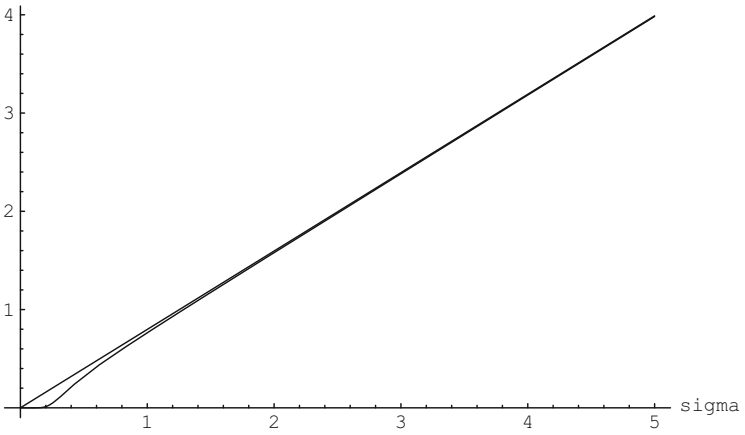


Fig. 9.2 Expected value of rounded and unrounded $|X|$ when x is $N(0, \text{sigma}\hat{}2)$

For example, if $\sigma = 1$, then this equals $2\sum_{i=1}^{\infty}[1-\Phi(i+1/2)] = .76358$, while the unrounded $|X|$ has the expectation $\sqrt{2/\pi} = .79789$. The effect of rounding is not serious when $\sigma = 1$.

A plot of the expected value of Y and the expected value of $|X|$ is shown in Figure 9.2 to study the effect of rounding.

We can see that the effect of rounding is uniformly small. There is classic literature on corrections needed in computing means, variances, and higher moments when data are rounded. These are known as *Sheppard's corrections*. Kendall and Stuart (1976) gives a thorough treatment of these necessary corrections.

Example 9.10 (Lognormal Distribution). Lognormal distributions are common models in studies of economic variables, such as income and wealth, because they can adequately describe the skewness that one sees in data on such variables. If $X \sim N(\mu, \sigma^2)$, then the distribution of $Y = e^X$ is called a *lognormal distribution with parameters* μ, σ^2. Note that the lognormal name can be confusing; a lognormal variable is *not* the logarithm of a normal variable. A better way to remember its meaning is *log is normal*.

Since $Y = e^X$ is a strictly monotone function of X, by the usual formula for the density of a monotone function, Y has the pdf

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\log y - \mu)^2}{2\sigma^2}}, y > 0;$$

this is called the lognormal density with parameters μ, σ^2. Since a lognormal variable is defined as e^X for a normal variable X, its mean and variance are easily found from the mgf of a normal variable. A simple calculation shows that

$$E(Y) = e^{\mu + \frac{\sigma^2}{2}}; \text{Var}(Y) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}.$$

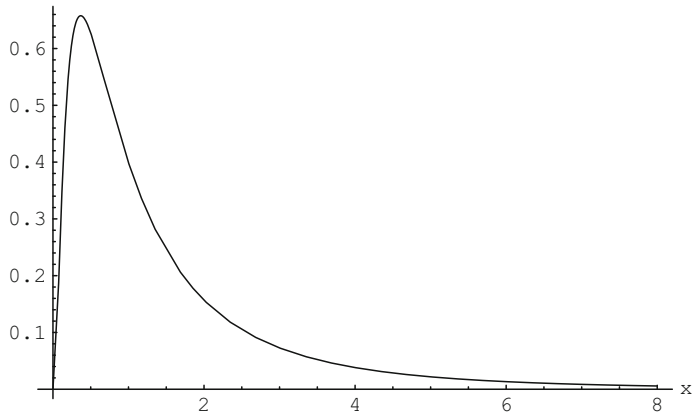


Fig. 9.3 Lognormal (0,1) density

One of the main reasons for the popularity of the lognormal distribution is its skewness; the lognormal density is extremely skewed for large values of σ. The coefficient of skewness has the formula

$$\beta = (2 + e^{\sigma^2})\sqrt{e^{\sigma^2} - 1},$$

$\to \infty$, as $\sigma \to \infty$. A plot of the lognormal density for $\mu = 0, \sigma = 1$ is shown in Figure 9.3 to illustrate the skewness. Note that the lognormal densities do not have a finite mgf at any $t > 0$, although all their moments are finite. It is also the only standard continuous distribution that is *not determined by its moments* (see Heyde (1963)). That is, there exist *other distributions besides the lognormal* whose moments exactly coincide with the moments of a given lognormal distribution. This is not true of any other distribution with a name that we have come across in this text. For example, the normal and Poisson distributions are determined by their moments.

9.4 Sums of Independent Normal Variables

We had remarked in the chapter introduction that sums of many independent variables tend to be approximately normally distributed. A precise version of this is the *central limit theorem*, which we will study in a later chapter. What is interesting is that sums of *any number* of independent normal variables are *exactly* normally distributed. Here is the result.

Theorem 9.2. *Let $X_1, X_2, \ldots, X_n, n \geq 2$ be independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$. Let $S_n = \sum_{i=1}^n X_i$. Then,*

$$S_n \sim N\left(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2\right).$$

Proof. The quickest proof of this uses the mgf technique. Since the X_i are independent, the mgf of S_n is

$$\begin{aligned}\psi_{S_n}(t) &= E(e^{tS_n}) = E(e^{tX_1}\cdots e^{tX_n})\\ &= \prod_{i=1}^{n} E(e^{tX_i}) = \prod_{i=1}^{n} e^{t\mu_i + t^2\sigma_i^2/2} = e^{t(\sum_{i=1}^{n}\mu_i) + (t^2/2)(\sum_{i=1}^{n}\sigma_i^2)},\end{aligned}$$

which agrees with the mgf of the $N(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2)$ distribution, and therefore by the distribution determining property of mgfs, it follows that $S_n \sim N(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2)$.

An important consequence is the following result.

Corollary. Suppose $X_i, 1 \le i \le n$ are independent and each is distributed as $N(\mu, \sigma^2)$. Then $\overline{X} = \frac{S_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$.

To prove it, simply note that, by the theorem, $S_n \sim N(n\mu, n\sigma^2)$, and therefore $\frac{S_n}{n} \sim N(\mu, \frac{n\sigma^2}{n^2}) = N(\mu, \frac{\sigma^2}{n})$.

This says that the distribution of $\overline{X}$ gets more concentrated around μ as n increases because the variance $\frac{\sigma^2}{n}$ decreases with n. When n gets very large, the normal distribution of $\overline{X}$ will get very spiky around μ. One way to think of it is that, for large n, the mean of the sample values, namely $\overline{X}$, will be very close to the mean of the distribution, namely μ; *$\overline{X}$ will be a good estimate of μ when n is large*.

Remark. The theorem above implies that any linear function of independent normal variables is also normal; i.e.,

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

Example 9.11. Suppose $X \sim N(-1, 4), Y \sim N(1, 5)$, and suppose that $X$and Y are independent. We want to find the CDF$_s$ of $X + Y$ and $X - Y$.

By the theorem above,

$$X + Y \sim N(0, 9) \text{ and } X - Y \sim N(-2, 9).$$

Therefore,

$$P(X + Y \le x) = \Phi\left(\frac{x}{3}\right) \text{ and } P(X - Y \le x) = \Phi\left(\frac{x+2}{3}\right).$$

For example,

$$P(X+Y \le 3) = \Phi(1) = .8413 \text{ and } P(X-Y \le 3) = \Phi\left(\frac{5}{3}\right) = .9525.$$

Example 9.12 (Confidence Interval and Margin of Error). Suppose some random variable $X \sim N(\mu, \sigma^2)$ and we have n independent observations $X_1, X_2, \ldots, X_n$ on this variable X; another way to put it is that $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. Therefore, $\overline{X} \sim N(\mu, \sigma^2/n)$, and we have

$$P(\overline{X} - 1.96\sigma/\sqrt{n} \le \mu \le \overline{X} + 1.96\sigma/\sqrt{n}) = P(-1.96\sigma/\sqrt{n} \le \overline{X} - \mu \le 1.96\sigma/\sqrt{n})$$

$$= P\left(-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) = \Phi(1.96) - \Phi(-1.96) = .95$$

from a standard normal table.

Thus, with a 95% probability, for any n, μ is between $\overline{X} \pm 1.96\sigma/\sqrt{n}$. Statisticians call the interval of values $\overline{X} \pm 1.96\sigma/\sqrt{n}$ *a 95% confidence interval for* μ, *with a margin of error* $1.96\sigma/\sqrt{n}$.

A tight confidence interval will correspond to a small margin of error. For example, if we want a margin of error $\le .1$, then we will need $1.96\sigma/\sqrt{n} \le .1 \Leftrightarrow \sqrt{n} \ge 19.6\sigma \Leftrightarrow n \ge 384.16\sigma^2$. Statisticians call such a calculation a *sample size calculation*.

9.5 Mills Ratio and Approximations for the Standard Normal CDF

The standard normal CDF cannot be represented in terms of the elementary functions. But it arises in many mathematical problems having to do with normal distributions. As such, it is important to know the behavior of the standard normal CDF $\Phi(x)$, especially for large x. We have seen in Chapter 7 that the Chernoff-Bernstein inequality implies that $1 - \Phi(x) \le e^{-x^2/2}$ for all positive x. However, actually we can prove better inequalities. The ratio

$$R(x) = \frac{1-\Phi(x)}{\phi(x)}$$

is called the *Mills ratio*. We will provide a selection of bounds and asymptotic expansions for $R(x)$ in this section.

Theorem 9.3 (Six Inequalities).

(a) ***Polýa Inequality***

$$\Phi(x) < \frac{1}{2}\left[1 + \sqrt{1 - e^{-\frac{2}{\pi}x^2}}\right].$$

(b) ***Chu Inequality***

$$\Phi(x) \geq \frac{1}{2}\left[1 + \sqrt{1 - e^{-\frac{x^2}{2}}}\right].$$

(c) ***Mitrinovic Inequality***
For $x > 0$,

$$\frac{2}{x + \sqrt{x^2 + 4}} < R(x) < \frac{2}{x + \sqrt{x^2 + \frac{8}{\pi}}}.$$

(d) ***Gordon Inequality***
For $x > 0$,

$$\frac{x}{x^2 + 1} \leq R(x) \leq \frac{1}{x}.$$

(e) ***Szarek-Werner Inequality***
For $x > -1$,

$$\frac{2}{x + \sqrt{x^2 + 4}} < R(x) < \frac{4}{3x + \sqrt{x^2 + 8}}.$$

(f) ***Boyd Inequality***
For $x > 0$,

$$R(x) < \frac{\pi}{2x + \sqrt{(\pi - 2)^2 x^2 + 2\pi}}.$$

See Patel and Read (1996) or DasGupta (2008) for these inequalities. A plot of the exact values of $\Phi(x)$*, the Polýa upper bound, and the Chu lower bound is given in Figure 9.4 as verification of the accuracy of the inequalities. The accuracy of the Polýa upper bound is remarkable, as we can see in the plot. The exact CDF* $\Phi(x)$ *and the Polýa upper bound are so close to each other that they look nearly indistinguishable in the plot.*

There are also pointwise asymptotic expansions available for both $\Phi(x)$ *and* $R(x)$*. We will give asymptotic expansions only for* $R(x)$*. These expansions have a long and detailed literature. Laplace was greatly interested in the problem, and a number of approximations and expansions for both* $\Phi(x)$ *and* $R(x)$ *are due to him. He obtained particularly accurate continued-fractions expansions for* $R(x)$ *but we do not present it here, as it would clearly be beyond the scope of this chapter. The expansion below is also due to Laplace. References to various other expansions, particularly those due to Laplace, can be found in Patel and Read (1996).*

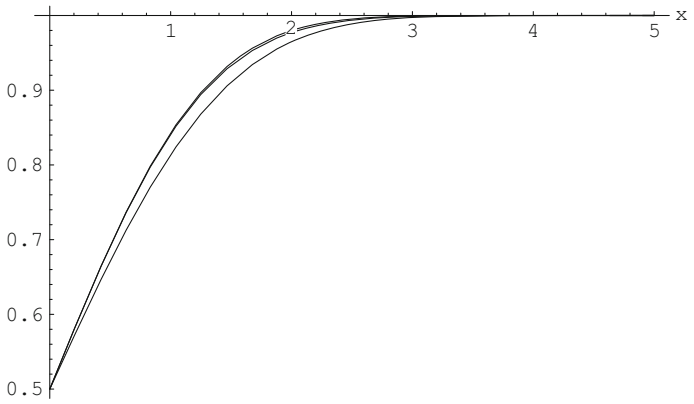


Fig. 9.4 Exact $N(0, 1)$ CDF, Polya upper bound, and Chu lower bound

Theorem 9.4 (Laplace's Expansions for the Mills Ratio).

(a) *For all* $x > 0$,

$$R(x) = \frac{1}{x} - \frac{1}{x^3} + \frac{1 \times 3}{x^5} - \cdots + (-1)^n \frac{1 \times 3 \times \cdots \times (2n-1)}{x^{2n+1}} + R_n(x),$$

where

$$|R_n(x)| < \min\left\{ \frac{1 \times 3 \times \cdots \times (2n-1)}{x^{2n+1}}, \frac{1 \times 3 \times \cdots \times (2n+1)}{x^{2n+3}} \right\}.$$

(b) *For* $x > 0$,

$$\frac{1}{x} - \frac{1}{x^3} < R(x) < \frac{1}{x}.$$

(c) *Let* $\overline{R}(x) = \frac{\Phi(x) - \frac{1}{2}}{\phi(x)}$. Then, for $x > 0$ and all $n \geq 1$,

$$\overline{R}(x) > x + \frac{x^3}{1 \times 3} + \frac{x^5}{1 \times 3 \times 5} + \cdots + \frac{x^{2n-1}}{1 \times 3 \times \cdots \times (2n-1)}.$$

Corollary. $1 - \Phi(x) \approx \frac{\phi(x)}{x}$ as $x \to \infty$.

In spite of this theoretical result, in practice $\frac{\phi(x)}{x}$ does not give a very accurate relative approximation for $1 - \Phi(x)$. For example, if $x = 3$, the exact value of $1 - \Phi(x)$ is .00135, while $\frac{\phi(x)}{x}$ equals .00148; the relative error is almost 10%. It is necessary to use more terms in the expansion to get an accurate approximation. It should be noted that if we end the expansion for $R(x)$ at a term ending in a "minus sign," we always get a lower bound to $R(x)$, while if we end the expansion at a term ending in a "plus sign," we always get an upper bound to $R(x)$.

9.6 Synopsis

(a) If $X \sim N(\mu, \sigma^2)$, then

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty; E(X) = \mu; \text{Var}(X) = \sigma^2.$$

(b) The $N(0, 1)$ density is called the standard normal density and is denoted as $\phi(x)$. The CDF of the standard normal distribution is denoted as $\Phi(x)$. Thus,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}; \Phi(x) = \int_{-\infty}^{x} \phi(z) dz.$$

The CDF $\Phi(x)$ cannot be written in terms of elementary functions but can be computed at a given value x.

(c) The mgf of the $N(\mu, \sigma^2)$ distribution equals

$$\psi(t) = e^{t\mu + \frac{t^2\sigma^2}{2}}, -\infty < t < \infty.$$

(d) If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$. Conversely, if $Z \sim N(0, 1)$, then for any real μ and any $\sigma > 0$, $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.

(e) If $X \sim N(\mu, \sigma^2)$, then $P(X \le x) = \Phi(\frac{x-\mu}{\sigma})\ \forall\, x$. More generally,

$$P(a \le X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

for all $a, b, a \le b$.

(f) If $X \sim N(\mu, \sigma^2)$, then for any $\alpha, 0 < \alpha < 1$, $x_\alpha = \mu + \sigma z_\alpha$, where x_α is the $(1-\alpha)$th quantile of X and z_α is the $(1-\alpha)$th quantile of the standard normal distribution.

(g) The tail probability $1 - \Phi(x)$ in a standard normal distribution converges to zero extremely quickly. Precisely, $1 - \Phi(x) \sim \frac{\phi(x)}{x}$ as $x \to \infty$. This means that $R(x) \sim \frac{1}{x}$ as $x \to \infty$, where $R(x) = \frac{1-\Phi(x)}{\phi(x)}$ is the Mills ratio. More accurate approximations are

$$R(x) \sim \left[\frac{1}{x} - \frac{1}{x^3}\right],$$

$$R(x) \sim \left[\frac{1}{x} - \frac{1}{x^3} + \frac{1 \times 3}{x^5}\right].$$

(h) If $X \sim N(\mu, \sigma^2)$, then the distribution of $Y = e^X$ is called a lognormal distribution with parameters μ, σ. It has the density, mean, and variance given by

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\log y - \mu)^2}{2\sigma^2}}, y > 0;$$

$$E(Y) = e^{\mu + \frac{\sigma^2}{2}}; \text{Var}(Y) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}.$$

Lognormal densities do not have a finite mgf at any $t > 0$, although all moments are finite. Lognormal densities also have a pronounced skewness.

(i) If $X_1, X_2, \ldots, X_n, n \geq 2$ are independent normal variables with $X_i \sim N(\mu_i, \sigma_i^2)$ then, for any constants $a_1, \ldots, a_n$,

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

In particular, if $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

(j) Based on a sample $X_1, X_2, \ldots, X_n$ of size n from a normal distribution with mean μ and variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is

$$\overline{X} \pm z_{\frac{\alpha}{2}} \sigma / \sqrt{n}.$$

For example, a 95% confidence interval for μ is $\overline{X} \pm 1.96\sigma/\sqrt{n}$.

9.7 Exercises

Exercise 9.1. Let $Z \sim N(0, 1)$. Find

$$P\left(.5 < \left|Z - \frac{1}{2}\right| < 1.5\right); P(1 + Z + Z^2 > 0); P\left(\frac{e^Z}{1+e^Z} > \frac{3}{4}\right); P(\Phi(Z) < .5).$$

Exercise 9.2. Let $Z \sim N(0, 1)$. Find the variance of Z^2 and Z^3.

Exercise 9.3. Let $Z \sim N(0, 1)$. Find the density of $\frac{1}{Z}$. Is the density bounded?

Exercise 9.4. Let $Z \sim N(0, 1)$. Find the mean, median, and mode of

$$Z + 1; 2Z - 3; Z^3.$$

Exercise 9.5. Let $Z \sim N(0, 1)$. Find the density of $\phi(Z)$. Does it have a finite mean?

Exercise 9.6. Let $Z \sim N(0, 1)$. Find the density of $\frac{1}{\phi(Z)}$. Does it have a finite mean?

Exercise 9.7. The 25th and 75th percentiles of a normally distributed random variable are -1 and $+1$. What is the probability that the random variable is between -2 and $+2$?

Exercise 9.8. Suppose X has an $N(\mu, \sigma^2)$ distribution, $P(X \leq 0) = 1/3$, and $P(X \leq 1) = 2/3$. What are the values of μ and σ?

Exercise 9.9 (Standard Normal CDF in Terms of the Error Function). In some places, instead of the standard normal CDF, one sees the *error function* $\text{erf}(x) = (2/\sqrt{\pi}) \int_0^x e^{-t^2} dt$ being used. Express $\Phi(x)$ in terms of $\text{erf}(x)$.

Exercise 9.10 (Grading on a Bell Curve). An instructor is going to give the grades A, B, C, D, F according to the following scale:

$$\begin{aligned} \text{grade} > \mu + 1.5\sigma &: A; \\ \mu + .5\sigma < \text{grade} < \mu + 1.5\sigma &: B; \\ \mu - .5\sigma < \text{grade} < \mu + .5\sigma &: C; \\ \mu - 2\sigma < \text{grade} < \mu - .5\sigma &: D; \\ \text{grade} < \mu - 2\sigma &: F. \end{aligned}$$

What percentage of students get each letter grade? Assume that the grades follow a normal distribution.

Exercise 9.11. Let $Z \sim N(0, 1)$. Find the smallest interval containing a probability of .9.

Exercise 9.12. * **(A Conditioning Problem).** Diameters of ball bearings made at a factory are normally distributed with mean 1.5 cm and s.d. 0.02 cm. Balls whose diameter exceeds 1.52 cm or is less than 1.48 cm are discarded. The rest are shipped for sale. What is the mean and the s.d. of balls that are sent for sale?

Exercise 9.13. * **(A Mixed Distribution).** Let $Z \sim N(0, 1)$ and let $g(Z)$ be the function

$$\begin{aligned} g(z) &= Z \text{ if } |Z| \leq a; \\ &= a \text{ if } Z > a; \\ &= -a \text{ if } Z < -a. \end{aligned}$$

Find and plot the CDF of $g(Z)$. Does $g(Z)$ have a continuous distribution? A discrete distribution? Or neither?

Exercise 9.14. The weights of instant coffee jars packed by a food processor are normally distributed with a standard deviation of 0.2 oz. The processor has set the mean such that about 2% of the jars weigh more than 8.41 oz. What is the mean setting?

Exercise 9.15. * **(An Interesting Calculation).** Suppose $X \sim N(\mu, \sigma^2)$. Prove that

$$E[\Phi(X)] = \Phi(\mu/\sqrt{1+\sigma^2}).$$

Exercise 9.16. * **(Useful Normal Distribution Formulas).** Prove the following primitive (indefinite integral) formulas:

(a) $\int x^2\phi(x)dx = \Phi(x) - x\phi(x)$.

(b) $\int[\phi(x)]^2dx = 1/(2\sqrt{\pi})\Phi(x\sqrt{2})$.

(c) $\int \phi(x)\phi(a+bx)dx = (1/t)\phi(a/t)\Phi(tx + a/t)$, where $t = \sqrt{1+b^2}$.

(d) $\int x\phi(x)\Phi(bx)dx = b/(\sqrt{2\pi}t)\Phi(tx) - \phi(x)\Phi(bx)$.

Exercise 9.17. * **(Useful Normal Distribution Formulas).** Prove the following definite integral formulas, with t as in the previous exercise:

(a) $\int_0^\infty x\phi(x)\Phi(bx)dx = 1/(2\sqrt{2\pi})[1 + b/t]$.

(b) $\int_{-\infty}^\infty x\phi(x)\Phi(bx)dx = b/(\sqrt{2\pi}t)$.

(c) $\int_{-\infty}^\infty \phi(x)\Phi(a+bx)dx = \Phi(a/t)$.

(d) $\int_0^\infty \phi(x)[\Phi(bx)]^2dx = 1/(2\pi)[\arctan b + \arctan\sqrt{1+2b^2}]$.

(e) $\int_{-\infty}^\infty \phi(x)[\Phi(bx)]^2dx = 1/\pi \arctan\sqrt{1+2b^2}$.

Exercise 9.18 (Median and Mode of lognormal). Show that a general lognormal density is unimodal, and find its mode and median.

Hint: For the median, remember that a lognormal variable is e^X, where X is a normal variable.

Exercise 9.19 (Kurtosis of lognormal). Find a formula for the coefficient of kurtosis of a general lognormal density.

Exercise 9.20. Suppose $X \sim N(0,1)$, $Y \sim N(0,9)$, and X, and Y are independent. Find the mean, variance, third moment, and fourth moment of $X + Y$.

Exercise 9.21. Suppose $X \sim N(0,1)$, $Y \sim N(0,9)$, and X, and Y are independent. Find the value of $P((X-Y)^2 > 5)$.

Exercise 9.22. * Suppose Cathy's pocket expenses per month are normally distributed with mean 900 dollars and standard deviation 200 dollars and those of her husband are normally distributed with mean 500 dollars and standard deviation 100 dollars. Assume that the respective pocket expenses are independent. Find the probability that:

(a) The total family pocket expense in one month exceeds 2000 dollars.
(b) Cathy spends twice as much as her husband in pocket expenses in some month.

Exercise 9.23 (Margin of Error of a Confidence Interval). Suppose $X_1, X_2, \ldots, X_n$ are independent $N(\mu, 10)$ variables. What is the smallest n such that the margin of error of a 95% confidence interval for μ is at most .05?

Exercise 9.24. * **(Maximum Error of Polýa's Inequality).** Find $\sup_{-\infty<x<\infty} [B(x) - \Phi(x)]$, where $B(x)$ is the upper bound of Polýa on $\Phi(x)$.

Exercise 9.25. * **(Accuracy of Laplace's Expansion for the Mills Ratio).** What is the smallest number of terms you must keep in Laplace's expansion for the Mills ratio in order to have the *percentage error* at most 4% for $x \geq 2$? This requires use of a computer.

Exercise 9.26 (Use Your Computer). Suppose you have simulated 100 values from a standard normal distribution. By cleverly using the quantile transformation method, convert these 100 values into 100 values from a standard exponential density.

References

Bryc, W. (1995). *The Normal Distribution*, Springer, New York.

Feller, W. (1971). *An Introduction to Probability Theory and Applications, Vol. II*, Wiley, New York.

Freedman, D. (2005). *Statistical Models*, Cambridge University Press, New York.

Heyde, C. (1963). On a property of the lognormal distribution, J.R. Statist. Soc. Ser. B, 25(2), 392–393.

Johnson, N., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions, Vol. II*, Wiley, New York.

Kendall, M. and Stuart, A. (1976). *Advanced Theory of Statistics, Vol. I*, Fourth ed., Macmillan, New York.

Patel, J. and Read, C. (1996). *Handbook of the Normal Distribution*, Marcel Dekker, New York.

Petrov, V. (1975). *Limit Theorems of Probability Theory*, Clarendon Press, London.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, Wiley, New York.

Stigler, S. (1975). Studies in the history of probability and statistics; Napoleonic statistics, the work of Laplace, Biometrika, 62, 503–517.

Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge, MA.

Tong, Y. (1990). *Multivariate Normal Distribution*, Springer, New York.

Chapter 10
Normal Approximations and the Central Limit Theorem

Many of the special discrete and special continuous distributions that we have discussed can be well approximated by a normal distribution for suitable configurations of their underlying parameters. Typically, the normal approximation works well when the parameter values are such that the skewness of the distribution is small. For example, binomial distributions are well approximated by a normal distribution when n is large and p is not too small or too large. Gamma distributions are well approximated by a normal distribution when the shape parameter α is large. Whenever we see a certain phenomenon empirically all too often, we might expect that there is a unifying mathematical result there, and in this case indeed there is. The unifying mathematical result is one of the most important results in all of mathematics and is called the *central limit theorem*. The subject of central limit theorems is incredibly diverse. In this chapter, we present the basic or the *canonical central limit theorem* and its applications to certain problems with which we are already familiar. Among numerous excellent references on central limit theorems, we recommend Feller (1968, 1971) and Pitman (1992) for lucid expositions and examples. The subject of central limit theorems also has a really interesting history; we recommend Le Cam (1986) and Stigler (1986) in this area. Careful and comprehensive mathematical treatments are available in Hall (1992) and Bhattacharya and Rao (1986). For a diverse selection of examples, see DasGupta (2008).

10.1 Some Motivating Examples

Example 10.1. Consider a binomial random variable X with parameters n and p; we will fix $p = .1$ and see the effect of increasing n on the pmf of X. Recall that the binomial pmf has the formula $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \ldots, n$. Using this formula, with $n = 10, 20, 50$, and 100, we have computed and plotted (see Figure 10.1) the pmf of X in the form of a *histogram*, which is a system of rectangles with the height of the rectangle corresponding to a specific x value equal to (or proportional to) the probability of that x value.

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_10,

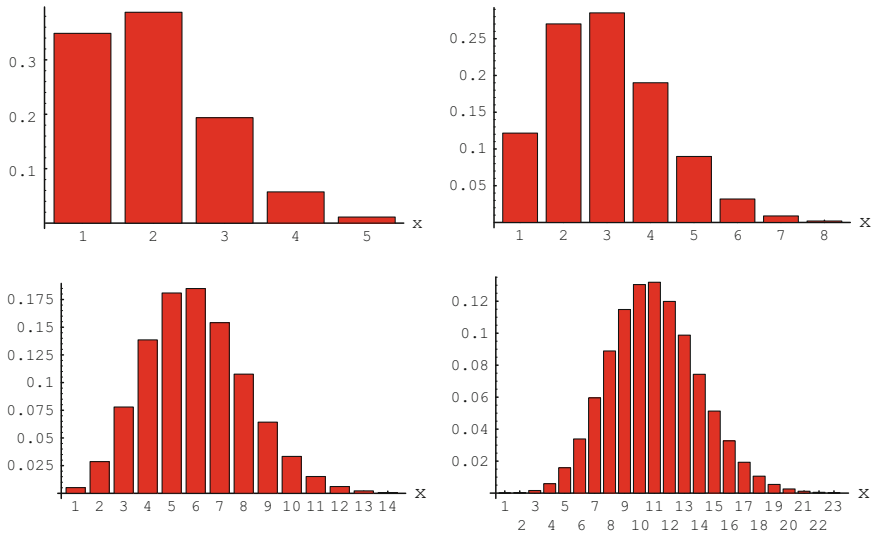


Fig. 10.1 Bin $(n, .1)$ pmf for $n = 10, 20, 50, 100$

We see that the histogram is rather skewed for the smallest n, namely $n = 10$. As n increases, the histogram gets less skewed, and for the largest value, $n = 100$, the histogram looks bell-shaped, centered at 10 and 11, resembling a normal density curve.

What is the explanation? The formula for the coefficient of skewness of a binomial distribution is $\frac{1-2p}{\sqrt{np(1-p)}}$, which goes to zero as $n \to \infty$ for any fixed p. That is, the distribution becomes nearly symmetric as n gets large, although it started out being very skewed when n was small. Indeed, it is true in general that the $Bin(n, p)$ distribution can be well approximated by the $N(np, np(1-p))$ distribution for any fixed p when n is large. If p is near .5, a normal-looking histogram will be produced even for n as small as 20; if p is closer to zero or one, a larger n is necessary to produce a normal-looking histogram. We will see this empirical illustration borne out by a theorem below.

Example 10.2. Recall that the sum of n independent exponential variables, each with mean λ, is distributed as $G(n, \lambda)$, the Gamma distribution with parameters n and λ. We will take $\lambda = 1$ and vary n, choosing $n = 1, 3, 10, 50$, respectively, and plot the density function of $G(n, \lambda)$.

We see a phenomenon similar to our previous binomial example. When n is small, the density is skewed. But, when n increases, the density becomes increasingly bell-shaped, resembling a normal density (see Figure 10.2).

What is common between the binomial and the Gamma examples? In the binomial example, a $Bin(n, p)$ variable is the sum of n independent $Ber(p)$ variables, while in the Gamma example a $G(n, 1)$ variable is the sum of n independent $Exp(1)$

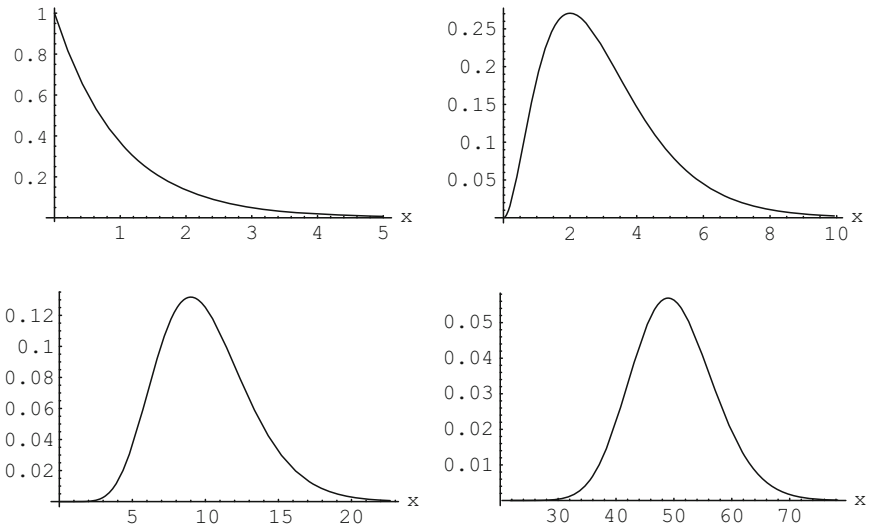


Fig. 10.2 Gamma $(n,1)$ Density for $n = 1, 3, 10, 50$

variables. In both cases, we see that when n is large (i.e., *when we add together a large number of independent random variables), the sum has a density that visually resembles a normal density*. This is in fact what the central limit theorem says. It does not really matter what kinds of variables you add; if you add a large number of independent variables, you will end up with a normal-like density.

10.2 Central Limit Theorem

Theorem 10.1. *For $n \geq 1$, let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having* the same distribution*, and suppose this common distribution, say F, has a finite mean μ and a finite variance σ^2. Let $S_n = X_1 + X_2 + \cdots + X_n$, $\bar{X} = \bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$. Then, as $n \to \infty$,*

(a) $P\left(\dfrac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq x\right) \to \Phi(x) \ \forall\, x \in \mathcal{R}$,

(b) $P\left(\dfrac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq x\right) \to \Phi(x) \ \forall\, x \in \mathcal{R}$.

In words, for large n,

$$S_n \approx N(n\mu, n\sigma^2),$$

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

The proof of this theorem in the generality stated here requires the use of certain tools in probability theory that we have not yet discussed. We will prove the theorem under the more restrictive condition that the underlying distribution F has a finite mgf in some open interval containing zero.

We will use the following notation in the proof. If a sequence of numbers $a_n \to 0$ as $n \to \infty$, we write $a_n = o(1)$. If a_n, and b_n are two sequences of numbers and $\frac{a_n}{b_n} \to 0$ as $n \to \infty$, we write $a_n = o(b_n)$. For example, $a_n = o(n^{-1})$ means not only that $a_n \to 0$ as $n \to \infty$ but even $na_n \to 0$ as $n \to \infty$. We will also need a fact about mgfs in this proof; we state this fact below.

Lemma (Continuity Theorem on MGFs). *Let Y_n be a sequence of random variables with Y_n having a finite mgf $\psi_n(t)$ in some open interval $(-a, a)$ containing zero. If $\psi_n(t) \to e^{t^2/2}$ for each $t \in (-a, a)$ as $n \to \infty$, then $P(Y_n \le x) \to \Phi(x)$ for all real numbers x as $n \to \infty$.*

Proof of the theorem. Part (b) of the theorem is in fact equivalent to part (a), so we just prove part (a). A crucial algebraic simplification that we can make is that we may assume, without loss of generality, that $\mu = 0$ and $\sigma^2 = 1$. This is because if we define a sequence of new random variables $W_i = \frac{X_i - \mu}{\sigma}, i \ge 1$, then the W_i are also iid, and they have mean zero and variance one. Furthermore, $\frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\sum_{i=1}^n W_i}{\sqrt{n}}, n \ge 1$. Thus, we have that $P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \le x\right) \to \Phi(x)$ if and only if $P\left(\frac{\sum_{i=1}^n W_i}{\sqrt{n}} \le x\right) \to \Phi(x)$. Therefore, we go ahead and set $\mu = 0, \sigma^2 = 1$, and show that $P\left(\frac{S_n}{\sqrt{n}} \le x\right) \to \Phi(x)$ as $n \to \infty$. We will prove this by appealing to the lemma concerning mgfs stated above.

Since the X_i are independent, the mgf of $Z_n := \frac{S_n}{\sqrt{n}}$ equals

$$\begin{aligned}
\psi_{Z_n}(t) &= E[e^{tZ_n}] = E\left[e^{t\left(\frac{S_n}{\sqrt{n}}\right)}\right] \\
&= E\left[e^{\sum_{i=1}^n \left[t\left(\frac{X_i}{\sqrt{n}}\right)\right]}\right] = \prod_{i=1}^n E\left[e^{t\frac{X_i}{\sqrt{n}}}\right] \\
&= \left(E\left[e^{\frac{t}{\sqrt{n}}X_1}\right]\right)^n = \psi^n\left(\frac{t}{\sqrt{n}}\right) \\
&\Rightarrow \log \psi_{Z_n}(t) = n \log \psi\left(\frac{t}{\sqrt{n}}\right) \\
&= n \log\left[1 + \frac{t}{\sqrt{n}}\psi'(0) + \frac{t^2}{2n}\psi''(0) + o(n^{-1})\right]
\end{aligned}$$

(by a Taylor expansion of $\psi(\frac{t}{\sqrt{n}})$ around $t = 0$)

$$= n\left(\frac{t}{\sqrt{n}}\psi'(0) + \frac{t^2}{2n}\psi''(0) - \frac{t^2}{2n}(\psi'(0))^2 + o(n^{-1})\right)$$

(by expanding $\log(1+x)$ around $x=0$, which gives $\log(1+x) \approx x - \frac{x^2}{2}$ for $x \approx 0$)

$$= \frac{t^2}{2}\left[\psi''(0) - (\psi'(0))^2\right] + o(1) = \frac{t^2}{2} + o(1)$$

(since $\psi'(0) = \mu = 0$ and $\psi''(0) - (\psi'(0))^2 = \sigma^2 = 1$)

$$\Rightarrow \psi_{Z_n}(t) \to e^{t^2/2}.$$

This proves, by the lemma that we stated above, that $P(Z_n \le x) \to \Phi(x)$ for all x, as was needed.

10.3 Normal Approximation to Binomial

A very important case in which the general central limit theorem applies is the binomial distribution. The CLT allows us to approximate clumsy binomial probabilities involving large factorials using simple and accurate normal approximations. We first give the exact result on normal approximation of the binomial.

Theorem 10.2 (de Moivre-Laplace Central Limit Theorem). *Let $X = X_n \sim Bin(n, p)$. Then, for any fixed p and $x \in \mathcal{R}$,*

$$P\left(\frac{X - np}{\sqrt{np(1-p)}} \le x\right) \to \Phi(x)$$

as $n \to \infty$.

Proof. Identify the binomial variable X with S_n and the X_i as independent $Ber(p)$ variables, so that $\mu = p$ and $\sigma^2 = p(1-p)$.

This theorem tells us how to approximate binomial probabilities of the type $\le$ by using a normal approximation. Sometimes, however, we want to know the probability that a binomial random variable is exactly equal to some value. This can be reduced to a problem of the $\le$ type on noting that $P(X = k) = P(X \le k) - P(X \le k-1)$. Theorems on approximations of probabilities of the $=$ type are called *local limit theorems* because the probability $P(X = k)$ that we are trying to approximate is a probability limited to a *local value*, namely the value k. Here is the binomial local limit theorem.

Theorem 10.3 (de Moivre-Laplace Local Limit Theorem). *Let $X \sim Bin(n, p)$. Then, for any fixed p and $k = 0, 1, \ldots, n$,*

$$P(X = k) = P\left(\frac{X - np}{\sqrt{np(1-p)}} = \frac{k - np}{\sqrt{np(1-p)}}\right)$$

$$\sim \frac{1}{\sqrt{np(1-p)}}\phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$$
$$= \frac{1}{\sqrt{2\pi np(1-p)}}e^{-\frac{(k-np)^2}{2np(1-p)}}.$$

The proof of this theorem uses Stirling's approximation for large factorials and involves only some algebra. It will be omitted.

10.3.1 Continuity Correction

The demoivre-Laplace CLT tells us that if $X \sim Bin(n, p)$, then we can approximate the $\leq$ type probability $P(X \leq k)$ as

$$P(X \leq k) = P\left(\frac{X-np}{\sqrt{np(1-p)}} \leq \frac{k-np}{\sqrt{np(1-p)}}\right)$$
$$\approx \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right).$$

Note that, in applying the normal approximation in the binomial case, we are using a *continuous* distribution to approximate a discrete distribution taking only integer values. The quality of the approximation improves, sometimes dramatically, if we *fill up the gaps between the successive integers*. That is, we *pretend* that an event of the form $X = x$ really corresponds to $x - \frac{1}{2} \leq X \leq x + \frac{1}{2}$. In that case, in order to approximate $P(X \leq k)$, we will in fact expand the domain of the event to $k + \frac{1}{2}$ and approximate $P(X \leq k)$ as

$$P(X \leq k) \approx \Phi\left(\frac{k+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right).$$

This adjusted normal approximation is called a *normal approximation with a continuity correction*. Continuity correction should always be done while computing a normal approximation to a binomial probability. Here are the continuity-corrected normal approximation formulas for easy reference:

$$P(X \leq k) \approx \Phi\left(\frac{k+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right),$$
$$P(m \leq X \leq k) \approx \Phi\left(\frac{k+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{m-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right).$$

We will now apply the continuity correction and the local limit theorem to find normal approximations to binomial probabilities in some examples.

Example 10.3 (Coin Tossing). This is the simplest example of a normal approximation of binomial probabilities. We will solve a number of problems by applying the normal approximation method.

First, suppose a fair coin is tossed 100 times. What is the probability that we obtain between 45 and 55 heads? Denoting X as the number of heads obtained in 100 tosses, $X \sim Bin(n, p)$ with $n = 100$, $p = .5$. Therefore, using the continuity-corrected normal approximation,

$$\begin{aligned} P(45 \le X \le 55) &\approx \Phi\left(\frac{55.5 - 50}{\sqrt{12.5}}\right) - \Phi\left(\frac{44.5 - -50}{\sqrt{12.5}}\right) \\ &= \Phi(1.56) - \Phi(-1.56) = .9406 - .0594 = .8812. \end{aligned}$$

So, the probability that the percentage of heads is between 45% and 55% is high but not very high if we toss the coin 100 times. The next question is: How many times do we need to toss a fair coin to be 99% sure that the percentage of heads will be between 45% and 55%? The percentage of heads is between 45% and 55% if and only if the number of heads is between $.45n$ and $.55n$. Using the continuity-corrected normal approximation, again we want

$$\begin{aligned} .99 &= \Phi\left(\frac{.55n + .5 - .5n}{\sqrt{.25n}}\right) - \Phi\left(\frac{.45n - .5 - .5n}{\sqrt{.25n}}\right) \\ \Rightarrow .99 &= 2\Phi\left(\frac{.55n + .5 - .5n}{\sqrt{.25n}}\right) - 1 \end{aligned}$$

(because, for any real number x, $\Phi(x) - \Phi(-x) = 2\Phi(x) - 1$)

$$\begin{aligned} &\Rightarrow \Phi\left(\frac{.55n + .5 - .5n}{\sqrt{.25n}}\right) = 995 \\ &\Rightarrow \Phi\left(\frac{.05n + .5}{\sqrt{.25n}}\right) = .995. \end{aligned}$$

Now, from a standard normal table, we find that $\Phi(2.575) = .995$. Therefore, we equate

$$\begin{aligned} \frac{.05n + .5}{\sqrt{.25n}} &= 2.575 \\ \Rightarrow .05n + .5 &= 2.575 \times .5\sqrt{n} = 1.2875\sqrt{n}. \end{aligned}$$

Writing $\sqrt{n} = x$, we have here a quadratic equation $.05x^2 - 1.2875x + .5 = 0$ to solve. The root we want is $x = 25.71$, and squaring it gives $n \ge (25.71)^2 = 661.04$. Thus, an *approximate value of n* such that in n tosses of a fair coin the percentage

of heads will be between 45% and 55% with a 99% probability is $n = 662$. Most people find that the value of n needed is higher than what they would have guessed.

Example 10.4 (Public Polling: Predicting the Correct Winner). Normal approximation to binomial probabilities is routinely used in designing polls on an issue, for example polls to predict a winner in an election. Suppose that in an election there are two candidates, A and B, and among *all voters*, 52% support A and 48% support B. A poll of 1400 voters is done; what is the probability that the poll will predict the correct winner?

Let X denote the number of respondents in the poll who favor A. The poll will predict the correct winner if $X > 700$. By using the continuity-corrected normal approximation,

$$\begin{aligned} P(X > 700) &= 1 - P(X \le 700) \approx 1 - \Phi\left(\frac{700.5 - 1400 \times .52}{\sqrt{1400 \times .52 \times .48}}\right) \\ &= 1 - \Phi(-1.5) = \Phi(1.5) = .9332. \end{aligned}$$

As long as the spread between the candidates' support is sufficiently large, say 4% or more, a poll that uses about 1500 respondents will predict the correct winner with a high probability. *But it takes much larger polls to predict the correct spread accurately. See the next example.*

Example 10.5 (Public Polling: Predicting the Vote Share). Consider again an election in which there are two candidates A and B, and suppose the proportion among all voters that support A is p. A poll of n respondents is to be conducted, and we want to know what the value of n should be if with a 95% probability we want to predict the true value of p within an error of at most 2%.

Let X denote the number of respondents in a poll of n people who favor A. We estimate the true value of p by the sample proportion value $\frac{X}{n}$. We want to ensure

$$\begin{aligned} & P\left(\left|\frac{X}{n} - p\right| \le .02\right) \ge .95 \\ \Leftrightarrow\ & P\left(p - .02 \le \frac{X}{n} \le p + .02\right) \ge .95 \\ \Leftrightarrow\ & P\left(np - .02n \le X \le np + .02n\right) \ge .95 \\ \Leftrightarrow\ & P\left(\frac{-.02n}{\sqrt{np(1-p)}} \le \frac{X - np}{\sqrt{np(1-p)}} \le \frac{.02n}{\sqrt{np(1-p)}}\right) \ge .95 \\ \Leftrightarrow\ & P\left(\frac{-.02\sqrt{n}}{\sqrt{p(1-p)}} \le \frac{X - np}{\sqrt{np(1-p)}} \le \frac{.02\sqrt{n}}{\sqrt{p(1-p)}}\right) \ge .95. \end{aligned}$$

Now, using the normal approximation to the binomial,

$$P\left(\frac{-.02\sqrt{n}}{\sqrt{p(1-p)}} \le \frac{X - np}{\sqrt{np(1-p)}} \le \frac{.02\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

$$\approx \Phi\left(\frac{.02\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{.02\sqrt{n}}{\sqrt{p(1-p)}}\right)$$
$$= 2\Phi\left(\frac{.02\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1.$$

From a standard normal table, $\Phi(z) - \Phi(-z) \geq .95$ when $z = 1,96$. We therefore set

$$\frac{.02\sqrt{n}}{\sqrt{p(1-p)}} = 1.96 \Rightarrow n = \left[\frac{1.96\sqrt{p(1-p)}}{.02}\right]^2 = 9604p(1-p).$$

However, the whole point of this calculation is that the true proportion p is not known, so the formula above cannot be used in practice. To circumvent this problem, we use *the most conservative value of p*, namely the value of p that gives the largest value of n in the formula above. That value is $p = .5$, giving ultimately $n = 9604 \times .25 = 2401$. This verifies our statement in the previous example that to predict the actual vote share accurately, one needs much larger polls than for just predicting the correct winner.

Example 10.6 (Random Walk). The theory of random walk is one of the most beautiful areas of probability. Here, we will give an introductory example that makes use of the normal approximation to a binomial.

Suppose a drunkard is standing at some point at time zero (say 11:00 PM) and every second he either moves one step to the right or one step to the left from where he is at that time with equal probability. What is the probability that after two minutes he will be ten or more steps away from where he started? Note that the drunkard will take 120 steps in two minutes.

Let the drunkard's movement at the ith step be denoted as X_i. Then, $P(X_i = \pm 1) = .5$. So, we can think of X_i as $X_i = 2Y_i - 1$, where $Y_i \sim Ber(.5), 1 \leq i \leq n = 120$. If we assume that the drunkard's successive movements $X_1, X_2, \ldots$ are independent, then $Y_1, Y_2, \ldots$ are also independent so $S_n = Y_1 + Y_2 + \cdots Y_n \sim Bin(n, .5)$. Furthermore,

$$|X_1 + X_2 + \cdots + X_n| \geq 10 \Leftrightarrow |2(Y_1 + Y_2 + \cdots + Y_n) - n| \geq 10,$$

so we want to find

$$P(|2(Y_1 + Y_2 + \cdots + Y_n) - n| \geq 10)$$
$$= P\left(S_n - \frac{n}{2} \geq 5\right) + P\left(S_n - \frac{n}{2} \leq -5\right)$$
$$= P\left(\frac{S_n - \frac{n}{2}}{\sqrt{.25n}} \geq \frac{5}{\sqrt{.25n}}\right) + P\left(\frac{S_n - \frac{n}{2}}{\sqrt{.25n}} \leq -\frac{5}{\sqrt{.25n}}\right).$$

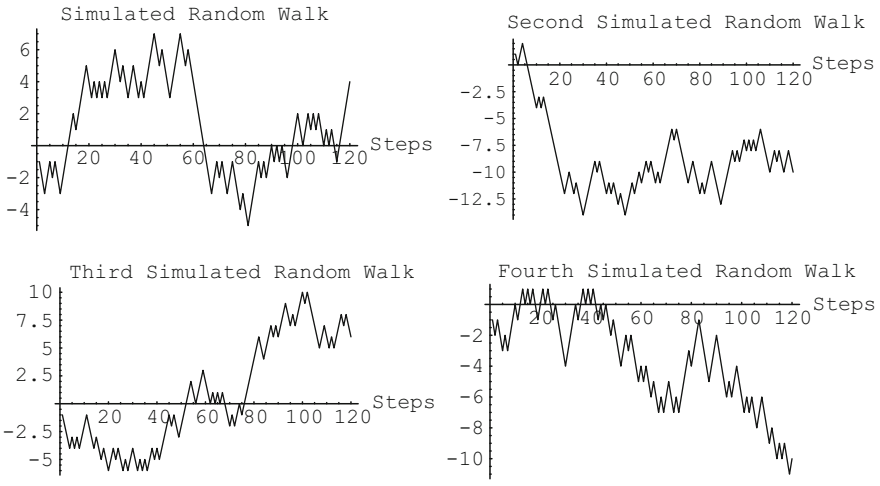


Fig. 10.3 Random walks

Using the normal approximation, this is approximately equal to $2\left[1-\Phi(\frac{5}{\sqrt{.25n}})\right] = 2[1-\Phi(.91)] = 2(1-.8186) = .3628$.

We present in Figure 10.3 four simulated walks of this drunkard over a two minute interval consisting of 120 steps. The different simulations show that the drunkard's random walk could evolve in different ways.

10.3.2 A New Rule of Thumb

A natural practical question is, when can the normal approximation to the binomial be safely applied? It depends on the accuracy of the approximation one wants in a particular problem. However, some general rules can be useful as guides. We provide such a *rule of thumb* below. First, we will show two examples.

Example 10.7 (Use of the Binomial Local Limit Theorem). Suppose $X \sim Bin(50, .4)$ and we want to find the probability that X is equal to 16. The exact value of the probability is $P(X = 16) = \binom{50}{16}.4^{16}.6^{50-16} = .0606$; this exact calculation requires calculations of some large factorials. On the other hand, the normal approximation from the local limit theorem is $P(X = 16) \approx \frac{1}{\sqrt{2\pi 50(.4)(.6)}} e^{-\frac{(16-50\times.4)^2}{2\times 50\times.4\times.6}} = .0591$. The error in the approximation is less than 2.5%. If we desire the normal approximation to be even better than this, then a larger n value will be necessary.

Example 10.8 (Normal Approximation or Poisson Approximation?). In Chapter 6, we discussed Poisson approximations to binomial probabilities when *n is large and p is small*. On the other hand, in this chapter, we discuss normal approximations

when *n is large and p is not very small or very large*. But small, very small, etc., are subjective words and open to interpretation. It is natural to ask when one should prefer a normal approximation and when a Poisson approximation should be preferred. We will offer a rule of thumb, but first we will show an example.

Example 10.9. It is estimated that the probability that a baby will be born on the date the obstetrician predicts is 1/40. What is the probability that of 400 babies born, 15 will be born on the date the doctor predicts?

Let X denote the number of babies among the 400 babies who are born on the predicted day. Then, assuming that the different childbirths are independent, $X \sim Bin(n, p)$, where $n = 400$, $p = 1/40$, so $np = 10, np(1-p) = 9.75$.

We have the exact value of $P(X = 15) = \binom{400}{15}(1/40)^{15}(39/40)^{385} = .0343$. If we do a Poisson approximation, we get the value

$$P(X = 15) \approx e^{-10}10^{15}/15! = .0347.$$

If we do a normal approximation, then by the de Moivre-Laplace local limit theorem

$$P(X = 15) \approx \frac{1}{\sqrt{2\pi \times 9.75}} e^{-(15-10)^2/(2\times 9.75)} = .0345.$$

Thus, although both approximations are very accurate, the normal approximation is even better, although here *n is large and p is small*. The reason that the normal approximation works even better is that, for these values of n and p, the skewness as well as the coefficient of kurtosis of the binomial distribution have become very small.

This idea can be used to write a practical rule for when the normal approximation to the binomial may be used. We use the normal approximation if n and p are such that the skewness and the kurtosis are both sufficiently small. From the formulas in Chapter 6, the coefficients of skewness and kurtosis in the binomial case are, respectively, $\frac{1-2p}{\sqrt{np(1-p)}}$ and $\frac{1-6p(1-p)}{np(1-p)}$. The following rule of thumb for using the normal approximation to the binomial is suggested.

Rule of Thumb for Normal Approximation to Binomial
Use a normal approximation to the binomial when

(a) $$\frac{|1-2p|}{\sqrt{np(1-p)}} \leq .15,$$

and

(b) $$\frac{|1-6p(1-p)|}{np(1-p)} \leq .075.$$

After a little algebra, this works out to

$$n \geq \max\left\{\frac{45(1-2p)^2}{p(1-p)}, \frac{14|1-6p(1-p)|}{p(1-p)}\right\}.$$

Needless to say, the choices of .15 and .075 are partly subjective. But these choices do lead to sensible answers for the value of n needed to produce an accurate normal approximation.

Example 10.10. We provide a table for the minimum n prescribed by this rule of thumb for some values of p.

p	Required n for Normal Approximation
.1	320
.2	100
.3	35
.4	30
.5	30

For p near .5, it will be important to control the kurtosis of the binomial distribution, while for p near 0 (or 1) it will be important to control the skewness. That is what the rule of thumb says.

A famous theorem in probability places an upper bound on the error of the normal approximation in the central limit theorem. If we make this upper bound itself small, then we can be confident that the normal approximation will be accurate. This upper bound on the error of the normal approximation is known as the *Berry-Esseen bound*. Specialized to the binomial case, it says the following; a proof can be seen in Bhattacharya and Rao (1986) or Feller (1968).

Theorem 10.4 (Berry-Esseen Bound for Normal Approximation). *Let* $X \sim Bin(n, p)$ *and let* $Y \sim N(np, np(1-p))$. *Then, for any real number* x,

$$|P(X \le x) - P(Y \le x)| \le \frac{4}{5} \frac{1 - 2p(1-p)}{\sqrt{np(1-p)}}.$$

It should be noted that the Berry-Esseen bound is rather conservative. Thus, accurate normal approximations are produced even when the upper bound, a conservative one, is .1 or so. We do not recommend the use of the Berry-Esseen bound to decide when a normal approximation to the binomial can be accurately done. The bound is simply too conservative. However, it is good to know this bound due to its classic nature.

10.4 Examples of the General CLT

We now give examples of applications of the general CLT for approximating probabilities related to *general sums of independent variables with a common distribution*, not necessarily sums of Bernoulli variables.

Example 10.11 (Distribution of Dice Sums). Suppose a fair die is rolled n times. In Chapter 5, we found the exact distribution of the sum of the n rolls by using de

Moivre's formula. It was a complicated sum. We will now use the CLT to approximate the distribution in a simple manner.

Let $X_i, 1 \le i \le n$ be the individual rolls. Then the sum of the n rolls is $S_n = X_1 + X_2 + \cdots + X_n$. The mean and variance of each individual roll are $\mu = 3.5$ and $\sigma^2 = 2.92$ (see Chapter 4). Therefore, by the CLT,

$$S_n \approx N(3.5n, 2.92n).$$

For example, suppose a fair die is rolled $n = 100$ times. Suppose we want to find the probability that the sum is 300 or more. Direct calculation using de Moivre's formula would be cumbersome at least and may be impossible. However, by the continuity-corrected normal approximation,

$$P(S_n \ge 300) = 1 - P\,(S_n \le 299) = 1 - \Phi\left(\frac{299.5 - 3.5 \times 100}{\sqrt{2.92 \times 100}}\right)$$

$$= 1 - \Phi(-2.96) = \Phi(2.96) = .9985.$$

Example 10.12 (Rounding Errors). Suppose n positive numbers are rounded to their nearest integers and that the rounding errors e_i = (true value of X_i − rounded value of X_i) are independently distributed as $U[-.5, .5]$. We want to find the probability that the total error is at most some number k in magnitude. An example would be a tax agency rounding off the exact refund amount to the nearest integer, in which case the total error would be the agency's loss or profit due to this rounding process.

From the general formulas for the mean and variance of a uniform distribution, each e_i has mean $\mu = 0$ and variance $\sigma^2 = \frac{1}{12}$. Therefore, by the CLT, the total error $S_n = \sum_{i=1}^{n} e_i$ has the approximate normal distribution

$$S_n \approx N\left(0, \frac{n}{12}\right).$$

For example, if $n = 1000$, then

$$P(|S_n| \le 20) = P(S_n \le 20) - P(S_n \le -20)$$

$$= P\left(\frac{S_n}{\sqrt{\frac{n}{12}}} \le \frac{20}{\sqrt{\frac{n}{12}}}\right) - P\left(\frac{S_n}{\sqrt{\frac{n}{12}}} \le \frac{-20}{\sqrt{\frac{n}{12}}}\right)$$

$$\approx \Phi(2.19) - \Phi(-2.19) = .9714.$$

We see from this that, due to the cancellations of positive and negative errors, the tax agency is unlikely to lose or gain much money from rounding.

Example 10.13 (Sum of Uniforms). In the previous example, we approximated the distribution of the sum of n independent uniforms on $[-.5, .5]$ by a normal distribution.

We can do exactly the same thing for the sum of n independent uniforms on any general interval $[a, b]$. It is interesting to ask what the exact density of the sum of n independent uniforms on a general interval $[a, b]$ is. Since a uniform random variable on a general interval $[a, b]$ can be transformed to a uniform on the unit interval $[-1, 1]$ by a linear transformation and vice versa (see Chapter 7), we ask what the exact density of the sum of n independent uniforms on $[-1, 1]$ is. We want to compare this exact density with a normal approximation for various values of n.

When $n = 2$, the density of the sum is a triangular density on $[-2, 2]$, which is a piecewise linear polynomial. In general, the density of the sum of n independent uniforms on $[-1, 1]$ is a piecewise polynomial of degree $n-1$, there being n different arcs in the graph of the density. The exact formula is

$$f_n(x) = \frac{1}{2^n(n-1)!} \sum_{k=0}^{\lfloor \frac{n+x}{2} \rfloor} (-1)^k \binom{n}{k} (n + x - 2k)^{n-1} \quad \text{if} \quad |x| \le n;$$

see Feller (1971).

On the other hand, the CLT approximates the density of the sum by the $N(0, \frac{n}{3})$ density. It would be interesting to compare plots of the exact and the approximating normal densities for various n. We see from Figures 10.4–10.6 that the normal approximation is already nearly exact when $n = 8$.

Example 10.14. A sprinter covers on average 140 cm, with a standard deviation of 5 cm, in each stride. What is the approximate probability that this runner will cover the 100 m distance in 70 or fewer steps? 72 or fewer steps?

Denote the distance covered by the sprinter in n strides by $X_1, X_2, \ldots, X_n$, and assume that, for any n, $X_1, X_2, \ldots, X_n$ are independent variables. Each X_i has mean $\mu = 140$ and $\sigma^2 = 25$. Therefore, by the CLT, the total distance covered in n strides,

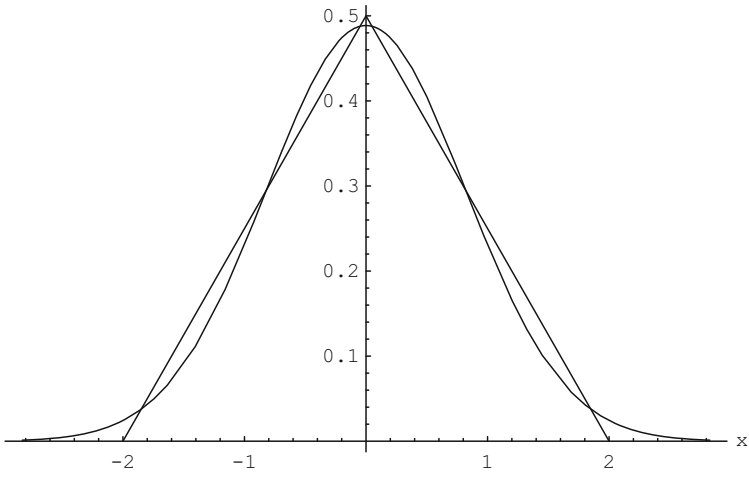


Fig. 10.4 Exact and approximating normal densities for sum of uniforms; $n = 2$

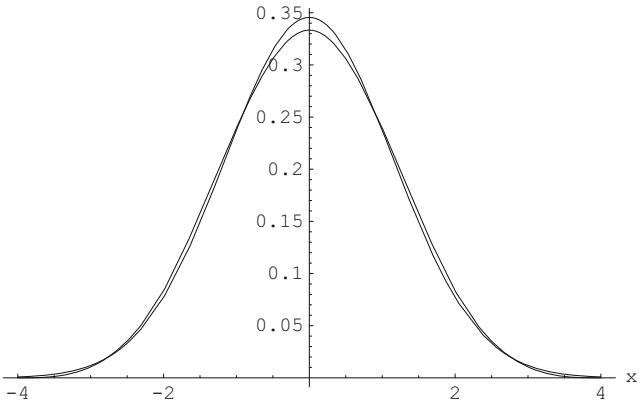


Fig. 10.5 Exact and approximating normal densities for sum of uniforms; $n = 4$

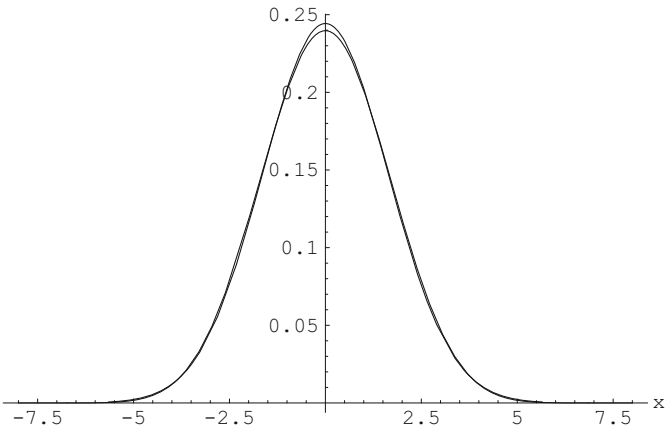


Fig. 10.6 Exact and approximating normal densities for sum of uniforms; $n = 8$

$S_n = X_1 + X_2 + \cdots + X_n$, is approximately $N(140n, 25n)$. To say that the sprinter can cover 100 m = 10,000 cm in n strides is the same as saying $S_n \geq 10{,}000$. Therefore, the probability of covering 100 m in 70 or fewer steps is

$$P(S_{70} \geq 10{,}000) = 1 - P(S_{70} < 10{,}000) \approx 1 - \Phi\left(\frac{10{,}000 - 140 \times 70}{\sqrt{25 \times 70}}\right)$$
$$= 1 - \Phi(4.78) \approx 0.$$

Now increase the number of steps to 72. Then,

$$P(S_{72} \geq 10{,}000) = 1 - P(S_{72} < 10{,}000) \approx 1 - \Phi\left(\frac{10{,}000 - 140 \times 72}{\sqrt{25 \times 72}}\right)$$
$$= 1 - \Phi(-1.89) = .9706.$$

Example 10.15 (Distribution of a Product). Suppose a fair die is rolled 20 times and you are promised a prize if the *geometric mean* of the 20 rolls exceeds 3.5. What are your chances of winning? Recall that the geometric mean of n positive numbers $a_1, a_2, \ldots, a_n$ is defined to be $(a_1 a_2 \cdots a_n)^{\frac{1}{n}}$.

First note that we do not have any means of finding the exact distribution of the product of 20 dice rolls, and enumeration of 6^{20} sample points is impossible. So we are basically forced to make an approximation. How do we find such an approximation?

By writing Y_i as the ith roll and $X_i = \log Y_i$, we get $\log(\prod_{i=1}^n Y_i)^{1/n} = \frac{1}{n}\sum_{i=1}^n \log Y_i = \frac{1}{n}\sum_{i=1}^n X_i$. This use of the logarithm turns our product problem into a problem about sums. Each X_i has the mean $\mu = \frac{1}{6}[\log 1 + \log 2 + \cdots + \log 6] = \frac{\log 6!}{6} = 1.097$. Also, the second moment of each X_i is $\frac{1}{6}[(\log 1)^2 + (\log 2)^2 + \cdots + (\log 6)^2] = 1.568$. Therefore, each X_i has the variance $\sigma^2 = 1.568 - 1.097^2 = .365$. Now, by the CLT,

$$\frac{1}{n}\sum_{i=1}^n X_i \approx N\left(1.097, \frac{.365}{n}\right).$$

Using $n = 20$,

$$\begin{aligned} P\left(\left(\prod_{i=1}^n Y_i\right)^{1/n} > 3.5\right) &= P\left(\log\left(\prod_{i=1}^n Y_i\right)^{1/n} > \log 3.5\right) \\ &= P\left(\frac{1}{n}\sum_{i=1}^n X_i > 1.25\right) \approx 1 - \Phi\left(\frac{1.25 - 1.097}{\sqrt{\frac{.365}{20}}}\right) \\ &= 1 - \Phi(1.13) = 1 - .8708 = .1292. \end{aligned}$$

Thus, there is only about a 13% chance that you will win the prize. What makes the offer unattractive is that the geometric mean of any set of positive numbers is smaller than their simple average. Thus, if the offer was to give a prize if the simple average of your 20 rolls exceeds 3.5, there would have been about a 50% chance of winning the prize, but phrasing the offer in terms of the geometric mean makes it an unattractive offer.

Example 10.16 (Risky Use of the CLT). Suppose the checkout time at a supermarket has a mean of four minutes and a standard deviation of one minute. You have just joined the queue in a lane, where there are eight people ahead of you. From just this information, can you say anything useful about the chances that you can be finished checking out within half an hour?

With the information provided being *only* on the mean and the variance of an individual checkout time but otherwise nothing about the distribution, a possibility

is to use the CLT, *although here n is only 9, which is not large*. Let $X_i, 1 \le i \le 8$, be the checkout times taken by the eight customers ahead of you and X_9 your time. If we use the CLT, then we will have

$$S_n = \sum_{i=1}^{9} X_i \approx N(36, 9).$$

Therefore,

$$P(S_n \le 30) \approx \Phi\left(\frac{30-36}{3}\right) = \Phi(-2) = .0228.$$

In situations such as this, where the information available is extremely limited, we sometimes use the CLT, but it is risky. It may be better to model the distribution of checkout times and answer the question under that chosen model.

10.5 Normal Approximation to Poisson and Gamma

A Poisson variable with an integer parameter $\lambda = n$ can be thought of as the sum of n independent Poisson variables each with mean 1. Likewise, a Gamma variable with parameters $\alpha = n$ and λ can be thought of as the sum of n independent exponential variables, each with mean λ. So, in these two cases the CLT already implies that a normal approxmation to the Poisson and Gamma distributions holds when n is large. However, *even if the Poisson parameter λ is not an integer and even if the Gamma parameter α is not an integer*, if λ or α is large, *a normal approximation still holds*. See Figure 10.7 for an illustration. These results can be proved directly by using the mgf technique. Theorems 10.5 and 10.6 give the normal approximation results for general Poisson and Gamma distributions.

Theorem 10.5. *Let $X \sim Poisson(\lambda)$. Then*

$$P\left(\frac{X-\lambda}{\sqrt{\lambda}} \le x\right) \to \Phi(x) \text{ as } \lambda \to \infty$$

for any real number x.

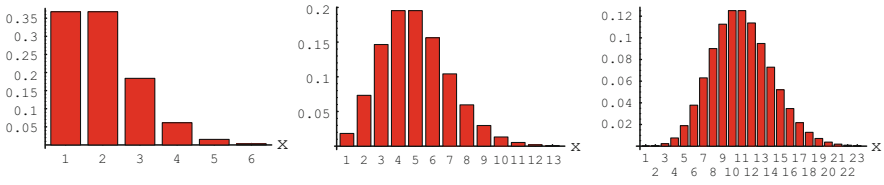


Fig. 10.7 Poisson pmf for lambda = 1, 4, 10

Notationally, for large λ,

$$X \approx N(\lambda, \lambda).$$

Theorem 10.6. *Let $X \sim G(\alpha, \lambda)$. Then, for every fixed λ,*

$$P\left(\frac{X - \alpha\lambda}{\lambda\sqrt{\alpha}} \le x\right) \to \Phi(x) \text{ as } \alpha \to \infty$$

for any real number x.

Notationally, for large α,

$$X \approx N(\alpha\lambda, \alpha\lambda^2).$$

Example 10.17. April receives three phone calls per day on average at her home. We want to find the probability that she will receive more than 100 phone calls next month.

Let X_i be the number of calls April receives on the ith day of the next month. Then the number of calls she will receive in the entire month is $\sum_{i=1}^{n} X_i$; we assume that $n = 30$. If each X_i is assumed to be Poisson with mean 3 and the days are independent, then $\sum_{i=1}^{n} X_i \sim Poi(\lambda)$ with $\lambda = 90$. By the normal limit theorem above, using a continuity correction,

$$P\left(\sum_{i=1}^{n} X_i > 100\right) = 1 - P\left(\sum_{i=1}^{n} X_i \le 100\right)$$
$$\approx 1 - \Phi\left(\frac{100.5 - 90}{\sqrt{90}}\right) = 1 - \Phi(1.11) = 1 - .8665 = .1335.$$

Exact calculation of this probability would be somewhat clumsy because of the large value of λ. That is the advantage in doing a normal approximation.

Example 10.18 (Nuclear Accidents). Suppose the probability of having any nuclear accidents in any nuclear plant during a given year is .0005 and that a country has 100 such nuclear plants. What is the probability that there will be at least six nuclear accidents in the country during the next 250 years?

Let X_{ij} be the number of accidents in the ith year in the jth plant. We assume that each X_{ij} has a common Poisson distribution. The parameter, say θ, of this common Poisson distribution is determined from the equation $e^{-\theta} = 1 - .0005 = .9995 \Rightarrow \theta = -\log(.9995) = .0005$. Assuming that these X_{ij} are all independent, the number of accidents T in the country during 250 years has a $Poi(\lambda)$ distribution, where $\lambda = \theta \times 100 \times 250 = .0005 \times 100 \times 250 = 12.5$. If we now do a normal approximation with continuity correction,

$$P(T \ge 6) \approx 1 - \Phi\left(\frac{5.5 - 12.5}{\sqrt{12.5}}\right)$$
$$= 1 - \Phi(-1.98) = .9761.$$

So we see that although the chances of having any accidents in a particular plant in any particular year are small, collectively and in the long run, the chances are high that there will be quite a few such accidents.

Example 10.19 (Confidence Interval for a Poisson Mean). The normal approximation to the Poisson distribution can be used to find a confidence interval for the mean of a Poisson distribution. We have already seen an example of a confidence interval for a normal mean in Chapter 9. We will now work out the Poisson case using the normal approximation to Poisson.

Suppose $X \sim Poi(\lambda)$. By the normal approximation theorem, if λ is large, then $\frac{X-\lambda}{\sqrt{\lambda}} \approx N(0,1)$. Now, a standard normal random variable Z has the property $P(-1.96 \le Z \le 1.96) = .95$. Since $\frac{X-\lambda}{\sqrt{\lambda}} \approx N(0,1)$, we have

$$\begin{aligned} & P\left(-1.96 \le \frac{X-\lambda}{\sqrt{\lambda}} \le 1.96\right) \approx .95 \\ \Leftrightarrow\ & P\left(\frac{(X-\lambda)^2}{\lambda} \le 1.96^2\right) \approx .95 \\ \Leftrightarrow\ & P((X-\lambda)^2 - 1.96^2\lambda \le 0) \approx .95 \\ \Leftrightarrow\ & P(\lambda^2 - \lambda(2X + 1.96^2) + X^2 \le 0) \approx .95. \quad (*) \end{aligned}$$

Now the quadratic equation

$$\lambda^2 - \lambda(2X + 1.96^2) + X^2 = 0$$

has the roots

$$\begin{aligned} \lambda = \lambda_{\pm} &= \frac{(2X + 1.96^2) \pm \sqrt{(2X + 1.96^2)^2 - 4X^2}}{2} \\ &= \frac{(2X + 1.96^2) \pm \sqrt{14.76 + 15.37X}}{2} \\ &= (X + 1.92) \pm \sqrt{3.69 + 3.84X}. \end{aligned}$$

The quadratic $\lambda^2 - \lambda(2X + 1.96^2) + X^2$ is ≤ 0 when λ is between these two values $\lambda_{\pm}$, so we can rewrite $(*)$ as

$$P((X+1.92) - \sqrt{3.69 + 3.84X} \le \lambda \le (X+1.92) + \sqrt{3.69 + 3.84X}) \approx .95 \quad (**).$$

In statistics, one often treats the parameter λ as unknown and uses the data value X to estimate the unknown λ. The statement $(**)$ is interpreted as saying that, with approximately 95% probability, λ will fall inside the interval of values

$$(X + 1.92) - \sqrt{3.69 + 3.84X} \le \lambda \le (X + 1.92) + \sqrt{3.69 + 3.84X},$$

so the interval

$$[(X + 1.92) - \sqrt{3.69 + 3.84X}, (X + 1.92) + \sqrt{3.69 + 3.84X}]$$

is called an *approximate 95% confidence interval for* λ. We see that it is derived from the normal approximation to a Poisson distribution.

Example 10.20 (Normal Approximation in a Gamma Case). Diabetes is one of the main causes for development of an eye disease known as retinopathy, which causes damage to the blood vessels in the retina and growth of abnormal blood vessels, potentially causing loss of vision. The average time to develop retinopathy after the onset of diabetes is 15 years, with a standard deviation of four years.

Suppose we let X be the time from onset of diabetes until development of retinopathy and that we model it as $X \sim G(\alpha, \lambda)$. Then, we have

$$\alpha\lambda = 15; \lambda\sqrt{\alpha} = 4 \Rightarrow \sqrt{\alpha} = \frac{15}{4} = 3.75 \Rightarrow \alpha = 14.06, \lambda = 1.07.$$

Suppose we want to know what percentage of diabetes patients develop retinopathy within 20 years. Since $\alpha = 14.06$ is large, we can use a normal approximation:

$$P(X \leq 20) \approx \Phi\left(\frac{20 - 15}{4}\right) = \Phi(1.25) = .8944;$$

i.e., under the Gamma model, approximately 90% develop diabetic retinopathy within 20 years.

10.6 ∗ Convergence of Densities and Higher-Order Approximations

If in the central limit theorem each individual X_i is a continuous random variable with a density $f(x)$, then the sum $S_n = \sum_{i=1}^{n} X_i$ also has a density for each n and hence so does the standardized sum $\frac{S_n - n\mu}{\sigma\sqrt{n}}$. It is natural to ask if the density of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to the standard normal density when $n \to \infty$. This is true under suitable conditions on the basic density $f(x)$. We will present a result in this direction. But first let us see an example. Recall that the notation $a_n = O(b_n)$ used in the example means that there is a finite positive constant K such that $|a_n| \leq Kb_n$ for all n. We do not worry about exactly what the constant K is; we only care that such a constant K exists.

Example 10.21 (Convergence of Chi-Square Density to Normal). Suppose $X_1, X_2, \ldots$ are iid $\chi^2(2)$ with density $\frac{1}{2}e^{-x/2}$; i.e., the $\chi^2(2)$ density is just an exponential density with mean two. We verify that in this example in fact the density of $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - 2n}{2\sqrt{n}}$ converges pointwise to the $N(0, 1)$ density,

Since $S_n = \sum_{i=1}^n X_i$ has the $\chi^2(2n)$ distribution with density $\frac{e^{-x/2}x^{n-1}}{2^n\Gamma(n)}$, Z_n has the density $f_n(z) = \frac{e^{-(z\sqrt{n}+n)}(1+\frac{z}{\sqrt{n}})^{n-1}n^{n-\frac{1}{2}}}{\Gamma(n)}$. Hence, by taking the logarithm and using the fact that $\log(1+x) = x - x^2/2 + O(x^{3/2})$ as $x \to 0$, we get

$$\begin{aligned}\log f_n(z) &= -z\sqrt{n} - n + (n-1)\left(\frac{z}{\sqrt{n}} - \frac{z^2}{2n} + O(n^{-3/2})\right) \\ &\quad + \left(n - \frac{1}{2}\right)\log n - \log\Gamma(n) \\ &= -z\sqrt{n} - n + (n-1)\left(\frac{z}{\sqrt{n}} - \frac{z^2}{2n} + O(n^{-3/2})\right) + \left(n - \frac{1}{2}\right)\log n \\ &\quad - \left(n\log n - n - \frac{1}{2}\log n + \log\sqrt{2\pi} + O(n^{-1})\right)\end{aligned}$$

on using Stirling's approximation for $\log\Gamma(n) = \log(n-1)!$.

On cancelling of terms, this gives

$$\log f_n(z) = -\frac{z}{\sqrt{n}} - \log\sqrt{2\pi} - \frac{(n-1)z^2}{2n} + O(n^{-1/2}),$$

implying that $\log f_n(z) \to -\log\sqrt{2\pi} - \frac{z^2}{2}$ and hence $f_n(z) \to \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$, establishing the pointwise density convergence to the standard normal density, which is what we wanted to show.

Of course, we really do not wish to treat each new example as a separate case. It is useful to have a general result that ensures that under suitable conditions, in the central limit theorem, the density of $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to the $N(0,1)$ density. The result below is not the best available result in this direction, but it often applies and is easy to state; a proof can be seen in Bhattacharya and Rao (1986).

Theorem 10.7 (Gnedenko's Local Limit Theorem). *Suppose $X_1, X_2, \ldots$ are independent random variables with a density $f(x)$, mean μ, and variance σ^2. If $f(x)$ is uniformly bounded, then the density function of $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges* uniformly *on the real line $\mathcal{R}$ to the standard normal density $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$.*

Remark. The preceding chi-square example is therefore a special case of Gnedenko's theorem because the χ^2_2 density is obviously uniformly bounded.

10.6.1 ∗ Refined Approximations

One criticism of the normal approximation in the various cases we have described is that any normal distribution is symmetric about its mean, so, by employing a

normal approximation, we necessarily ignore any skewness that may be present in the true distribution that we are approximating. For instance, if the individual X_i's have exponential densities, then the true density of the sum S_n is a Gamma density, which always has a skewness. But a normal approximation ignores that skewness, and, as a result, the quality of the approximation can be poor unless n is quite large. Refined approximations that address this criticism are available. These refined approximations were formally introduced in Edgeworth (1904) and Charlier (1931). As such, they are usually called *Edgeworth densities* and the *Gram-Charlier series*. Although they are basically the same thing, there is a formal difference between the formulas in the Edgeworth density and the Gram-Charlier series. Modern treatments of these refined approximations are carefully presented in Bhattacharya and Rao (1986) and Hall (1992). We present here a refined density approximation that adjusts the normal approximation for skewness and another one that also adjusts for kurtosis. Some discussion of their pros and cons will follow the formulas and the theorem below.

Suppose $X_1, X_2, \ldots, X_n$ are continuous random variables with a density $f(x)$. Suppose each individual X_i has four finite moments. Let $\mu, \sigma^2, \beta, \gamma$ denote the mean, variance, coefficient of skewness, and coefficient of kurtosis of the common distribution of the X_i's. Let $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. Define the following three successively more refined density approximations for the density of Z_n:

$$\hat{f}_{n,0}(x) = \phi(x).$$

$$\hat{f}_{n,1}(x) = \left(1 + \frac{\beta(x^3 - 3x)}{6\sqrt{n}}\right)\phi(x),$$

$$\hat{f}_{n,2}(x) = \left(1 + \frac{\beta(x^3 - 3x)}{6\sqrt{n}} + \left[\gamma\frac{x^4 - 6x^2 + 3}{24} + \beta^2\frac{x^6 - 15x^4 + 45x^2 - 15}{72}\right]\frac{1}{n}\right)\phi(x).$$

The functions $\hat{f}_{n,0}(x)$, $\hat{f}_{n,1}(x)$, and $\hat{f}_{n,2}(x)$ are called the *CLT approximation, the first-order Edgeworth expansion, and the second-order Edgeworth expansion for the density of the mean*.

Remark. Of the three approximations, *only* $\hat{f}_{n,0}(x)$ is truly a density function. The functions $\hat{f}_{n,1}(x)$ and $\hat{f}_{n,2}(x)$ become negative for some values of x for a given n. As a result, if they are integrated to obtain approximations for the probability $P(Z_n \le x)$, then the approximations are not monotonically nondecreasing functions of x and can even become negative (or larger than 1). For any given n, the refined approximations give inaccurate and even nonsensical answers for values of x far from zero. However, at any given x, the approximations become more accurate as n increases.

It is important to note that the approximations are of the form $\phi(x) + \frac{P_1(x)}{\sqrt{n}}\phi(x) + \frac{P_2(x)}{n}\phi(x) + \cdots$ for suitable polynomials $P_1(x)$, $P_2(x)$, etc. The

relevant polynomials $P_1(x), P_2(x)$ are related to some very special polynomials, known as *Hermite polynomials*. Hermite polynomials are obtained from successive differentiations of the standard normal density $\phi(x)$. Precisely, the jth Hermite polynomial $H_j(x)$ is defined by the relation

$$\frac{d^j}{dx^j}\phi(x) = (-1)^j H_j(x)\phi(x).$$

In particular,

$$H_1(x) = x; H_2(x) = x^2 - 1; H_3(x) = x^3 - 3x; H_4(x) = x^4 - 6x^2 + 3;$$
$$H_5(x) = x^5 - 10x^3 + 15x; H_6(x) = x^6 - 15x^4 + 45x^2 - 15.$$

By comparing the formulas for the refined density approximations with the formulas for the Hermite polynomials, the connection becomes obvious. They arise in the density approximation formulas as a matter of fact; there is no intuition for it.

Example 10.22. Suppose $X_1, X_2, \ldots, X_n$ are independent $Exp(1)$ variables, and let $n = 15$. The exact density of the sum S_n is $G(n, 1)$, a Gamma density. By a simple linear transformation, the exact density of Z_n is

$$f_n(x) = \frac{\sqrt{n}e^{-n-x\sqrt{n}}(n + x\sqrt{n})^{n-1}}{(n-1)!}, x \geq -\sqrt{n}.$$

For the standard exponential density, $\beta = 4$, so the first-order Edgeworth expansion is

$$\hat{f}_{n,1}(x) = \left(1 + \frac{4(x^3 - 3x)}{6\sqrt{n}}\right)\phi(x).$$

The exact density, the CLT approximation, and the first-order Edgeworth expansion are plotted in Figure 10.8 to explore the quality of the approximations. The exact density is visibly skewed. The CLT approximation of course completely misses the skewness. The Edgeworth approximation does capture the skewness nicely. But, on close inspection, we find that it becomes negative when x is less than about -2.5.

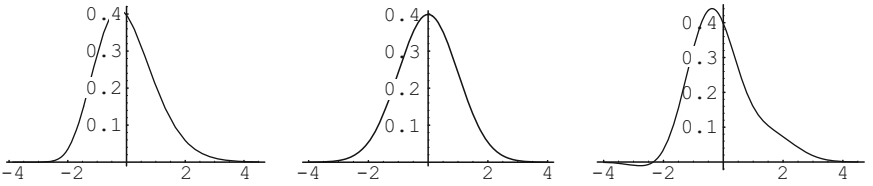


Fig. 10.8 Exact, CLT approximation, and first-order Edgeworth approximation in EXP (1) case

As a result, if the Edgeworth approximation is used to approximate tail probabilities of the form $P(Z_n \leq x)$, then the probability will actually increase when x is decreased below $x = -2.5$.

10.7 Practical Recommendations for Normal Approximations

We have presented normal approximations to the binomial, Poisson, uniform, and Gamma distributions as specific examples in this chapter by appealing to the central limit theorem. In the binomial and Poisson cases, we presented approximations with or without a continuity correction. Normal approximations are useful for some other standard distributions because these distributions do not have any analytical formula, or not an easy one, for the CDF. An example of such a distribution is a general Beta distribution. For practical use by students and other practitioners, we put together a set of normal approximation formulas for a selection of standard distributions. In some cases, we provide two formulas, and a user may use both for comparison. These formulas are based on comparative research work of numerous people on their effectiveness; references to many of them can be seen in Abramowitz and Stegun (1970) and in Chapter 7 of Patel and Read (1996).

For the approximation in the negative binomial case, the following relationship with binomial distributions has been used in our list of formulas:

$$\begin{gathered}\text{Let } X \sim NB(r,\theta), Y \sim Bin(m,\theta), Z \sim Bin(m,1-\theta); \text{ then,} \\ P(X > m) = P(Y < r) = P(Z > m - r) \\ \Leftrightarrow P(X \leq m) = P(Z \leq m - r).\end{gathered}$$

Similarly, Beta densities with integer parameters also have a relationship with binomial distributions; see the exercises in Chapter 8. However, we treat Beta densities with general parameters below.

Distribution	Quantity Approximated	Approximation Formula
$Bin(n, p)$	$P(X \leq k)$	$\Phi\left(\dfrac{k + .5 - np}{\sqrt{np(1-p)}}\right)$ $\Phi(z) - \dfrac{1}{6\sqrt{np(1-p)}}(z^2 - 1)\phi(z)$, $z = \dfrac{k + .5 - np}{\sqrt{np(1-p)}}$
$Poi(\lambda)$	$P(X \leq k)$	$\Phi\left(\dfrac{k + .5 - \lambda}{\sqrt{\lambda}}\right)$ $\Phi(2\sqrt{k + .75} - 2\sqrt{\lambda})$
$NB(r, \theta)$	$P(X \leq m)$	Use formula for binomial case using $k = m - r, n = m, p = 1 - \theta$

$Hypergeo(n, D, N)$	$P(X \le k)$	$\Phi(z),\ z = \dfrac{k + .5 - n\dfrac{D}{N}}{\sqrt{n\dfrac{D}{N}\left(1 - \dfrac{D}{N}\right)\dfrac{N-n}{N-1}}}$
$Be(a, b)$	$P(X \le x)$	$\Phi(z),\ z = \dfrac{3\left[u\left(1 - \dfrac{1}{9b}\right) - v\left(1 - \dfrac{1}{9a}\right)\right]}{\sqrt{u^2/b + v^2/a}}$ $a + b > 6, (a + b - 1)(1 - x) \le .8,$ $u = (bx)^{1/3}, v = (a(1-x))^{1/3}$
χ^2_m	$P(X \le x)$	$\Phi(\sqrt{2x} - \sqrt{2m-1})$ $\Phi\left(\sqrt{\dfrac{9m}{2}}\left[\left(\dfrac{x}{m}\right)^{1/3} - 1 + \dfrac{2}{9m}\right]\right)$

10.8 Synopsis

(a) The central limit theorem (CLT) for iid random variables says that if $X_1, X_2, \ldots$ are iid random variables with finite mean μ and finite variance σ^2, and if, for $n \ge 1$, $S_n = X_1 + \cdots + X_n$ and $\bar{X} = \bar{X}_n = \frac{S_n}{n}$, then, for any real x,

$$P\left(\frac{Sn - n\mu}{\sqrt{n\sigma^2}} \le x\right) = P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \le x\right) \to \Phi(x)$$

as $n \to \infty$. Colloquially, for large n, $S_n \approx N(n\mu, n\sigma^2)$ and $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$.

(b) In particular, in the binomial, Poisson, and Gamma cases, the following normal approximations hold:

If $X = X_n \sim Bin(n, p)$, then, for any real x, $P\left(\frac{X-np}{\sqrt{np(1-p)}} \le x\right) \to \Phi(x)$ as $n \to \infty$

If $X \sim Poi(\lambda)$, then, for any real x, $P\left(\frac{X-\lambda}{\sqrt{\lambda}} \le x\right) \to \Phi(x)$ as $\lambda \to \infty$.

If $X \sim Gamma(\alpha, \lambda)$, then, for any real x, $P\left(\frac{X-\alpha\lambda}{\lambda\sqrt{\alpha}} \le x\right) \to \Phi(x)$ as $\alpha \to \infty$.

(c) There are also *local limit theorems* that give approximate values for $P(X = k)$ in the binomial case and assure convergence of the density of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ to the standard normal density in the continuous case.

(d) For the normal approximation to the binomial, two practical rules to follow are the following:

(i) Use continuity correction.

(ii) Use the rule of thumb given in the text to decide if in a particular case the normal approximation is safe to apply.

The continuity-corrected normal approximation to the binomial says that

$$P(m \le X \le k) \approx \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{m - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

(e) The normal approximation to the density of $\bar{X}$ ignores the skewness in the true density of $\bar{X}$ when such skewness is present. Higher-order density approximations, known as *Edgeworth expansions*, that adjust for the skewness and the kurtosis are available. At any given x, the Edgeworth density approximations become more accurate as n increases. But, for a given n, there are values of x at which the Edgeworth density approximations become negative, so the Edgeworth densities are not truly densities.

10.9 Exercises

Exercise 10.1. Suppose a fair coin is tossed ten times. Find the probability of obtaining six or more heads and compare it with a normal approximation with and without a continuity correction.

Exercise 10.2. A fair die is rolled 25 times. Let X be the number of times a six is obtained. Find the exact value of $P(X = 6)$ and compare it with a normal approximation of $P(X = 6)$.

Exercise 10.3. A basketball player has a history of converting 80% of his free throws. Find a normal approximation with a continuity correction of the probability that he will make between 18 and 22 out of 25 free throws.

Exercise 10.4 (Rule of Thumb). Suppose $X \sim Bin(n, p)$. For $p = .1, .25, .5$, find the values of n that satisfy the *rule of thumb* for the applicability of a normal approximation.

Exercise 10.5. Two persons have 16 and 32 dollars, respectively. They bet one dollar on the outcome of each of 900 independent tosses of a fair coin. What is an approximation to the probability that neither person is in debt at the end of the 900 tosses?

Exercise 10.6 (Poll). In an election to the U.S. Senate, one candidate has popular support of 53% and the other has support of 47%. On election eve, a newspaper will conduct a poll of 750 voters. Compute a normal approximation with continuity correction of the probability that the poll will predict the correct winner.

Exercise 10.7. A new elevator in a large hotel is designed to carry up to 5000 lbs. of weight. The weights of the elevator's users have an average of 150 lbs. and a standard deviation of 55 lbs. If 30 people get into the elevator, find an approximation to the probability that the elevator will be overloaded.

Exercise 10.8. The cost of a textbook at the college level is on average 50 dollars and a standard deviation of 7 dollars. In a four year bachelors program, a student will need to buy 25 textbooks. Find an approximation to the probability that he or she will have to spend more than 1300 dollars on textbooks.

Exercise 10.9. * Suppose $X_1, X_2, \ldots, X_n$ are independent $N(0, 1)$ variables. Find an approximation to the probability that $\sum_{i=1}^{n} X_i$ is larger than $\sum_{i=1}^{n} X_i^2$ when $n = 10, 20, 30$.

Exercise 10.10 (Airline Overbooking). An airline knows from past experience that 10% of fliers with a confirmed reservation do not show up for the flight. Suppose a flight has 250 seats. How many reservations over 250 can the airline permit if they want to be 95% sure that no more than two passengers with a confirmed reservation would have to be bumped?

Exercise 10.11. * **(Breaking Exactly Even is Unlikely).** Suppose a fair coin is tossed $2n$ times. Prove that the probability of getting exactly n heads converges to zero as $n \to \infty$. How about the probability of getting between $n - 1$ and $n + 1$ heads? Can you generalize to the case of getting between $n - k$ and $n + k$ heads for any fixed number k?

Exercise 10.12 (Dice Sums). Suppose a fair die is rolled 1000 times. Compute an approximation to the probability that the sum of the 1000 rolls will exceed 3600.

Exercise 10.13 (Dice Sums). How many times should a fair die be rolled if you want to be 99% sure that the sum of all the rolls will exceed 100?

Exercise 10.14. For your desk lamp, you have an inventory of 25 bulbs. The lifetime of one bulb has an exponential distribution with mean 1 (in thousands of hours).

(a) What is the *exact* distribution of the total time you can manage with these 25 bulbs?
(b) Find an approximate probability that you can manage more than 30,000 hours with these 25 bulbs.

Exercise 10.15 (A Product Problem). Suppose $X_1, X_2, \ldots, X_{30}$ are 30 independent variables, each distributed as $U[0, 1]$. Find an approximation to the probability that their *geometric mean* (a) exceeds .4; (b) exceeds .5.

Exercise 10.16. There are 100 counties in a particular state. The average number of traffic accidents per week is four for each county. Find an approximation to the probability that there are more than 450 traffic accidents in the state in one week.

Exercise 10.17 (Comparing a Poisson Approximation and a Normal Approximation). Suppose 1.5% of residents of a town never read a newspaper. Compute the exact value, a Poisson approximation, and a normal approximation of the probability that at least one resident in a sample of 50 never reads a newspaper.

Exercise 10.18. * (Comparing a Poisson Approximation and a Normal Approximation). One hundred people will each toss a fair coin 200 times. Compute a Poisson approximation and a normal approximation with a continuity correction of the probability that at least 10 of the 100 people would each obtain exactly 100 heads and 100 tails.

Exercise 10.19 (Confidence Interval for Poisson mean). Derive a formula for an approximate 99% confidence interval for a Poisson mean by using the normal approximation to a Poisson distribution. Compare your formula with the formula for an approximate 95% confidence interval that was worked out in the text. Compute the 95% and 99% confidence intervals if $X = 5, 8, 12$.

Exercise 10.20 (Anything that Can Happen Will Eventually Happen). If you predict in advance the outcomes of ten tosses of a fair coin, the probability that you get them all correct is $(.5)^{10}$, which is very small. Show that if each of 2,000 people each try to predict the ten outcomes correctly, the chance that at least one of them succeeds is better than 85%.

Exercise 10.21. *(A Back Calculation). A psychologist would like to find the average time required for a two-year-old to complete a simple maze. He knows from experience that if he samples $n = 36$ children at random, then in about 2.5% of such samples, the mean time to complete the maze will be larger than 3.65 minutes, and in about 5% of such samples, the mean time to complete the maze will be smaller than 3.35 minutes. What are μ and σ, the mean and the standard deviation of the time taken by one child to finish the maze?

Exercise 10.22 (A Gambling Example). It costs one dollar to play a certain slot machine in Las Vegas. The machine is set by the house to pay two dollars with probability .45 (and to pay nothing with probability .55). Let X_i be the house's net winnings on the ith play of the machine. Then $S_n = \sum_{i=1}^{n} X_i$ is the house's winnings after n plays of the machine. Assuming that successive plays are independent, find

(a) $E(S_n)$;
(b) $\text{Var}(S_n)$;
(c) the approximate probability that after 10,000 plays of the machine the house's winnings are between 800 and 1100 dollars.

Exercise 10.23. * (A Problem on Difference). Tom tosses a fair die 40 times and Sara tosses a fair die 45 times. Tom wins if he can score a larger total than Sara. Find an approximation to the probability that Tom wins.

Exercise 10.24. The proportion of impurities in a sample of water from a lake has a Beta distribution with parameters $\alpha = \beta = 2$. Suppose 25 such water samples are taken. Find an approximation to the probabilities that:

(a) The average proportion of impurities in the samples exceeds .54.
(b) The number of samples for which the proportion of impurities exceeds .54 is at most 15.

Exercise 10.25. * **(Density of Uniform Sums).** Give a direct proof that the density of $\frac{S_n}{\sqrt{\frac{n}{3}}}$ at zero converges to $\phi(0)$, where S_n is the sum of n independent $U[-1, 1]$ variables.

Exercise 10.26. * **(Uniform Sums).** Find the third moment of $\frac{S_n}{\sqrt{\frac{n}{3}}}$, where S_n is the sum of n independent $U[-1, 1]$ variables. Does it converge to zero? Would you expect it to converge to zero?

Exercise 10.27 (Roundoff Errors). Suppose you balance your checkbook by rounding amounts to the nearest dollar. Between 0 and 49 cents, drop the cents; between 50 and 99 cents, drop the cents and add a dollar. Find the approximate probability that the accumulated error in 100 transactions is greater than five dollars (either way), assuming that the number of cents involved is independent and uniformly distributed between 0 and 99.

Exercise 10.28. * **(Random Walk)** Consider the drunkard's random walk example. Find the probability that the drunkard will be at least ten steps over on the right from his starting point after 200 steps. Compute a normal approximation.

Exercise 10.29. *(**Random Walk).** Consider again the drunkard's random walk example. Find the probability that more than 125 times in 200 steps the drunkard steps toward his right. Compute a normal approximation.

Exercise 10.30 (Test Your Intuition). Suppose a fair coin is tossed 100 times. Is it more likely that you will get exactly 50 heads or that you will get more than 60 heads?

Exercise 10.31 (Test Your Intuition). Suppose a fair die is rolled 60 times. Is it more likely that you will get at least 20 sixes or that you will score a total of at least 250?

Exercise 10.32 (Test Your Intuition). Suppose a fair coin is tossed 120 times and a fair die is rolled 120 times. Is it more likely that you will get exactly 60 heads or that you will get exactly 20 sixes?

Exercise 10.33 (Computing an Edgeworth Approximation). Suppose X_1, $X_2, \ldots, X_n$ are independent $U[-1, 1]$ variables. For $n = 5$, plot the exact density of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$, the CLT approximation, and the first-order Edgeworth approximation. For the exact density, use the formula for the density of S_n given in the text. Comment on the accuracy of the two approximations.

Exercise 10.34 (Use Your Computer). Simulate the roll of a fair die 50 times, and evaluate the sum of the 50 values. Repeat the simulation 500 times. Use a software package to draw a histogram of these 500 values of the sum. Do you see a normal-looking distribution?

Exercise 10.35 (Use Your Computer). Simulate the problem of rounding $n = 40$ numbers to their nearest integer when the numbers are chosen uniformly from the interval $[0, 100]$. Find the rounding errors and their sum. Repeat the simulation 500 times. Use a software package to draw a histogram of these 500 values of the sum; remember that you are summing the rounding errors. Do you see a normal-looking distribution?

References

Abramowitz, M. and Stegun, I. (1970). *Handbook of Mathematical Functions*, Dover, New York.
Bhattacharya, R. and Rao, R. (1986). *Normal Approximation and Asymptotic Expansions*, Wiley, New York.
Charlier, C. (1931). *Applications de la théorie des probabilités á l'astronomie*, Gauthier-Villars, Paris.
DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer, New York.
Edgeworth, F. (1904). The law of error, Trans. Cambridge Philos. Soc., 20, 36–65, 113–141.
Feller, W. (1968). *Introduction to Probability Theory and Its Applications*, *Vol. I*, Wiley, New York.
Feller, W. (1971). *Introduction to Probability Theory and Its Applications*, *Vol. II*, Wiley, New York.
Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer, New York.
Le Cam, L. (1986). The central limit theorem around 1935, Statist. Sci., 1, 78–91.
Patel, J. and Read, C. (1996). *Handbook of the Normal Distribution*, Marcel Dekker, New York.
Pitman, J. (1992). *Probability*, Springer, New York.
Stigler, S. (1986). *History of Statistics: Measurement of Uncertainty before 1900*, Harvard University press, Cambridge, MA.

Chapter 11
Multivariate Discrete Distributions

We have provided a detailed treatment of distributions of one discrete or one continuous random variable in the previous chapters. But often in applications we are just naturally interested in two or more random variables simultaneously. We may be interested in them simultaneously because they provide information about each other or because they arise simultaneously as part of the data in some scientific experiment. For instance, on a doctor's visit, the physician may check someone's blood pressure, pulse rate, blood cholesterol level, and blood sugar level because together they give information about the general health of the patient. Or, in agricultural studies, one may want to study the effect of the amount of rainfall and the temperature on the yield of a crop and therefore study all three random variables simultaneously. At other times, several independent measurements of the same object may be available as part of an experiment and we may want to combine the various measurements into a single index or function. In all such cases, it becomes essential to know how to operate with many random variables simultaneously. This is done by using *joint distributions*. Joint distributions naturally lead to considerations of *marginal and conditional distributions*. We will study joint, marginal, and conditional distributions for discrete random variables in this chapter. The concepts of joint, marginal, and conditional distributions for continuous random variables are not different, but the techniques are mathematically more sophisticated. The continuous case will be treated in the next chapter.

11.1 Bivariate Joint Distributions and Expectations of Functions

We present the fundamentals of joint distributions of two variables in this section. The concepts in the multivariate case are the same, although the technicalities are somewhat more involved. We will treat the multivariate case in a later section. The idea is that there is still an underlying experiment ξ with an associated sample space Ω. But now we have two or more random variables on the sample space Ω. Random variables being functions on the sample space Ω, we now have multiple functions, say $X(\omega), Y(\omega), \ldots$, etc., on Ω. We want to study their *joint behavior*.

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_11,

Example 11.1 (Coin tossing). Consider the experiment ξ of tossing a fair coin three times. Let X be the number of heads among the first two tosses and Y the number of heads among the last two tosses. If we consider X and Y *individually*, we realize immediately that they are each $Bin(2, .5)$ random variables. But the individual distributions hide part of the full story. For example, if we knew that X was 2, then that would imply that Y must be at least 1. Thus, their joint behavior cannot be fully understood from their individual distributions; we must study their *joint distribution*.

Here is what we mean by their joint distribution. The sample space Ω of this experiment is

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

Each sample point has an equal probability, $\frac{1}{8}$. Denoting the sample points as $\omega_1, \omega_2, \ldots, \omega_8$, we see that if ω_1 prevails, then $X(\omega_1) = Y(\omega_1) = 2$, but if ω_2 prevails, then $X(\omega_2) = 2, Y(\omega_2) = 1$. The combinations of *all* possible values of (X, Y) are

$$(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2).$$

The joint distribution of (X, Y) provides the probability $p(x, y) = P(X = x, Y = y)$ for each such combination of possible values (x, y). Indeed, by direct counting using the eight equally likely sample points, we see that

$$p(0,0) = \frac{1}{8}, p(0,1) = \frac{1}{8}, p(0,2) = 0, p(1,0) = \frac{1}{8}, p(1,1) = \frac{1}{4};$$
$$p(1,2) = \frac{1}{8}, p(2,0) = 0, p(2,1) = \frac{1}{8}, p(2,2) = \frac{1}{8}.$$

For example, why is $p(0,1) = \frac{1}{8}$? This is because the combination $(X = 0, Y = 1)$ is favored by only one sample point, namely TTH. It is convenient to present these nine different probabilities in the form of a table as follows.

	Y		
X	0	1	2
0	$\frac{1}{8}$	$\frac{1}{8}$	0
1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
2	0	$\frac{1}{8}$	$\frac{1}{8}$

Such a layout is a convenient way to present the joint distribution of two discrete random variables with a small number of values. The distribution itself is called *the joint pmf*; here is a formal definition.

Definition 11.1. Let X and Y be two discrete random variables with respective sets of values $x_1, x_2, \ldots,$ and $y_1, y_2, \ldots,$ defined on a common sample space Ω. The *joint pmf* of X, Y is defined to be the function $p(x_i, y_j) = P(X = x_i, Y = y_j), i, j \geq 1$, and $p(x, y) = 0$ at any other point (x, y) in $\mathcal{R}^2$.

The requirements of a joint pmf are that

(i) $p(x, y) \geq 0 \ \forall (x, y)$;
(ii) $\sum_i \sum_j p(x_i, y_j) = 1$.

Thus, if we write the joint pmf in the form of a table, then all entries should be nonnegative and the sum of all the entries in the table should be 1.

As in the case of a single variable, we can define a CDF for more than one variable also. For the case of two variables, here is the definition of a CDF.

Definition 11.2. Let X and Y be two discrete random variables defined on a common sample space Ω. The joint CDF, or simply the CDF, of (X, Y) is a function $F : \mathcal{R}^2 \to [0, 1]$ defined as $F(x, y) = P(X \leq x, Y \leq y), x, y \in \mathcal{R}$.

Like the joint pmf, the CDF also characterizes the joint distribution of two discrete random variables. But it is not very convenient or even interesting to work with the CDF in the case of discrete random variables. It is much preferred to work with the pmf when dealing with discrete random variables.

Example 11.2 (Maximum and Minimum in Dice Rolls). Suppose a fair die is rolled twice, and let X and Y be the larger and the smaller of the two rolls (note that X can be equal to Y), respectively. Each of X and Y takes the individual values $1, 2, \ldots, 6$, but we have necessarily $X \geq Y$. The sample space of this experiment is

$$\{11, 12, 13, \ldots, 64, 65, 66\}.$$

By direct counting, for example, $p(2, 1) = \frac{2}{36}$. Indeed, $p(x, y) = \frac{2}{36}$ for each $x, y = 1, 2, \ldots, 6, x > y$, and $p(x, y) = \frac{1}{36}$ for $x = y = 1, 2, \ldots, 6$. Here is how the joint pmf looks in the form of a table:

			Y			
X	1	2	3	4	5	6
1	$\frac{1}{36}$	0	0	0	0	0
2	$\frac{1}{18}$	$\frac{1}{36}$	0	0	0	0
3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	0	0	0
4	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	0	0
5	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	0
6	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$

The individual pmfs of X, Y are easily recovered from the joint distribution. For example, $P(X = 1) = \sum_{y=1}^{6} P(X = 1, Y = y) = \frac{1}{36}$, and $P(X = 2) = \sum_{y=1}^{6} P(X = 2, Y = y) = \frac{1}{18} + \frac{1}{36} = \frac{1}{12}$, etc. The individual pmfs are obtained by summing the joint probabilities over all values of the other variable. They are:

$$\begin{array}{cccccccc} x & 1 & 2 & 3 & 4 & 5 & 6 \\ p_X(x) & \frac{1}{36} & \frac{3}{36} & \frac{5}{36} & \frac{7}{36} & \frac{9}{36} & \frac{11}{36} \end{array}$$

$$\begin{array}{cccccccc} y & 1 & 2 & 3 & 4 & 5 & 6 \\ p_Y(y) & \frac{11}{36} & \frac{9}{36} & \frac{7}{36} & \frac{5}{36} & \frac{3}{36} & \frac{1}{36} \end{array}$$

From the individual pmf of X, we can find the expectation of X. Indeed, $E(X) = 1 \times \frac{1}{36} + 2 \times \frac{3}{36} + \cdots + 6 \times \frac{11}{36} = \frac{161}{36}$. Similarly, $E(Y) = \frac{91}{36}$. The individual pmfs are called *marginal pmfs*, and here is the formal definition.

Definition 11.3. Let $p(x, y)$ be the joint pmf of (X, Y). The *marginal pmf* of a function $Z = g(X, Y)$ is defined as $p_Z(z) = \sum_{(x,y):g(x,y)=z} p(x, y)$. In particular,

$$p_X(x) = \sum_y p(x, y); \;\; p_Y(y) = \sum_x p(x, y),$$

and for any event A,

$$P(A) = \sum_{(x,y)\in A} p(x, y).$$

Here is another example.

Example 11.3 (Bridge Hands). Let X be the number of aces in the hands of North and Y the number of aces in the hands of South in a bridge game. Then, $0 \le X$ and Y, and $X + Y \le 4$. If North gets x aces, South can get y aces in $\binom{4-x}{y}$ ways. Also, North has to get $13 - x$ non-ace cards and South has to get $13 - y$ non-ace cards. Thus,

$$p(x, y) = \frac{\binom{4}{x}\binom{48}{13-x}\binom{4-x}{y}\binom{35+x}{13-y}}{\binom{52}{13}\binom{39}{13}},$$
$$x, y \ge 0, x + y \le 4.$$

For example,

$$p(1, 0) = .1249; p(1, 1) = .2029; p(1, 2) = .0974; p(1, 3) = .0137; p(1, 4) = 0.$$

Summing, we get

$$P(X = 1) = \sum_y p(1, y) = .4389,$$

which is what we get from the direct formula for the pmf of X:

$$P(X = 1) = \frac{\binom{4}{1}\binom{48}{12}}{\binom{52}{13}} = .4389.$$

Likewise,

$$p(2,0) = .0936;\ p(2,1) = .0974;\ p(2,2) = .0225,$$

and adding, we get $P(X = 2) = .2135$, which is what we get directly from the formula

$$P(X = 2) = \frac{\binom{4}{2}\binom{48}{11}}{\binom{52}{13}} = .2135.$$

Now suppose we want to find the probability of the event A that $X + Y = 2$. Then,

$$P(A) = \sum_{(x,y)\in A} p(x,y) = p(0,2) + p(1,1) + p(2,0)$$

$$= .0936 + .2029 + .0936 = .3901.$$

There is no way to compute the probability of the event A except by using the joint distribution of X and Y and by adding up the probabilities of all the favorable combinations (x, y) for the event A. This exemplifies the importance of studying joint distributions, which carry all the information about X and Y, while the marginal distributions, in general, do not.

Example 11.4. Consider a joint pmf given by the formula

$$p(x,y) = c(x+y), 1 \le x, y \le n,$$

where c is a normalizing constant.

First of all, we need to evaluate c by equating

$$\sum_{x=1}^{n}\sum_{y=1}^{n} p(x,y) = 1$$

$$\Leftrightarrow c\sum_{x=1}^{n}\sum_{y=1}^{n}(x+y) = 1$$

$$\Leftrightarrow c\sum_{x=1}^{n}\left[nx + \frac{n(n+1)}{2}\right] = 1$$

$$\Leftrightarrow c\left[\frac{n^2(n+1)}{2} + \frac{n^2(n+1)}{2}\right] = 1$$

$$\Leftrightarrow cn^2(n+1) = 1$$

$$\Leftrightarrow c = \frac{1}{n^2(n+1)}.$$

The joint pmf is symmetric between x and y (since $x + y = y + x$), so X and Y have the same marginal pmf. For example, X has the pmf

$$\begin{aligned} p_X(x) &= \sum_{y=1}^{n} p(x,y) = \frac{1}{n^2(n+1)} \sum_{y=1}^{n} (x+y) \\ &= \frac{1}{n^2(n+1)} \left[nx + \frac{n(n+1)}{2} \right] \\ &= \frac{x}{n(n+1)} + \frac{1}{2n}, 1 \le x \le n. \end{aligned}$$

Suppose now that we want to compute $P(X > Y)$. This can be found by summing $p(x, y)$ over all combinations for which $x > y$. But this longer calculation can be avoided by using a *symmetry argument* that is often very useful. Note that because the joint pmf is symmetric between x and y, we must have $P(X > Y) = P(Y > X) = p$ (say). But also

$$\begin{aligned} &P(X > Y) + P(Y > X) + P(X = Y) = 1 \Rightarrow 2p + P(X = Y) = 1 \\ &\Rightarrow p = \frac{1 - P(X = Y)}{2}. \end{aligned}$$

Now,

$$\begin{aligned} P(X = Y) &= \sum_{x=1}^{n} p(x,x) = c \times \sum_{x=1}^{n} 2x \\ &= \frac{1}{n^2(n+1)} n(n+1) = \frac{1}{n}. \end{aligned}$$

Therefore, $P(X > Y) = p = \frac{n-1}{2n} \approx \frac{1}{2}$ for large n.

Example 11.5 (Dice Rolls Revisited). Consider again the example of two rolls of a fair die, and suppose X and Y are the larger and the smaller of the two rolls, respectively. We have worked out the joint distribution of (X, Y) in Example 11.2. Suppose we want to find the distribution of the difference, $X - Y$. The possible values of $X - Y$ are $0, 1, \ldots, 5$, and we find $P(X - Y = k)$ by using the joint distribution of (X, Y):

$$P(X - Y = 0) = p(1,1) + p(2,2) + \cdots + p(6,6) = \frac{1}{6};$$

$$P(X - Y = 1) = p(2,1) + p(3,2) + \cdots + p(6,5) = \frac{5}{18};$$

$$P(X - Y = 2) = p(3,1) + p(4,2) + p(5,3) + p(6,4) = \frac{2}{9};$$

$$P(X-Y=3) = p(4,1) + p(5,2) + p(6,3) = \frac{1}{6};$$
$$P(X-Y=4) = p(5,1) + p(6,2) = \frac{1}{9};$$
$$P(X-Y=5) = p(6,1) = \frac{1}{18}.$$

Again, there is no way to find the distribution of $X-Y$ except by using the joint distribution of (X,Y).

Suppose now that we also want to know the expected value of $X-Y$. Now that we have the distribution of $X-Y$ worked out, we can find the expectation by directly using the definition of expectation:

$$\begin{aligned} E(X-Y) &= \sum_{k=0}^{5} kP(X-Y=k) \\ &= \frac{5}{18} + \frac{4}{9} + \frac{1}{2} + \frac{4}{9} + \frac{5}{18} = \frac{35}{18}. \end{aligned}$$

But we can also use linearity of expectations and find $E(X-Y)$ as

$$E(X-Y) = E(X) - E(Y) = \frac{161}{36} - \frac{91}{36} = \frac{35}{18}$$

(see Example 11.2 for $E(X)$, $E(Y)$).

A third possible way to compute $E(X-Y)$ is to treat $X-Y$ as a function of (X,Y) and *use the joint pmf of* (X,Y) to find $E(X-Y)$ as $\sum_x \sum_y (x-y)p(x,y)$. In this particular example, this will be an unnecessarily laborious calculation because luckily we can find $E(X-Y)$ by other quicker means in this example, as we just saw. But in general one has to resort to the joint pmf to calculate the expectation of a function of (X,Y). Here is the formal formula.

Theorem 11.1 (Expectation of a Function). *Let (X,Y) have the joint pmf $p(x,y)$ and let $g(X,Y)$ be a function of (X,Y). We say that the expectation of $g(X,Y)$ exists if $\sum_x \sum_y |g(x,y)|p(x,y) < \infty$, in which case*

$$E[g(X,Y)] = \sum_x \sum_y g(x,y)p(x,y).$$

Example 11.6. Consider the example of three tosses of a fair coin, and let X and Y be the number of heads in the first two and the last two tosses, respectively. Let $g(X,Y) = |X-Y|$. We want to find the expectation of $g(X,Y)$. Because of the absolute value, we cannot find this expectation from the marginal distributions of X and Y; we must use the joint pmf in this case.

Using the joint pmf of (X, Y) from Example 11.1,

$$\begin{aligned} E(|X-Y|) &= \sum_{x=0}^{2}\sum_{y=0}^{2}|x-y|p(x,y) \\ &= 1\times[p(0,1)+p(1,0)+p(1,2)+p(2,1)]+2 \\ &\quad \times[p(0,2)+p(2,0)] = \frac{4}{8} = \frac{1}{2}. \end{aligned}$$

How about $E[\max\{X, Y\}]$? Again, this can only be found from the joint pmf of (X, Y). By using the joint pmf,

$$\begin{aligned} E[\max\{X,Y\}] &= (p(0,1)+p(1,0)+p(1,1))+(2p(1,2)+2p(2,1)+2p(2,2)) \\ &= \frac{1}{4}+\frac{1}{4}+\frac{3}{4} = \frac{5}{4} = 1.25. \end{aligned}$$

Thus, each of $E(X)$ and $E(Y)$ is one, but the expectation of the maximum of X, Y is bigger than one:

$$E[\max\{X,Y\}] > \max\{E(X), E(Y)\}.$$

11.2 Conditional Distributions and Conditional Expectations

Sometimes we want to know the expected value of one of the variables, say X, if we knew the value of the other variable Y. For example, in the die-tossing experiment above, what should we expect the larger of the two rolls to be if the smaller roll is known to be 2?

To answer this question, we have to find the probabilities of the various values of X, *conditional on* knowing that Y equals some given y, and then average by using these conditional probabilities. Here are the formal definitions.

Definition 11.4 (Conditional Distribution). Let (X, Y) have the joint pmf $p(x, y)$. The *conditional distribution of X given $Y = y$* is defined to be

$$p(x|y) = P(X = x|Y = y) = \frac{p(x,y)}{p_Y(y)},$$

and the *conditional expectation of X given $Y = y$* is defined to be

$$E(X|Y=y) = \sum_x xp(x|y) = \frac{\sum_x xp(x,y)}{p_Y(y)} = \frac{\sum_x xp(x,y)}{\sum_x p(x,y)}.$$

The conditional distribution of Y given $X = x$ and the conditional expectation of Y given $X = x$ are defined analogously by switching the roles of X and Y in the definitions above .

We often casually write $E(X|y)$ to mean $E(X|Y = y)$. Two easy facts that are nevertheless often useful are the following.

Proposition 11.1. *Let X and Y be random variables defined on a common sample space Ω. Then,*

(a) $E(g(Y)|Y = y) = g(y), \ \forall y, \quad$ for any function g;
(b) $E(Xg(Y)|Y = y) = g(y)E(X|Y = y) \ \forall y, \quad$ for any function g.

Recall that in Chapter 4 we defined two random variables to be independent if $P(X \le x, Y \le y) = P(X \le x)P(Y \le y) \ \forall x, y \in \mathcal{R}$. This is of course a correct definition, but in the case of discrete random variables, it is more convenient to think of independence in terms of the pmf. The definition below puts together some equivalent definitions of independence of two discrete random variables.

Definition 11.5 (Independence). Let (X, Y) have the joint pmf $p(x, y)$. Then X and Y are said to be independent if

$$\begin{aligned} & p(x|y) = p_X(x), \ \forall x, y \text{ such that } p_Y(y) > 0; \\ \Leftrightarrow \ & p(y|x) = p_Y(y), \ \forall x, y \text{ such that } p_X(x) > 0; \\ \Leftrightarrow \ & p(x, y) = p_X(x)p_Y(y), \forall x, y; \\ \Leftrightarrow \ & P(X \le x, Y \le y) = P(X \le x)P(Y \le y) \ \forall x, y. \end{aligned}$$

The third equivalent condition in the list above is usually the most convenient one to verify and use.

One more frequently useful fact about conditional expectations is the following.

Proposition 11.2. *Suppose X and Y are independent random variables. Then, for any function $g(X)$ such that the expectations below exist, and for any y,*

$$E[g(X)|Y = y] = E[g(X)].$$

11.2.1 Examples on Conditional Distributions and Expectations

Example 11.7. In the experiment of three tosses of a fair coin, we have worked out the joint mass function of X, Y, where X is the number of heads in the first two tosses and Y the number of heads in the last two tosses. Using this joint mass function, we now find

$$\begin{aligned} P(X = 0|Y = 0) &= \frac{p(0,0)}{p_Y(0)} = \frac{1/8}{1/4} = \frac{1}{2}; \\ P(X = 1|Y = 0) &= \frac{p(1,0)}{p_Y(0)} = \frac{1/8}{1/4} = \frac{1}{2}; \\ P(X = 2|Y = 0) &= \frac{p(2,0)}{p_Y(0)} = \frac{0}{1/4} = 0. \end{aligned}$$

That is, the conditional distribution of X given $Y = 0$ is a *two-point distribution*, although X by itself takes three values. We can also similarly find

$$P(Y = 0|X = 0) = \frac{p(0,0)}{p_X(0)} = \frac{1/8}{1/4} = \frac{1}{2};$$
$$P(Y = 1|X = 0) = \frac{p(0,1)}{p_X(0)} = \frac{1/8}{1/4} = \frac{1}{2};$$
$$P(Y = 2|X = 0) = \frac{p(0,2)}{p_X(0)} = \frac{0}{1/4} = 0.$$

Thus, the conditional distribution of Y given $X = 0$ is also a two-point distribution, and in fact, as distributions, the two conditional distributions that we worked out in this example are the same.

Example 11.8 (Maximum and Minimum in Dice Rolls). In the experiment of two rolls of a fair die, we have worked out the joint distribution of X, Y, where X is the larger and Y the smaller of the two rolls. Using this joint distribution, we can now find the conditional distributions. For instance,

$$P(Y = 1|X = 1) = 1; P(Y = y|X = 1) = 0 \text{ if } y > 1;$$
$$P(Y = 1|X = 2) = \frac{1/18}{1/18 + 1/36} = \frac{2}{3};$$
$$P(Y = 2|X = 2) = \frac{1/36}{1/18 + 1/36} = \frac{1}{3};$$
$$P(Y = y|X = 2) = 0 \text{ if } y > 2;$$
$$P(Y = y|X = 6) = \frac{1/18}{5/18 + 1/36} = \frac{2}{11} \text{ if } 1 \le y \le 5;$$
$$P(Y = 6|X = 6) = \frac{1/36}{5/18 + 1/36} = \frac{1}{11}.$$

Example 11.9 (Conditional Expectation in a 2×2 Table). Suppose X and Y are binary variables, each taking only the values $0, 1$ with the following joint distribution.

	Y	
X	0	1
0	s	t
1	u	v

We want to evaluate the conditional expectation of X given $Y = 0, 1$, respectively. By using the definition of conditional expectation,

$$E(X|Y = 0) = \frac{0 \times p(0,0) + 1 \times p(1,0)}{p(0,0) + p(1,0)} = \frac{u}{s+u};$$
$$E(X|Y = 1) = \frac{0 \times p(0,1) + 1 \times p(1,1)}{p(0,1) + p(1,1)} = \frac{v}{t+v}.$$

Therefore,

$$E(X|Y=1) - E(X|Y=0) = \frac{v}{t+v} - \frac{u}{s+u} = \frac{vs-ut}{(t+v)(s+u)}.$$

It follows that we can now have the single formula

$$E(X|Y=y) = \frac{u}{s+u} + \frac{vs-ut}{(t+v)(s+u)}y,$$

$y = 0, 1$. We now realize that the conditional expectation of X given $Y = y$ is a linear function of y in this example. This will be the case whenever both X and Y are binary variables, as they were in this example.

Example 11.10 (Conditional Expectation of Number of Aces). Consider again the example of the number of aces X, Y in the hands of North and South in a bridge game. We want to find $E(X|Y = y)$ for $y = 0, 1, 2, 3, 4$. Of these, note that $E(X|Y = 4) = 0$.

For the rest, from the definition,

$$E(X|Y=y) = \frac{\sum_x xp(x,y)}{\sum_x p(x,y)} = \frac{\sum_{x=0}^{4-y} xp(x,y)}{\sum_{x=0}^{4-y} p(x,y)},$$

where $p(x,y) = \frac{\binom{4}{x}\binom{48}{13-x}\binom{4-x}{y}\binom{35+x}{13-y}}{\binom{52}{13}\binom{39}{13}}$ from Example 11.3.

For example,

$$E(X|Y=2) = \frac{0 \times p(0,2) + 1 \times p(1,2) + 2 \times p(2,2)}{p(0,2) + p(1,2) + p(2,2)}$$

$$= \frac{.0974 + 2 \times .0225}{.0936 + .0974 + .0225} = .67.$$

Note that the .67 value is actually $\frac{2}{3}$, and this makes intuitive sense. If South already has two aces, then the remaining two aces should be divided among East, West, and North equitably, which would give $E(X|Y = 2)$ as $\frac{2}{3}$.

Example 11.11 (Conditional Expectation in Dice Experiment). Consider again the example of the joint distribution of the maximum and the minimum of two rolls of a fair die. Let X denote the maximum and Y the minimum. We will find $E(X|Y = y)$ for various values of y.

By using the definition of $E(X|Y = y)$, we have, for example,

$$E(X|Y=1) = \frac{1 \times \frac{1}{36} + \frac{1}{18}[2 + \cdots + 6]}{\frac{1}{36} + \frac{5}{18}} = \frac{41}{11} = 3.73,$$

as another example

$$E(X|Y=3) = \frac{3 \times \frac{1}{36} + \frac{1}{18} \times 15}{\frac{1}{36} + \frac{3}{18}} = \frac{33}{7} = 4.71,$$

and

$$E(X|Y=5) = \frac{5 \times \frac{1}{36} + 6 \times \frac{1}{18}}{\frac{1}{36} + \frac{1}{18}} = \frac{17}{3} = 5.77.$$

We notice that $E(X|Y=5) > E(X|Y=3) > E(X|Y=1)$; in fact, it is true that $E(X|Y=y)$ is increasing in y in this example. Again, it does make intuitive sense.

Just as in the case of a distribution of a single variable, we often also want a measure of variability in addition to a measure of average for conditional distributions. This motivates defining a *conditional variance*.

Definition 11.6 (Conditional Variance). Let (X, Y) have the joint pmf $p(x, y)$. Let $\mu_X(y) = E(X|Y=y)$. The *conditional variance* of X given $Y=y$ is defined to be

$$\text{Var}(X|Y=y) = E[(X-\mu_X(y))^2|Y=y] = \sum_x (x-\mu_X(y))^2 p(x|y).$$

We often casually write $\text{Var}(X|y)$ to mean $\text{Var}(X|Y=y)$.

Example 11.12 (Conditional Variance in Dice Experiment). We will work out the conditional variance of the maximum of two rolls of a die given the minimum. That is, suppose a fair die is rolled twice and X and Y are the larger and the smaller of the two rolls respectively; we want to compute $\text{Var}(X|y)$.

For example, if $y=3$, then $\mu_X(y) = E(X|Y=y) = E(X|Y=3) = 4.71$ (see the previous example). Therefore,

$$\begin{aligned}
\text{Var}(X|y) &= \sum_x (x-4.71)^2 p(x|3) \\
&= \frac{(3-4.71)^2 \times \frac{1}{36} + (4-4.71)^2 \times \frac{1}{18} + (5-4.71)^2 \times \frac{1}{18} + (6--4.71)^2 \times \frac{1}{18}}{\frac{1}{36} + \frac{1}{18} + \frac{1}{18} + \frac{1}{18}} \\
&= 1.06.
\end{aligned}$$

To summarize, given that the minimum of two rolls of a fair die is 3, the expected value of the maximum is 4.71 and the variance of the maximum is 1.06.

These two values, $E(X|y)$ and Var(X|y), change as we change the given value y. Thus, E(X|y) and Var(X|y) are functions of y and, for each separate y, a new

calculation is needed. If X and Y happen to be independent, then of course whatever be y, E(X|y) = E(X) and Var(X|y) = Var(X).

The next result is an important one in many applications.

Theorem 11.2 (Poisson Conditional Distribution). *Let X and Y be independent Poisson random variables with means λ, μ. Then the conditional distribution of X given $X + Y = t$ is $Bin(t, p)$, where $p = \frac{\lambda}{\lambda+\mu}$.*

Proof. Clearly, $P(X = x|X + Y = t) = 0\,\forall x > t$. For $x \le t$,

$$\begin{aligned} P(X = x|X + Y = t) &= \frac{P(X = x, X + Y = t)}{P(X + Y = t)} \\ &= \frac{P(X = x, Y = t - x)}{P(X + Y = t)} \\ &= \frac{e^{-\lambda}\lambda^x}{x!}\frac{e^{-\mu}\mu^{t-x}}{(t-x)!}\frac{t!}{e^{-(\lambda+\mu)}(\lambda+\mu)^t} \end{aligned}$$

(on using the fact that $X + Y \sim Poi(\lambda + \mu)$; see Chapter 6)

$$\begin{aligned} &= \frac{t!}{x!(t-x)!}\frac{\lambda^x\mu^{t-x}}{(\lambda+\mu)^t} \\ &= \binom{t}{x}\left(\frac{\lambda}{\lambda+\mu}\right)^x\left(\frac{\mu}{\lambda+\mu}\right)^{t-x}, \end{aligned}$$

which is the pmf of the $Bin(t, \frac{\lambda}{\lambda+\mu})$ distribution.

11.3 Using Conditioning to Evaluate Mean and Variance

Conditioning is often an extremely effective tool for calculating probabilities, means, and variances of random variables with a complex or clumsy joint distribution. Thus, in order to calculate the mean of a random variable X, it is sometimes very convenient to follow an *iterative process* whereby we first evaluate the mean of X *after conditioning on the value y of some suitable random variable Y* and then average over y. The random variable Y has to be chosen judiciously but is often clear from the context of the specific problem. Here are the precise results on how this technique works; it is important to note that the next two results hold for any kind of random variable, not just discrete ones.

Theorem 11.3 (Iterated Expectation Formula). *Let X and Y be random variables defined on the same probability space Ω. Suppose $E(X)$ and $E(X|Y = y)$ exist for each y. Then,*

$$E(X) = E_Y[E(X|Y = y)];$$

thus, in the discrete case,

$$E(X) = \sum_y \mu_X(y) p_Y(y),$$

where $\mu_X(y) = E(X|Y = y)$.

Proof. We prove this for the discrete case. By the definition of conditional expectation,

$$\mu_X(y) = \frac{\sum_x xp(x, y)}{p_Y(y)}$$

$$\Rightarrow \sum_y \mu_X(y) p_Y(y) = \sum_y \sum_x xp(x, y) = \sum_x \sum_y xp(x, y)$$

$$= \sum_x x \sum_y p(x, y) = \sum_x xp_X(x) = E(X).$$

The corresponding variance calculation formula is the following. The proof of this uses the iterated mean formula above and applies it to $(X - \mu_X)^2$.

Theorem 11.4 (Iterated Variance Formula). *Let X and Y be random variables defined on the same probability space Ω. Suppose* $\text{Var}(X)$ *and* $\text{Var}(X|Y = y)$ *exist for each y. Then,*

$$\text{Var}(X) = E_Y[\text{Var}(X|Y = y)] + \text{Var}_Y[E(X|Y = y)].$$

Remark. These two formulas for iterated expectation and iterated variance are valid for all types of variables, not just the discrete ones. Thus, these same formulas will still hold when we discuss joint distributions for continuous random variables in the next chapter.

Some operational formulas that one should be familiar with are summarized below.

Conditional Expectation and Variance Rules

$$E(g(X)|X = x) = g(x); E(g(X)h(Y)|Y = y) = h(y)E(g(X)|Y = y);$$

$E(g(X)|Y = y) = E(g(X))$ if X and Y are independent;

$$\text{Var}(g(X)|X = x) = 0; \text{Var}(g(X)h(Y)|Y = y) = h^2(y)\text{Var}(g(X)|Y = y);$$

$\text{Var}(g(X)|Y = y) = \text{Var}(g(X))$ if X and Y are independent.

Let us see some applications of the two iterated expectation and iterated variance formulas.

Example 11.13 (A Two-Stage Experiment). Suppose n fair dice are rolled. Those that show a six are rolled again. What are the mean and the variance of the number of sixes obtained in the second round of this experiment?

Define Y to be the number of dice in the first round that show a six and X the number of dice in the second round that show a six. Given $Y = y$, $X \sim Bin(y, \frac{1}{6})$ and Y itself is distributed as $Bin(n, \frac{1}{6})$. Therefore,

$$E(X) = E[E(X|Y = y)] = E_Y\left[\frac{y}{6}\right] = \frac{n}{36}.$$

Also,

$$\begin{aligned}\mathrm{Var}(X) &= E_Y[\mathrm{Var}(X|Y = y)] + \mathrm{Var}_Y[E(X|Y = y)] \\ &= E_Y\left[y\frac{1}{6}\frac{5}{6}\right] + \mathrm{Var}_Y\left[\frac{y}{6}\right] \\ &= \frac{5}{36}\frac{n}{6} + \frac{1}{36}n\frac{1}{6}\frac{5}{6} \\ &= \frac{5n}{216} + \frac{5n}{1296} = \frac{35n}{1296}.\end{aligned}$$

Example 11.14. Suppose that in a certain population 30% of couples have one child, 50% have two children, and 20% have three children. One family is picked at random from this population. What is the expected number of boys in this family?

Let Y denote the number of children in the family that was picked, and let X be the number of boys it has. Making the usual assumption of a childbirth being equally likely to be a boy or a girl,

$$E(X) = E_Y[E(X|Y = y)] = .3 \times .5 + .5 \times 1 + .2 \times 1.5 = .95.$$

Example 11.15. Suppose a chicken lays a Poisson number of eggs per week with mean λ. Each egg, independently of the others, has a probability p of being fertilized. We want to find the mean and the variance of the number of eggs fertilized in a week.

Let N denote the number of eggs hatched and X the number of eggs fertilized. Then, $N \sim Poi(\lambda)$, and given $N = n$, $X \sim Bin(n, p)$. Therefore,

$$E(X) = E_N[E(X|N = n)] = E_N[np] = p\lambda$$

and

$$\begin{aligned}\mathrm{Var}(X) &= E_N[\mathrm{Var}(X|N = n)] + \mathrm{Var}_N(E(X|N = n) \\ &= E_N[np(1 - p)] + \mathrm{Var}_N(np) = \lambda p(1 - p) + p^2\lambda = p\lambda.\end{aligned}$$

Interestingly, the number of eggs actually fertilized has the same mean and variance $p\lambda$. (Can you see why?)

Remark. In all of these examples, it was important to choose the variable Y on which one should condition wisely. The efficiency of the technique depends on this very crucially.

Sometimes, a formal generalization of the iterated expectation formula when a third variable Z is present is useful. It is particularly useful in hierarchical statistical modeling of distributions, where an ultimate marginal distribution for some X is constructed by first conditioning on a number of auxiliary variables and then gradually unconditioning them. We state the more general iterated expectation formula; its proof is similar to that of the usual iterated expectation formula.

Theorem 11.5 (Higher-Order Iterated Expectation). *Let X, Y, Z be random variables defined on the same sample space Ω. Assume that each conditional expectation below and the marginal expectation $E(X)$ exist. Then,*

$$E(X) = E_Y[E_{Z|Y}\{E(X|Y = y, Z = z)\}].$$

11.4 Covariance and Correlation

We know that variance is additive for independent random variables; i.e., if $X_1, X_2, \ldots, X_n$ are independent random variables, then $\text{Var}(X_1 + X_2 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n)$. In particular, for two independent random variables X, Y, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. However, *in general*, variance is not additive. Let us do the general calculation for $\text{Var}(X + Y)$:

$$\begin{aligned}
\text{Var}(X + Y) &= E(X + Y)^2 - [E(X + Y)]^2 \\
&= E(X^2 + Y^2 + 2XY) - [E(X) + E(Y)]^2 \\
&= E(X^2) + E(Y^2) + 2E(XY) - [E(X)]^2 - [E(Y)]^2 - 2E(X)E(Y) \\
&= E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2 + 2[E(XY) - E(X)E(Y)] \\
&= \text{Var(X)} + \text{Var(Y)} + 2[\text{E(XY)} - \text{E(X)E(Y)}].
\end{aligned}$$

We thus have the extra term $2[E(XY) - E(X)E(Y)]$ in the expression for $\text{Var}(X + Y)$; of course, when X and Y are independent, $E(XY) = E(X)E(Y)$, so the extra term drops out. But, in general, one has to keep the extra term. The quantity $E(XY) - E(X)E(Y)$ is called the *covariance* of X and Y.

Definition 11.7 (Covariance). Let X and Y be two random variables defined on a common sample space Ω such that $E(XY)$, $E(X)$, $E(Y)$ all exist. The *covariance* of X and Y is defined as

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E[(X - E(X))(Y - E(Y))].$$

Remark. Covariance is a measure of whether two random variables X and Y tend to increase or decrease together. If a larger value of X generally causes an increment

in the value of Y, then often (but not always) they have a positive covariance. For example, taller people tend to weigh more than shorter people, and height and weight usually have a positive covariance.

Unfortunately, however, covariance can take arbitrary positive and negative values. Therefore, by looking at its value in a particular problem, we cannot judge whether it is a large value or not. We cannot compare a covariance with a standard to judge if it is large or small. A renormalization of the covariance cures this problem and calibrates it to a scale of -1 to $+1$. We can judge such a quantity as large, small, or moderate; for example, .95 would be large positive, .5 moderate, and .1 small. The renormalized quantity is the *correlation coefficient* or simply the *correlation* between X and Y.

Definition 11.8 (Correlation). Let X and Y be two random variables defined on a common sample space Ω such that $\text{Var}(X)$ and $\text{Var}(Y)$ are both finite. The *correlation* between X and Y is defined to be

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}.$$

Some important properties of covariance and correlation are put together in the next theorem.

Theorem 11.6 (Properties of Covariance and Correlation). *Provided that the required variances and covariances exist,*

(a) $\text{Cov}(X,c) = 0$ *for any* X *and any constant* c,

(b) $\text{Cov}(X,X) = \text{var}(X)$ *for any* X,

(c) $$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \,\text{Cov}(X_i, Y_j),$$

and, in particular,

$$\begin{aligned}\text{Var}(aX+bY) &= \text{Cov}(aX+bY, aX+bY)\\ &= a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y)\end{aligned}$$

and

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\text{Var}(X_i) + 2\sum\sum_{i<j=1}^{n}\text{Cov}(X_i, X_j).$$

(d) *For any two independent random variables X and Y,* $\text{Cov}\,(X,Y) = \rho_{X,Y} = 0$.

(e) $\rho_{a+bX,c+dY} = \text{sgn}(bd)\rho_{X,Y}$, *where* $\text{sgn}(bd) = 1$ *if* $bd > 0$ *and* $\text{sgn}(bd) = -1$ *if* $bd < 0$.

(f) *Whenever* $\rho_{X,Y}$ *is defined,* $-1 \le \rho_{X,Y} \le 1$.

(g) $\rho_{X,Y} = 1$ *if and only if for some* a *and some* $b > 0$, $P(Y = a + bX) = 1$*; and* $\rho_{X,Y} = -1$ *if and only if for some* a *and some* $b < 0$, $P(Y = a + bX) = 1$.

Proof. For part (a), $\text{Cov}(X, c) = E(cX) - E(c)E(X) = cE(X) - cE(X) = 0$. For part (b), $\text{Cov}(X, X) = E(X^2) - [E(X)]^2 = \text{Var}(X)$. For part (c),

$$\begin{aligned}
\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) &= E\left[\sum_{i=1}^{n} a_i X_i \times \sum_{j=1}^{m} b_j Y_j\right] \\
&\quad - E\left(\sum_{i=1}^{n} a_i X_i\right) E\left(\sum_{j=1}^{m} b_j Y_j\right) \\
&= E\left(\sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j X_i Y_j\right) - \left[\sum_{i=1}^{n} a_i E(X_i)\right] \\
&\quad \times \left[\sum_{j=1}^{m} b_j E(Y_j)\right] = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j E(X_i, Y_j) \\
&\quad - \sum_{i=1}^{n} a_i \sum_{j=1}^{m} b_j E(X_i)E(Y_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j [E(X_i, Y_j) - E(X_i)E(Y_j)] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j).
\end{aligned}$$

Part (d) follows on noting that $E(XY) = E(X)E(Y)$ if X and Y are independent. For part (e), first note that $\text{Cov}(a + bX, c + dY) = bd\,\text{Cov}(X, Y)$ by using part (a) and part (c). Also, $\text{Var}(a + bX) = b^2\text{Var}(X)$ and $\text{Var}(c + dY) = d^2\text{Var}(Y)$

$$\begin{aligned}
\Rightarrow \rho_{a+bX,c+dY} &= \frac{bd\,\text{Cov}(X, Y)}{\sqrt{b^2\text{Var}(X)}\sqrt{d^2\text{Var}(Y)}} = \frac{bd\,\text{Cov}(X, Y)}{|b|\sqrt{\text{Var}(X)}|d|\sqrt{\text{Var}(Y)}} \\
&= \frac{bd}{|bd|}\rho_{X,Y} = \text{sgn}(bd)\rho_{X,Y}.
\end{aligned}$$

The proof of part (f) uses the Cauchy-Schwartz inequality (see Chapter 4) that for any two random variables U and V, $[E(UV)]^2 \le E(U^2)E(V^2)$. Let $U = \frac{X-E(X)}{\sqrt{\text{Var(X)}}}$, $V = \frac{Y-E(Y)}{\sqrt{\text{Var(Y)}}}$. Then, $E(U^2) = E(V^2) = 1$ and

$$\rho_{X,Y} = E(UV) \le E(U^2)E(V^2) = 1.$$

The lower bound $\rho_{X,Y} \ge -1$ follows similarly.

Part (g) uses the condition for equality in the Cauchy-Schwartz inequality; i.e., that in order that $\rho_{X,Y} = \pm 1$, one must have $[E(UV)]^2 = E(U^2)E(V^2)$ in the argument above, which implies the statement in part (g).

Example 11.16 (Correlation between Minimum and Maximum in Dice Rolls). Consider again the experiment of rolling a fair die twice, and let X and Y be the maximum and the minimum of the two rolls respectively. We want to find the correlation between X and Y.

The joint distribution of (X, Y) was worked out in Example 11.2. From the joint distribution,

$$E(XY) = 1/36+2/18+4/36+3/18+6/18+9/36+\cdots+30/18+36/36 = 49/4.$$

The marginal pmfs of X, Y were also worked out in Example 11.2. From the marginal pmfs, by direct calculation, $E(X) = 161/36, E(Y) = 91/36$, and $\text{Var}(X) = \text{Var}(Y) = 2555/1296$. Therefore,

$$\begin{aligned}\rho_{X,Y} &= \frac{E(XY) - E(X)E(Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \\ &= \frac{49/4 - 161/36 \times 91/36}{2555/1296} = \frac{35}{73} = .48.\end{aligned}$$

The correlation between the maximum and the minimum is in fact positive for *any* number of rolls of a die, although the correlation will converge to zero when the number of rolls converges to ∞.

Example 11.17 (Correlation in the Chicken-Eggs Example). Consider again the example of a chicken laying a Poisson number of eggs, N, with mean λ and each egg fertilizing, independently of others, with probability p. If X is the number of eggs actually fertilized, we want to find the correlation between the number of eggs laid and the number fertlized; i.e., the correlation between X and N.

First,

$$\begin{aligned}E(XN) &= E_N[E(XN|N = n)] = E_N[nE(X|N = n)] \\ &= E_N[n^2 p] = p(\lambda + \lambda^2).\end{aligned}$$

Next, from our previous calculations, $E(X) = p\lambda, E(N) = \lambda$, $\text{Var}(X) = p\lambda$, and $\text{Var}(N) = \lambda$. Therefore,

$$\begin{aligned}\rho_{X,N} &= \frac{E(XN) - E(X)E(N)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(N)}} \\ &= \frac{p(\lambda + \lambda^2) - p\lambda^2}{\sqrt{p\lambda}\sqrt{\lambda}} = \sqrt{p}.\end{aligned}$$

Thus, the correlation goes up with the fertility rate of the eggs.

Example 11.18 (Best Linear Predictor). Suppose X and Y are two jointly distributed random variables and either by necessity or omission the variable Y was not observed. But X was observed, and there may be some information in the X value about Y. The problem is to predict Y by using X. Linear predictors, because of their functional simplicity, are appealing. The mathematical problem is to choose the *best linear predictor* $a + bX$ of Y, where best is defined as the predictor that minimizes *the mean squared error* $E[Y - (a + bX)]^2$. We will see that the answer has something to do with the covariance between X and Y.

By breaking the square,

$$R(a,b) = E[Y-(a+bX)]^2 = a^2 + b^2E(X^2) + 2abE(X) - 2aE(Y) \\ -2bE(XY) + E(Y^2).$$

To minimize this with respect to a, b, we partially differentiate $R(a,b)$ with respect to a, b and set the derivatives equal to zero:

$$\begin{aligned} \frac{\partial}{\partial a}R(a,b) &= 2a + 2bE(X) - 2E(Y) = 0 \\ &\Leftrightarrow a + bE(X) = E(Y); \\ \frac{\partial}{\partial b}R(a,b) &= 2bE(X^2) + 2aE(X) - 2E(XY) = 0 \\ &\Leftrightarrow aE(X) + bE(X^2) = E(XY). \end{aligned}$$

Solving these two equations simultaneously, we get

$$b = \frac{E(XY) - E(X)E(Y)}{\mathrm{Var}(X)}, a = E(Y) - \frac{E(XY) - E(X)E(Y)}{\mathrm{Var}(X)}E(X).$$

These values do minimize $R(a,b)$ by an easy application of the second derivative test. So, the best linear predictor of Y based on X is

$$\begin{aligned} \text{best linear predictor of } Y &= E(Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}E(X) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}X \\ &= E(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}[X - E(X)]. \end{aligned}$$

The best linear predictor is also known as *the regression line of Y on X*. It is of widespread use in statistics.

Example 11.19 (Zero Correlation Does Not Mean Independence). If X and Y are independent, then necessarily $\mathrm{Cov}(X,Y) = 0$, and hence the correlation is also zero. The converse is not true. Take a three-valued random variable X with the pmf $P(X = \pm 1) = p, P(X = 0) = 1 - 2p, 0 < p < \frac{1}{2}$. Let the other variable

Y be $Y = X^2$. Then, $E(XY) = E(X^3) = 0$ and $E(X)E(Y) = 0$ because $E(X) = 0$. Therefore, $\text{Cov}(X, Y) = 0$. But X and Y are certainly not independent; e.g., $P(Y = 0|X = 0) = 1$, but $P(Y = 0) = 1 - 2p \neq 0$.

Indeed, if X has a distribution symmetric around zero and has three finite moments, then X and X^2 always have a zero correlation, although they are not independent.

11.5 Multivariate Case

The extension of the concepts for the bivariate discrete case to the multivariate discrete case is straightforward. We will give the appropriate definitions and an important example, namely that of the *multinomial distribution*, an extension of the binomial distribution.

Definition 11.9. Let $X_1, X_2, \ldots, X_n$ be discrete random variables defined on a common sample space Ω, with X_i taking values in some countable set $\mathcal{X}_i$. The *joint pmf* of $(X_1, X_2, \ldots, X_n)$ is defined as $p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n), x_i \in \mathcal{X}_i$, and zero otherwise..

Definition 11.10. Let $X_1, X_2, \ldots, X_n$ be random variables defined on a common sample space Ω. The *joint CDF* of $X_1, X_2, \ldots, X_n$ is defined as $F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n), x_1, x_2, \ldots, x_n \in \mathcal{R}$.

The requirements of a joint pmf are the usual:

(i) $p(x_1, x_2, \ldots, x_n) \geq 0 \ \forall\, x_1, x_2, \ldots, x_n \in \mathcal{R}$;
(ii) $\sum_{x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n} p(x_1, x_2, \ldots, x_n) = 1.$

The requirements of a joint CDF are somewhat more complicated.

The requirements of a CDF are that

(i) $0 \leq F \leq 1\, \forall (x_1, \ldots, x_n)$;
(ii) F is nondecreasing in each coordinate;
(iii) F equals zero if one or more of the $x_i = -\infty$;
(iv) F equals one if all the $x_i = +\infty$;
(v) F assigns a nonnegative probability to every n-dimensional rectangle

$$[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n].$$

This last condition (v) is a notationally clumsy condition to write down. If $n = 2$, it reduces to the simple inequality that

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \geq 0\, \forall a_1 \leq b_1, a_2 \leq b_2.$$

Once again, we mention that it is not convenient or interesting to work with the CDF for discrete random variables; for discrete variables, it is preferable to work with the pmf.

11.5.1 * Joint MGF

Analogous to the case of one random variable, we can define the joint mgf for several random variables. The definition is the same for all types of random variables, discrete or continuous, or other mixed types. As in the one-dimensional case, the joint mgf of several random variables is also a very useful tool. First, we repeat the definition of expectation of a function of several random variables; see Chapter 4, where it was first introduced and defined. The definition below is equivalent to what was given in Chapter 4.

Definition 11.11. Let $X_1, X_2, \ldots, X_n$ be discrete random variables defined on a common sample space Ω, with X_i taking values in some countable set $\mathcal{X}_i$. Let the joint pmf of $X_1, X_2, \ldots, X_n$ be $p(x_1, \ldots, x_n)$. Let $g(x_1, \ldots, x_n)$ be a real-valued function of n variables. We say that $E[g(X_1, X_2, \ldots, X_n)]$ exists if $\sum_{x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n} |g(x_1, \ldots, x_n)| p(x_1, \ldots, x_n) < \infty$, in which case, the expectation is defined as

$$E[g(X_1, X_2, \ldots, X_n)] = \sum_{x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n} g(x_1, \ldots, x_n) p(x_1, \ldots, x_n).$$

A corresponding definition when $X_1, X_2, \ldots, X_n$ are all continuous random variables will be given in the next chapter.

Definition 11.12. Let $X_1, X_2, \ldots, X_n$ be n random variables defined on a common sample space Ω. The *joint moment generating function* of $X_1, X_2, \ldots, X_n$ is defined to be

$$\psi(t_1, t_2, \ldots, t_n) = E[e^{t_1 X_1 + t_2 X_2 + \cdots + t_n X_n}] = E[e^{\mathbf{t}'\mathbf{X}}],$$

provided the expectation exists, and $\mathbf{t}'\mathbf{X}$ denotes the inner product of the vectors $\mathbf{t} = (t_1, \ldots, t_n), \mathbf{X} = (X_1, \ldots, X_n)$.

Note that the joint moment generating function (mgf) always exists at the origin, namely $\mathbf{t} = (0, \ldots, 0)$, and equals 1 at that point. It may or may not exist at other points $\mathbf{t}$. If it does exist in a nonempty rectangle containing the origin, then many important characteristics of the joint distribution of $X_1, X_2, \ldots, X_n$ can be derived by using the joint mgf. As in the one-dimensional case, it is a very useful tool. Theorem 11.7 gives the moment generating property of a joint mgf.

Theorem 11.7. *Suppose $\psi(t_1, t_2, \ldots, t_n)$ exists in a nonempty open rectangle containing the origin $\mathbf{t} = \mathbf{0}$. Then a partial derivative of $\psi(t_1, t_2, \ldots, t_n)$ of every order with respect to each t_i exists in that open rectangle, and furthermore,*

$$E(X_1^{k_1} X_2^{k_2} \cdots X_n^{k_n}) = \frac{\partial^{k_1 + k_2 + \cdots + k_n}}{\partial t_1^{k_1} \cdots \partial t_n^{k_n}} \psi(t_1, t_2, \ldots, t_n) | t_1 = 0, t_2 = 0, \ldots, t_n = 0.$$

A corollary of this result is sometimes useful in determining the covariance between two random variables.

Corollary. Let X and Y have a joint mgf in some open rectangle around the origin $(0, 0)$. Then,

$$\text{Cov}(X, Y) = \frac{\partial^2}{\partial t_1 \partial t_2}\psi(t_1, t_2)|_{0,0} - \left(\frac{\partial}{\partial t_1}\psi(t_1, t_2)|_{0,0}\right)\left(\frac{\partial}{\partial t_2}\psi(t_1, t_2)|_{0,0}\right).$$

We also have the distribution-determining property, as in the one-dimensional case.

Theorem 11.8. *Suppose $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_n)$ are two sets of jointly distributed random variables, such that their mgfs $\psi_{\mathbf{X}}(t_1, t_2, \ldots, t_n)$ and $\psi_{\mathbf{Y}}(t_1, t_2, \ldots, t_n)$ exist and coincide in some nonempty open rectangle containing the origin. Then $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_n)$ have the same joint distribution.*

Remark. It is important to note that the last two theorems are not limited to discrete random variables; they are valid for general random variables. The proofs of these two theorems follow the same arguments as in the one-dimensional case, namely that when an mgf exists in a nonempty open rectangle, it can be differentiated infinitely often with respect to each variable t_i inside the expectation; i.e., the order of the derivative and the expectation can be interchanged. See Chapter 5 for this argument.

11.5.2 Multinomial Distribution

One of the most important multivariate discrete distributions is the multinomial distribution. The multinomial distribution corresponds to n balls being distributed to k cells independently, with each ball having the probability p_i of being dropped into the ith cell. The random variables under consideration are $X_1, X_2, \ldots, X_k$, where X_i is the number of balls that get dropped into the ith cell. Then their joint pmf is the *multinomial pmf* defined below.

Definition 11.13. A multivariate random vector $(X_1, X_2, \ldots, X_k)$ is said to have a multinomial distribution with parameters $n, p_1, p_2, \ldots, p_k$ if it has the pmf

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, x_i \geq 0,$$

$$\sum_{i=1}^{k} x_i = n, p_i \geq 0, \sum_{i=1}^{k} p_i = 1.$$

We write $(X_1, X_2, \ldots, X_k) \sim \textit{Mult}(n, p_1, \ldots, p_k)$ to denote a random vector with a multinomial distribution.

Example 11.20 (Dice Rolls). Suppose a fair die is rolled 30 times. We want to find the probabilities that

(i) each face is obtained exactly five times and
(ii) the number of sixes is at least five.

If we denote as X_i the number of times face number i is obtained, then $(X_1, X_2, \ldots, X_6) \sim Mult(n, p_1, \ldots, p_6)$, where $n = 30$ and each $p_i = \frac{1}{6}$. Therefore,

$$\begin{aligned} P(X_1 = 5, X_2 = 5, \cdots, X_6 = 5) &= \frac{30!}{(5!)^6}\left(\frac{1}{6}\right)^5 \cdots \left(\frac{1}{6}\right)^5 \\ &= \frac{30!}{(5!)^6}\left(\frac{1}{6}\right)^{30} = .0004. \end{aligned}$$

Next, each of the thirty rolls will either be a six or not, independently of the other rolls, with probability $\frac{1}{6}$, and so $X_6 \sim Bin(30, \frac{1}{6})$. Therefore,

$$P(X_6 \geq 5) = 1 - P(X_6 \leq 4) = 1 - \sum_{x=0}^{4} \binom{30}{x}\left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{30-x} = .5757.$$

Example 11.21 (Bridge). Consider a bridge game with four players, North, South, East, and West. We want to find the probability that North and South together have two or more aces. Let X_i denote the number of aces in the hands of player i, $i = 1, 2, 3, 4$; we let i = 1, 2 mean North and South. Then, we want to find $P(X_1 + X_2 \geq 2)$.

The joint distribution of (X_1, X_2, X_3, X_4) is $Mult(4, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ (think of each ace as a ball and the four players as cells). Then, $(X_1 + X_2, X_3 + X_4) \sim Mult(4, \frac{1}{2}, \frac{1}{2})$. Therefore,

$$\begin{aligned} P(X_1 + X_2 \geq 2) &= \frac{4!}{2!2!}\left(\frac{1}{2}\right)^4 + \frac{4!}{3!1!}\left(\frac{1}{2}\right)^4 + \frac{4!}{4!0!}\left(\frac{1}{2}\right)^4 \\ &= \frac{11}{16}. \end{aligned}$$

Important formulas and facts about the multinomial distribution are given in the next theorem.

Theorem 11.9. *Let* $(X_1, X_2, \ldots, X_k) \sim Mult(n, p_1, p_2, \ldots, p_k)$. *Then,*

(a) $E(X_i) = np_i$; $\mathrm{Var}(X_i) = np_i(1 - p_i)$;
(b) $\forall\, i, X_i \sim Bin(n, p_i)$;
(c) $\mathrm{Cov}(X_i, X_j) = -np_i p_j$, $\forall i \neq j$;
(d) $\rho_{X_i, X_j} = -\sqrt{\frac{p_i p_j}{(1 - p_i)(1 - p_j)}}$, $\forall i \neq j$;

(e) $\forall m, 1 \le m < k, (X_1, X_2, \ldots, X_m)|(X_{m+1} + X_{m+2} + \cdots + X_k) = s \sim$ *Mult* $(n - s, \theta_1, \theta_2, \ldots, \theta_m)$, *where* $\theta_i = \frac{p_i}{p_1 + p_2 + \cdots + p_m}$.

Proof. Define W_{ir} as the indicator of the event that the rth ball lands in the ith cell. Note that, for a given i, the variables W_{ir} are independent. Then,

$$X_i = \sum_{r=1}^{n} W_{ir},$$

and therefore $E(X_i) = \sum_{r=1}^{n} E[W_{ir}] = np_i$ and $\text{Var}(X_i) = \sum_{r=1}^{n} \text{Var}(W_{ir}) = np_i(1 - p_i)$. Part (b) follows from the definition of a multinomial experiment (the trials are identical and independent, and each ball either lands or does not and in the ith cell). For part (c),

$$\begin{aligned}
\text{Cov}(X_i, X_j) &= \text{Cov}\left(\sum_{r=1}^{n} W_{ir}, \sum_{s=1}^{n} W_{js}\right) \\
&= \sum_{r=1}^{n}\sum_{s=1}^{n} \text{Cov}(W_{ir}, W_{js}) \\
&= \sum_{r=1}^{n} \text{Cov}(W_{ir}, W_{jr})
\end{aligned}$$

(because $\text{Cov}(W_{ir}, W_{js})$ would be zero when $s \neq r$)

$$\begin{aligned}
&= \sum_{r=1}^{n} [E(W_{ir}W_{jr}) - E(W_{ir})E(W_{jr})] \\
&= \sum_{r=1}^{n} [0 - p_i p_j] = -np_i p_j.
\end{aligned}$$

Part (d) follows immediately from part (c) and part (a). Part (e) is a calculation and is omitted.

Example 11.22 (MGF of the Multinomial Distribution). Let $(X_1, X_2, \ldots, X_k) \sim$ *Mult*$(n, p_1.p_2, \ldots, p_k)$. Then the mgf $\psi(t_1, t_2, \ldots, t_k)$ exists at all $\mathbf{t}$, and a formula follows easily. Indeed,

$$\begin{aligned}
E[e^{t_1X_1 + \cdots + t_kX_k}] &= \sum_{x_i \ge 0, \sum_{i=1}^{k} x_i = n} \frac{n!}{x_1! \cdots x_k!} e^{t_1x_1} e^{t_2x_2} \cdots e^{t_kx_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \\
&= \sum_{x_i \ge 0, \sum_{i=1}^{k} x_i = n} \frac{n!}{x_1! \cdots x_k!} (p_1e^{t_1})^{x_1} (p_2e^{t_2})^{x_2} \cdots (p_ke^{t_k})^{x_k} \\
&= (p_1e^{t_1} + p_2e^{t_2} + \cdots + p_ke^{t_k})^n
\end{aligned}$$

by the *multinomial expansion identity*

$$(a_1 + a_2 + \cdots + a_k)^n = \sum_{x_i \geq 0, \sum_{i=1}^k x_i = n} \frac{n!}{x_1! \cdots x_k!} a_1^{x_1} a_2^{x_2} \cdots a_k^{x_k}.$$

11.6 Synopsis

(a) The joint pmf of two discrete random variables X and Y must satisfy

$$p(x, y) = P(X = x, Y = y) \geq 0 \ \forall (x, y); \sum_i \sum_j p(x_i, y_j) = 1.$$

The joint CDF is defined as $F(x, y) = P(X \leq x, Y \leq y), x, y \in \mathcal{R}$.

(b) The marginal pmfs of X and Y can be found from the joint pmf as

$$p_X(x) = \sum_y p(x, y); \ \ p_Y(y) = \sum_x p(x, y).$$

More generally, for any set A, $P((X, Y) \in A) = \sum_{(x,y) \in A} p(x, y)$.

(c) The expectation of a function $g(X, Y)$ is given by $E[g(X, Y)] = \sum_x \sum_y g(x, y) p(x, y)$.
In particular, the marginal expectations $E(X)$ and $E(Y)$ can be found in any of the following ways:

$$E(X) = \sum_x x p_X(x); E(X) = \sum_x \sum_y x p(x, y),$$

and similarly for $E(Y)$.

(d) The conditional distribution of X given $Y = y$ is defined as $p(x|y) = P(X = x|Y = y) = \frac{p(x,y)}{p_Y(y)}$. The conditional expectation of X given $Y = y$ is defined as $E(X|Y = y) = \sum_x x p(x|y) = \frac{\sum_x x p(x,y)}{p_Y(y)}$. The conditional distribution of Y given $X = x$ and the conditional expectation of Y given $X = x$ are defined similarly.

(e) The conditional variance of X given $Y = y$ is defined as

$$\text{Var}(X|Y = y) = E[(X - \mu_X(y))^2 | Y = y] = \sum_x (x - \mu_X(y))^2 p(x|y),$$

where $\mu_X(y) = E(X|Y = y)$. In other words, the conditional variance of X given $Y = y$ is just the variance of the conditional distribution of X given $Y = y$.

(f) Conditional expectations and conditional variances satisfy numerous rules, which are given in the text. Two special rules are the following:
iterated expectation formula: $E(X) = E_Y[E(X|Y = y)]$;
iterated variance formula: $\text{Var}(X) = E_Y[\text{Var}(X|Y = y)] + \text{Var}_Y[E(X|Y = y)]$.

(g) Two discrete random variables X and Y are independent if and only if $p(x|y) = p_X(x)$ for all x, y, or equivalently $p(y|x) = p_Y(y)$ for all x, y. Both of these are equivalent to $p(x, y) = p_X(x)p_Y(y)$ for all x, y.

(h) The definitions of the joint pmf, the joint CDF, and independence all extend to the case of more than two variables in the obvious way. For example, the joint pmf of $X_1, \ldots, X_n$ is defined as $p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$.

(i) The covariance of X and Y is defined as $\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = E[(X - E(X))(Y - E(Y))]$. Covariance enters naturally into expressing the variance of sums and linear combinations of random variables. For example,

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y).$$

In particular,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

and

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y).$$

More generally,

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{i<j=1}^{n} a_i a_j \text{Cov}(X_i, X_j).$$

(j) Covariance is additive. That is,

$$\text{Cov}(X + Y, Z + W) = \text{Cov}(X, Z) + \text{Cov}(X, W) + \text{Cov}(Y, Z) + \text{Cov}(Y, W).$$

Other properties of the covariance are given in the text.

(k) The correlation between X and Y is defined as $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$. Correlation is always a number between -1 and 1, and its magnitude remains the same under linear transformations. Precisely, $\rho_{a+bX,c+dY} = \text{sgn}(bd)\rho_{X,Y}$ (see the text).

(l) If X and Y are independent random variables, then both $\text{Cov}(X, Y)$ and $\rho_{X,Y}$ are zero; but the converse in general is not true.

(m) A special multivariate discrete distribution is the multinomial distribution, which corresponds to a distribution of n balls into k cells, the balls being

distributed independently and with probabilities $p_1, p_2, \ldots, p_k$ of being distributed into the k cells. The multinomial pmf is given by

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

$x_i \geq 0, \sum_{i=1}^{k} x_i = n, p_i \geq 0, \sum_{i=1}^{k} p_i = 1.$

(n) If $(X_1, X_2, \ldots, X_k)$ has a joint multinomial distribution, then

$$E(X_i) = np_i; \text{Var}(X_i) = np_i(1 - p_i); \text{Cov}(X_i, X_j) = -np_i p_j, \ \forall i \neq j;$$
$$\rho_{X_i, X_j} = -\sqrt{\frac{p_i p_j}{(1 - p_i)(1 - p_j)}}, \ \forall i \neq j.$$

Many other properties are given in the text.

(o) An important characteristic of a joint distribution is the joint mgf. The joint mgf of $X_1, X_2, \ldots, X_n$ is defined as

$$\psi(t_1, t_2, \cdots, t_n) = E[e^{t_1 X_1 + t_2 X_2 + \cdots + t_n X_n}] = E[e^{\mathbf{t}' \mathbf{X}}].$$

The joint mgf determines a joint distribution in the same way that the mgf determines the distribution in the case of one variable. If $X_1, X_2, \ldots, X_n$ are independent, then the joint mgf factorizes as

$$\psi(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} E[e^{t_i X_i}] = \prod_{i=1}^{n} \psi_i(t_i).$$

11.7 Exercises

Exercise 11.1. Consider the experiment of picking one word at random from the sentence

ALL IS WELL IN THE NEWELL FAMILY

Let X be the length of the word selected and Y the number of L's in it. Find in tabular form the joint pmf of X and Y, their marginal pmfs, means, and variances, and the correlation between X and Y.

Exercise 11.2. A fair coin is tossed four times. Let X be the number of heads, Z the number of tails, and $Y = |X - Z|$. Find the joint pmf of (X, Y), and $E(Y)$.

Exercise 11.3. Consider the joint pmf $p(x, y) = cxy, 1 \leq x \leq 3, 1 \leq y \leq 3$.

(a) Find the normalizing constant c.
(b) Are X and Y independent? Prove your claim.
(c) Find the expectations of X, Y, XY.

Exercise 11.4. Consider the joint pmf $p(x, y) = cxy, 1 \le x \le y \le 3$.

(a) Find the normalizing constant c.
(b) Are X and Y independent? Prove your claim.
(c) Find the expectations of X, Y, XY.

Exercise 11.5. A fair die is rolled twice. Let X be the maximum and Y the minimum of the two rolls. By using the joint pmf of (X, Y) worked out in the text, find the pmf of $\frac{X}{Y}$ and hence the mean of $\frac{X}{Y}$.

Exercise 11.6. A hat contains four slips of paper, numbered 1, 2, 3, and 4. Two slips are drawn at random without replacement. X is the number on the first slip and Y the sum of the two numbers drawn. Write in tabular form the joint pmf of (X, Y). Hence find the marginal pmfs. Are X and Y independent?

Exercise 11.7. * **(Conditional Expectation in Bridge).** Let X be the number of clubs in the hand of North and Y the number of clubs in the hand of South in a bridge game. Write a general formula for $E(X|Y = y)$, and compute $E(X|Y = 3)$. Can you compute $E(Y|X = 3)$?

Exercise 11.8. A fair die is rolled four times. Find the probabilities that

(a) at least one six is obtained;
(b) exactly one six and exactly one two is obtained;
(c) exactly one six, one two, and two fours are obtained.

Exercise 11.9 (Iterated Expectation). A household has a Poisson number of cars with mean 1. Each car that a household possesses has, independently of the other cars, a 20% chance of being an SUV. Find the mean number of SUVs a household possesses.

Exercise 11.10 (Iterated Variance). Suppose $N \sim Poi(\lambda)$, and given $N = n$, X is distributed as uniform on $\{0, 1, \ldots, n\}$. Find the variance of the marginal distribution of X.

Exercise 11.11. Suppose X and Y are independent $Geo(p)$ random variables. Find $P(X \ge Y)$ and $P(X > Y)$.

Exercise 11.12. * Suppose X and Y are independent $Poi(\lambda)$ random variables. Find $P(X \ge Y)$ and $P(X > Y)$.

Hint: This will involve a Bessel function of a suitable kind.

Exercise 11.13. Suppose X and Y are independent, and take the values 1, 2, 3, 4 with probabilities .2, .3, .3, .2, respectively Find the pmf of $X + Y$.

Exercise 11.14. Two random variables have the joint pmf $p(x, x+1) = \frac{1}{n+1}, x = 0, 1, \ldots, n$. Answer the following questions with as little calculation as possible.

(a) Are X and Y independent?
(b) What is the variance of $Y - X$?
(c) What is $\mathrm{Var}(Y|X = 1)$?

Exercise 11.15 (Binomial Conditional Distribution). Suppose X and Y are independent random variables and that $X \sim Bin(m, p), Y \sim Bin(n, p)$. Show that the conditional distribution of X given $X + Y = t$ is a hypergeometric distribution. Identify the parameters of this hypergeometric distribution.

Exercise 11.16. * **(Poly-hypergeometric Distribution).** A box has D_1 red, D_2 green, and D_3 blue balls. Suppose n balls are picked from the box at random without replacement. Let X, Y, Z be the number of red, green, and blue balls selected, respectively. Find the joint pmf of (X, Y, Z).

Exercise 11.17 (Bivariate Poisson). Suppose U, V, W are independent Poisson random variables with means λ, μ, η, respectively. Let $X = U + W, Y = V + W$.

(a) Find the marginal pmfs of X and Y.
(b) Find the joint pmf of (X, Y).

Exercise 11.18. Suppose a fair die is rolled twice. Let X and Y be the two rolls. Find the following with as little calculation as possible:

(a) $E(X + Y|Y = y)$.
(b) $E(XY|Y = y)$.
(c) $\text{Var}(X^2Y|Y = y)$.
(d) $\rho_{X+Y,X-Y}$.

Exercise 11.19. * **(A Waiting Time Problem).** In repeated throws of a fair die, let X be the throw in which the first six is obtained and Y the throw in which the second six is obtained.

(a) Find the joint pmf of (X, Y).
(b) Find the expectation of $Y - X$.
(c) Find $E(Y - X|X = 8)$.
(d) Find $\text{Var}(Y - X|X = 8)$.

Exercise 11.20. * **(Family Planning).** A couple wants to have a child of each sex, but they will have at most four children. Let X be the total number of children they will have and Y the number of girls at the second childbirth. Find the joint pmf of (X, Y) and the conditional expectation of X given $Y = y, y = 0, 2$.

Exercise 11.21 (A Standard Deviation Inequality). Let X and Y be two random variables. Show that $\sigma_{X+Y} \leq \sigma_X + \sigma_Y$.

Exercise 11.22. * **(A Covariance Fact).** Let X and Y be two random variables. Suppose $E(X|Y = y)$ is nondecreasing in y. Show that $\rho_{X,Y} \geq 0$, assuming the correlation exists.

Exercise 11.23 (Another Covariance Fact). Let X and Y be two random variables. Suppose $E(X|Y = y)$ is a finite constant c. Show that $\text{Cov}(X, Y) = 0$.

Exercise 11.24 (Two-Valued Random Variables). Suppose X and Y are both two-valued random variables. Prove that X and Y are independent if and only if they have a zero correlation.

Exercise 11.25 (A Correlation Inequality). Suppose X and Y each have mean 0 and variance 1 and a correlation ρ. Show that $E(\max\{X^2, Y^2\}) \leq 1 + \sqrt{1-\rho^2}$.

Exercise 11.26. * **(A Covariance Inequality).** Let X be any random variable and $g(X)$ and $h(X)$ two functions such that they are both nondecreasing or both nonincreasing. Show that $\text{Cov}(g(X), h(X)) \geq 0$.

Exercise 11.27 (Joint MGF). Suppose a fair die is rolled four times. Let X be the number of ones and Y the number of sixes. Find the joint mgf of X and Y and hence the covariance between X and Y.

Exercise 11.28 (MGF of Bivariate Poisson). Suppose U, V, W are independent Poisson random variables with means λ, μ, η, respectively. Let $X = U + W$, $Y = V + W$. Find the joint mgf of X, Y and hence $E(XY)$.

Exercise 11.29 (Joint MGF). In repeated throws of a fair die, let X be the throw in which the first six is obtained and Y the throw in which the second six is obtained. Find the joint mgf of X, Y and hence the covariance between X and Y.

Chapter 12
Multidimensional Densities

Similar to the case of several discrete random variables, in applications we are frequently interested in studying several continuous random variables simultaneously. An example would be a physician's measurement of a patient's height, weight, blood pressure, electrolytes, and blood sugar. Analogous to the case of one continuous random variable, again we do not talk of pmfs of several continuous variables but of a pdf jointly for all the continuous random variables. The joint density function completely characterizes the joint distribution of the full set of continuous random variables. We refer to the entire set of random variables as a random vector. Both the calculation aspects and the application aspects of multidimensional density functions are generally sophisticated. As such, using and operating with multidimensional densities are among the most important skills one needs to have in probability and statistics. The general concepts and calculations are discussed in this chapter. Some special multidimensional densities, and in particular the multivariate normal density, are introduced separately in the next chapter.

12.1 Joint Density Function and Its Role

Exactly as in the one-dimensional case, it is important to note the following points

(a) The joint density function of all the variables does not equal the probability of a specific point in the multidimensional space; *the probability of any specific point is still zero*.
(b) The joint density function reflects the relative importance of a particular point. Thus, the probability that the variables together belong to a small set around a specific point, say $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, is roughly equal to the volume of that set multiplied by the density function at the specific point $\mathbf{x}$. *This volume interpretation for probabilities is useful for intuitive understanding of the distributions of multidimensional continuous random variables*.
(c) For a general set A in the multidimensional space, the probability that the random vector $\mathbf{X}$ belongs to A is obtained by integrating the joint density function over the set A.

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_12,

These are all just the most natural extensions of the corresponding one-dimensional facts to the present multidimensional case. We now formally define a joint density function.

Definition 12.1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be an n-dimensional random vector taking values in $\mathcal{R}^n$ for some $n, 1 < n < \infty$. We say that $f(x_1, x_2, \ldots, x_n)$ is the *joint density* or simply the density of $\mathbf{X}$ if, for all $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$, $-\infty < a_i \le b_i < \infty$,

$$P(a_1 \le X_1 \le b_1, a_2 \le X_2 \le b_2, \ldots, a_n \le X_n \le b_n) = \int_{a_n}^{b_n} \ldots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n.$$

In order that a function $f : \mathcal{R}^n \to \mathcal{R}$ be a density function of some n-dimensional random vector, it is necessary and sufficient that

(i) $f(x_1, x_2, \ldots, x_n) \ge 0 \ \forall\, (x_1, x_2, \ldots, x_n) \in \mathcal{R}^n$;
(ii) $\int_{\mathcal{R}^n} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n = 1$.

The definition of the joint CDF is the same as that given in the discrete case. But now the joint CDF is an integral of the density rather than a sum. Here is the precise definition.

Definition 12.2. Let $\mathbf{X}$ be an n-dimensional random vector with the density function $f(x_1, x_2, \ldots, x_n)$. The *joint CDF*, or simply the CDF, of $\mathbf{X}$ is defined as

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(t_1, \cdots, t_n) dt_1 \cdots dt_n.$$

As in the one-dimensional case, *both the CDF and the density completely specify the distribution of a continuous random vector* and one can be obtained from the other. We know how to obtain the CDF from the density; the reverse relation is that (for *almost all* $(x_1, x_2, \ldots, x_n)$)

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_1, x_2, \ldots, x_n).$$

Again, the qualification *almost all* is necessary for a rigorous description of the interrelation between the CDF and the density, but we will operate as though the identity above holds for *all* $(x_1, x_2, \ldots, x_n)$.

Analogous to the case of several discrete variables, the marginal densities are obtained by integrating out (instead of summing) all the other variables. In fact, all lower-dimensional marginals are obtained that way. The precise statement is the following.

Proposition. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a continuous random vector with a joint density $f(x_1, x_2, \ldots, x_n)$. Let $1 \le p < n$. Then the *marginal joint density* of $(X_1, X_2, \ldots, X_p)$ is given by

$$f_{1,2,\ldots,p}(x_1, x_2, \ldots, x_p) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_{p+1} \ldots dx_n.$$

At this stage, it is useful to give a characterization of independence of a set of n continuous random variables by using the density function.

Proposition. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a continuous random vector with a joint density $f(x_1, x_2, \ldots, x_n)$. Then, $X_1, X_2, \ldots, X_n$ are independent if and only if the joint density factorizes as

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i),$$

where $f_i(x_i)$ is the marginal density function of X_i.

Proof. If the joint density factorizes as above, then on integrating both sides of this factorization identity, one gets $F(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i)\ \forall\ (x_1, x_2, \ldots, x_n)$, which is the definition of independence.

Conversely, if they are independent, then take the identity

$$F(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i)$$

and partially differentiate both sides successively with respect to $x_1, x_2, \ldots, x_n$, and it follows that the joint density factorizes as $f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$.

Let us see some initial examples.

Example 12.1 (Bivariate Uniform). Consider the function

$$\begin{aligned} f(x, y) &= 1 \text{ if } 0 \le x \le 1, 0 \le y \le 1, \\ &= 0 \text{ otherwise.} \end{aligned}$$

Clearly, f is always nonnegative, and

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= \int_0^1 \int_0^1 f(x, y) dx dy \\ &= \int_0^1 \int_0^1 dx dy = 1. \end{aligned}$$

Therefore, f is a valid bivariate density function. The marginal density of X is

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\ &= \int_0^1 f(x, y) dy = \int_0^1 dy = 1 \end{aligned}$$

if $0 \le x \le 1$ and zero otherwise. Thus, marginally, $X \sim U[0,1]$, and similarly, marginally, $Y \sim U[0,1]$. Furthermore, clearly, for *all* x, y the joint density $f(x,y)$ factorizes as $f(x,y) = f_1(x) f_2(y)$, so X, and Y are independent, too. The joint density $f(x,y)$ of this example is called *the bivariate uniform density*. It gives the constant density of 1 to all points (x,y) in the unit square $[0,1] \times [0,1]$ and zero density outside of the unit square. The bivariate uniform therefore is the same as putting two independent $U[0,1]$ variables together as a bivariate vector.

Example 12.2 (Uniform in a Triangle). Consider the function

$$\begin{aligned} f(x,y) &= c \text{ if } x, y \ge 0, x + y \le 1, \\ &= 0 \text{ otherwise.} \end{aligned}$$

The set of points $x, y \ge 0, x + y \le 1$ form a triangle in the plane with vertices at $(0,0)$, $(1,0)$, and $(0,1)$; thus, it is just half the unit square (see Figure 12.1). The normalizing constant c is easily evaluated:

$$\begin{aligned} 1 &= \int_{x,y:x,y\ge 0,x+y\le 1} c dx dy \\ &= \int_0^1 \int_0^{1-y} c dx dy \\ &= c \int_0^1 (1-y) dy \\ &= \frac{c}{2} \\ &\Rightarrow c = 2. \end{aligned}$$

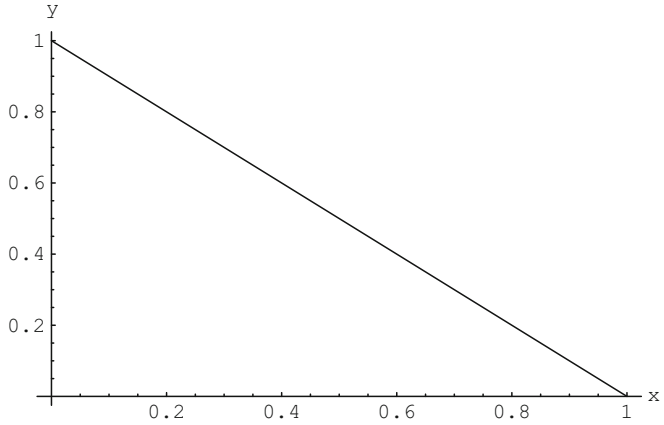


Fig. 12.1 Uniform density on a triangle equals $c = 2$ for this set

The marginal density of X is

$$f_1(x) = \int_0^{1-x} 2dy = 2(1-x), 0 \le x \le 1.$$

Similarly, the marginal density of Y is

$$f_2(y) = 2(1-y), 0 \le y \le 1.$$

Contrary to the previous example, X and Y are *not independent* now. There are many ways to see this. For example,

$$P\left(X > \frac{1}{2} | Y > \frac{1}{2}\right) = 0.$$

But, $P(X > \frac{1}{2}) = \int_{\frac{1}{2}}^1 2(1-x)dx = \frac{1}{4} \neq 0$, so X and Y cannot be independent. We can also see that the joint density $f(x, y)$ does not factorize as the product of the marginal densities, so X and Y cannot be independent.

Example 12.3. Consider the function $f(x, y) = xe^{-x(1+y)}, x, y \ge 0$. First, let us verify that it is a valid density function.

It is obviously nonnegative. Furthermore,

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy &= \int_0^{\infty}\int_0^{\infty} xe^{-x(1+y)}dxdy \\ &= \int_0^{\infty} \frac{1}{(1+y)^2}dy \\ &= \int_1^{\infty} \frac{1}{y^2}dy = 1.\end{aligned}$$

Hence, $f(x, y)$ is a valid joint density. Next, let us find the marginal densities

$$\begin{aligned}f_1(x) &= \int_0^{\infty} xe^{-x(1+y)}dy = x\int_0^{\infty} e^{-x(1+y)}dy \\ &= x\int_1^{\infty} e^{-xy}dy = x\frac{e^{-x}}{x} = e^{-x}, x \ge 0.\end{aligned}$$

Therefore, marginally, X is a standard exponential. Next,

$$f_2(y) = \int_0^{\infty} xe^{-x(1+y)}dx = \frac{1}{(1+y)^2}, y \ge 0.$$

Clearly, we do not have the factorization identity $f(x, y) = f_1(x)f_2(y) \; \forall \, x, y$; thus, X and Y are not independent. The joint density is plotted in Figure 12.2.

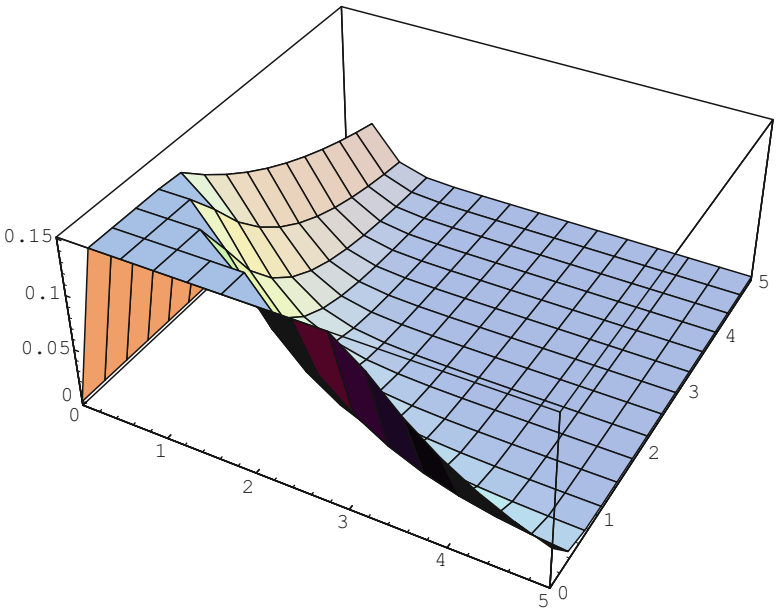


Fig. 12.2 The density $f(x, y) = xe^{-x(1+y)}$

Example 12.4 (Nonuniform Joint Density with Uniform Marginals). Let (X, Y) have the joint density function $f(x, y) = c - 2(c - 1)(x + y - 2xy), x, y \in [0, 1], 0 < c < 2$. This is nonnegative in the unit square, as can be seen by considering the cases $c < 1, c = 1, c > 1$ separately. Also,

$$\begin{aligned}
&\int_0^1 \int_0^1 f(x, y)dxdy \\
&= c - 2(c - 1) \int_0^1 \int_0^1 (x + y - 2xy)dxdy \\
&= c - 2(c - 1) \int_0^1 \left(\frac{1}{2} + y - y\right) dy = c - (c - 1) = 1.
\end{aligned}$$

Now, the marginal density of X is

$$\begin{aligned}
f_1(x) &= \int_0^1 f(x, y)dy \\
&= c - 2(c - 1)\left[x + \frac{1}{2} - x\right] = 1.
\end{aligned}$$

Similarly, the marginal density of Y is also the constant function 1, so each marginal is uniform, although the joint density is not uniform if $c \neq 1$.

Example 12.5 (Using the Density to Calculate a Probability). Suppose (X, Y) has the joint density $f(x, y) = 6xy^2, x, y \geq 0, x + y \leq 1$. Thus, this is yet another density on the triangle with vertices at $(0, 0), (1, 0)$, and $(0, 1)$. We want to find $P(X + Y < \frac{1}{2})$. By definition,

$$\begin{aligned} P\left(X + Y < \frac{1}{2}\right) &= \int_{(x,y);x,y\geq 0, x+y<\frac{1}{2}} 6xy^2 dxdy \\ &= 6\int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}-y} xy^2 dxdy \\ &= 6\int_0^{\frac{1}{2}} y^2 \frac{\left(\frac{1}{2}-y\right)^2}{2} dy \\ &= 3\int_0^{\frac{1}{2}} y^2 \left(\frac{1}{2}-y\right)^2 dy \\ &= 3 \times \frac{1}{960} = \frac{1}{320}. \end{aligned}$$

This example gives an elementary illustration of the need to work out the limits of the iterated integrals carefully while using a joint density to calculate the probability of some event. In fact, properly finding the limits of the iterated integrals is the part that requires the greatest care when working with joint densities.

Example 12.6 (Uniform Distribution in a Circle). Suppose C denotes the unit circle in the plane:

$$C = \{(x, y) : x^2 + y^2 \leq 1\}.$$

We pick a point (X, Y) at random from C. What that means is that (X, Y) has the density

$$f(x, y) = c \text{ if } (x, y) \in C$$

and is zero otherwise. Since

$$\int_C f(x, y)dxdy = c\int_C dxdy = c \times \text{ area of } C = c\pi = 1,$$

we have that the normalizing constant $c = \frac{1}{\pi}$. Let us find the marginal densities. First,

$$\begin{aligned} f_1(x) &= \int_{y:x^2+y^2\leq 1} \frac{1}{\pi} dy = \frac{1}{\pi} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy \\ &= \frac{2\sqrt{1-x^2}}{\pi}, -1 \leq x \leq 1. \end{aligned}$$

Since the joint density $f(x, y)$ is symmetric between x and y (i.e., $f(x, y) = f(y, x)$), Y has the same marginal density as X,

$$f_2(y) = \frac{2\sqrt{1-y^2}}{\pi}, -1 \le y \le 1.$$

Since $f(x, y) \neq f_1(x) f_2(y)$, X and Y are not independent. Note that if X and Y have a joint uniform density in the unit square, we have found them to be independent, but now, when they have a uniform density in the unit circle, we find them not to be independent. In fact, the following general rule holds:

Suppose a joint density $f(x, y)$ can be written in a form $g(x)h(y), (x, y) \in S$, and $f(x, y)$ zero otherwise. Then, X and Y are independent if and only if S is a rectangle (including squares).

Example 12.7. Cathy and Jen have agreed to meet at a cafe between 10:00 AM and 11:00 AM. Cathy will arrive at a random time during that hour, and Jen's arrival time has a Beta density with each parameter equal to 2. They arrive independently, and the first to arrive waits 15 minutes for the other. We want to find the probability that they will meet.

Take 10:00 AM as time zero. We let X and Y denote the arrival times of Cathy and Jen, respectively, so that $X \sim U[0, 1], Y \sim Be(2, 2)$, and X and Y are independent. We want to find $P(|X - Y| \le \frac{1}{4})$. Again, this problem uses the fact that the probability of an event is found by integrating the joint density over the event. In evaluating the iterated integral, the limits of the integral have to be found carefully. The part of the unit square that corresponds to the event $|X - Y| \le \frac{1}{4}$ is plotted in Figure 12.3.

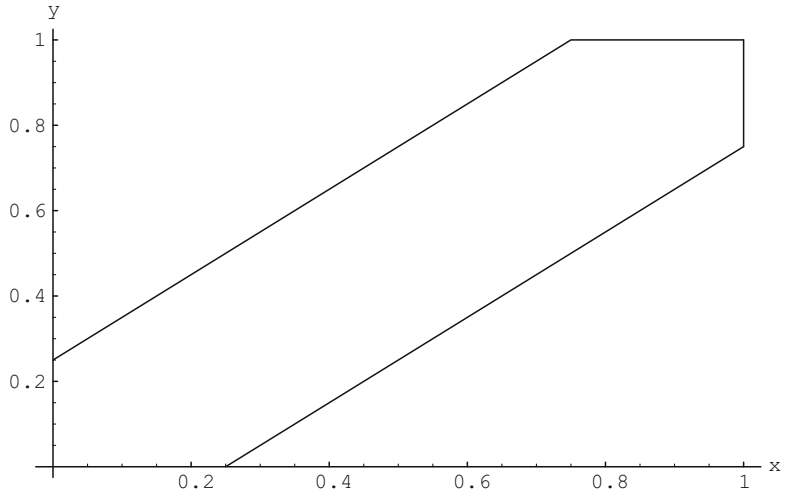


Fig. 12.3 The area $|x - y|$ smaller than 1/4 in the unit square

The necessay probability calculation is

$$
\begin{aligned}
P\left(|X-Y| \le \frac{1}{4}\right) &= P\left(Y - \frac{1}{4} \le X \le Y + \frac{1}{4}\right) \\
&= \int_{(x,y):0\le x,y\le 1, y-\frac{1}{4}\le x\le y+\frac{1}{4}} f(x,y)dxdy \\
&= \int_0^{\frac{1}{4}} \int_0^{y+\frac{1}{4}} f(x,y)dxdy + \int_{\frac{1}{4}}^1 \int_{y-\frac{1}{4}}^{y+\frac{1}{4}} f(x,y)dxdy \\
&= \int_0^{\frac{1}{4}} \int_0^{y+\frac{1}{4}} 6y(1-y)dxdy + \int_{\frac{1}{4}}^1 \int_{y-\frac{1}{4}}^{y+\frac{1}{4}} 6y(1-y)dxdy \\
&= \int_0^{\frac{1}{4}} \left(y+\frac{1}{4}\right) 6y(1-y)dy + \int_{\frac{1}{4}}^1 \frac{1}{2} 6y(1-y)dy \\
&= \frac{33}{512} + \frac{27}{64} = \frac{249}{512} = .486.
\end{aligned}
$$

Example 12.8 (An Interesting Property of Exponential Variables). Suppose X and Y are independent $Exp(\lambda), Exp(\mu)$ variables. We want to find $P(X \le Y)$. A possible application is the following. Suppose you have two televisions at your home, a plasma unit with a mean lifetime of five years and an ordinary unit with a mean lifetime of ten years. What is the probability that the plasma TV will fail before the ordinary one?

From our general definition of probabilities of events, we need to calculate $\int_{x,y>0,x\le y} f(x,y)dxdy$. In general, there need not be an interesting answer for this integral. But here, in the independent exponential case, there is.

Since X and Y are independent, the joint density is $f(x,y) = \frac{1}{\lambda\mu} e^{-x/\lambda - y/\mu}$, $x, y > 0$. Therefore,

$$
\begin{aligned}
P(X \le Y) &= \int_{x,y>0,x\le y} \frac{1}{\lambda\mu} e^{-x/\lambda - y/\mu} dxdy \\
&= \frac{1}{\lambda\mu} \int_0^\infty \int_0^y e^{-x/\lambda - y/\mu} dxdy \\
&= \frac{1}{\mu} \int_0^\infty e^{-y/\mu} \int_0^{y/\lambda} e^{-x} dxdy \\
&= \frac{1}{\mu} \int_0^\infty e^{-y/\mu}(1 - e^{-y/\lambda})dy \\
&= 1 - \frac{1}{\mu} \int_0^\infty e^{-y(1/\mu + 1/\lambda)} dy
\end{aligned}
$$

$$= 1 - \frac{\frac{1}{\mu}}{\frac{1}{\mu} + \frac{1}{\lambda}} = 1 - \frac{\lambda}{\lambda + \mu}$$

$$= \frac{\mu}{\lambda + \mu} = \frac{1}{1 + \frac{\lambda}{\mu}}.$$

Thus, the probability that X is less than Y depends in a very simple way on just the quantity $\frac{E(X)}{E(Y)}$.

Example 12.9 (Curse of Dimensionality). A phenomenon that complicates the work of a probabilist in high dimensions (i.e., when dealing with a large number of random variables simultaneously) is that the major portion of the probability in the joint distribution lies *away from the central region of the variable space*. As a consequence, sample observations taken from the high-dimensional distribution tend to leave the central region sparsely populated. Therefore, it becomes difficult to learn about what the distribution is doing in the central region. This phenomenon has been called *the curse of dimensionality*.

As an example, consider n independent $U[-1, 1]$ random variables, $X_1, X_2, \ldots, X_n$, and suppose we ask what the probability is that $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ lies in the inscribed sphere

$$B_n = \{(x_1, x_2, \ldots, x_n) : x_1^2 + x_2^2 + \ldots + x_n^2 \leq 1\}.$$

By definition, the joint density of $X_1, X_2, \ldots, X_n$ is

$$f(x_1, x_2, \ldots, x_n) = c, -1 \leq x_i \leq 1, 1 \leq i \leq n,$$

where $c = \frac{1}{2^n}$. Also, by the definition of probability,

$$P(\mathbf{X} \in B_n) = \int_{B_n} c dx_1 dx_2 \ldots dx_n$$

$$= \frac{\text{Vol}(B_n)}{2^n},$$

where $\text{Vol}(B_n)$ is the volume of the n-dimensional unit sphere B_n and equals

$$\text{Vol}(B_n) = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2} + 1\right)}.$$

Thus, finally,

$$P(\mathbf{X} \in B_n) = \frac{\pi^{\frac{n}{2}}}{2^n \Gamma\left(\frac{n}{2} + 1\right)}.$$

This is a very pretty formula. Let us evaluate this probability for various values of n and examine the effect of increasing the number of dimensions on this probability. Here is a table.

n	$P(\mathbf{X} \in B_n)$
2	.785
3	.524
4	.308
5	.164
6	.081
10	.002
12	.0003
15	.00001
18	3.13×10^{-7}

We see that in ten dimensions there is a 1 in 500 chance that a uniform random vector will fall in the central inscribed sphere, and in 18 dimensions the chance is much less than one in a million. *Therefore, when you are dealing with a large number of random variables at the same time, you will need a huge amount of sample data to learn about the behavior of their joint distribution in the central region; most of the data will come from the corners! You must have a huge amount of data to have at least some data points in your central region. This phenomenon has been termed the curse of dimensionality.*

12.2 Expectation of Functions

Expectations for multidimensional densities are defined analogously to the one-dimensional case. Here is the definition.

Definition 12.3. Let $(X_1, X_2, \ldots, X_n)$ have a joint density function $f(x_1, x_2, \ldots, x_n)$ and let $g(x_1, x_2, \ldots, x_n)$ be a real-valued function of $x_1, x_2, \ldots, x_n$. We say that the expectation of $g(X_1, X_2, \ldots, X_n)$ exists if

$$\int_{\mathcal{R}^n} |g(x_1, x_2, \ldots, x_n)| f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n < \infty,$$

in which case the expected value of $g(X_1, X_2, \ldots, X_n)$ is defined as

$$E[g(X_1, X_2, \ldots, X_n)] = \int_{\mathcal{R}^n} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n.$$

Remark. It is clear from the definition that the expectation of each individual X_i can be evaluated by either interpreting X_i as a function of the full vector $(X_1, X_2, \ldots, X_n)$ or by simply using the marginal density $f_i(x)$ of X_i; that is,

$$E(X_i) = \int_{\mathcal{R}^n} x_i f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n$$

$$= \int_{-\infty}^{\infty} x f_i(x) dx.$$

A similar comment applies to any function $h(X_i)$ of just X_i alone. All the properties of expectations that we have previously established—for example, the linearity of expectations—continue to hold in the multidimensional case. Thus,

$$E[ag(X_1, X_2, \ldots, X_n) + bh(X_1, X_2, \ldots, X_n)]$$
$$= aE[g(X_1, X_2, \ldots, X_n)] + bE[h(X_1, X_2, \ldots, X_n)].$$

We work out some examples now.

Example 12.10 (Bivariate Uniform). Two numbers X and Y are picked independently at random from $[0, 1]$. What is the expected distance between them? Thus, if X and Y are independent $U[0, 1]$, we want to compute $E(|X - Y|)$, which is

$$\begin{aligned} E(|X - Y|) &= \int_0^1 \int_0^1 |x - y| dx dy \\ &= \int_0^1 \left[\int_0^y (y - x) dx + \int_y^1 (x - y) dx \right] dy \\ &= \int_0^1 \left[\left(y^2 - \frac{y^2}{2} \right) + \left(\frac{1 - y^2}{2} - y(1 - y) \right) \right] dy \\ &= \int_0^1 \left[\frac{1}{2} - y + y^2 \right] dy \\ &= \frac{1}{2} - \frac{1}{2} + \frac{1}{3} = \frac{1}{3}. \end{aligned}$$

Example 12.11 (Uniform in a Triangle). Let (X, Y) have the uniform density

$$f(x, y) = 2 \text{ if } x, y \geq 0, x + y \leq 1,$$

and zero otherwise.

We have previously worked out the marginal density of X to be $f_1(x) = 2(1-x)$, $0 \leq x \leq 1$. Therefore,

$$E(X) = \int_0^1 2x(1 - x) dx = \frac{1}{3}.$$

The marginal expectation of Y is also $\frac{1}{3}$. Let us next calculate the variance of X and Y. The second moment of X is

$$E(X^2) = 2\int_0^1 x^2(1-x)dx = \frac{1}{6}.$$

Therefore, $\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}$, which is also the variance of Y.

What is the expected value of XY? By the definition of expectation,

$$E(XY) = 2\int_0^1 \int_0^{1-y} xydxdy = \int_0^1 y(1-y)^2 dy = \frac{1}{12}.$$

Therefore,

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{12} - \frac{1}{9} = -\frac{1}{36}.$$

Therefore, the correlation of X and Y is

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{-\frac{1}{36}}{\frac{1}{18}} = -\frac{1}{2}.$$

Example 12.12 (Independent Exponentials). Suppose X and Y are independently distributed as $Exp(\lambda)$ and $Exp(\mu)$, respectively. We want to find the expectation of the minimum of X and Y. The calculation below requires patience, but is not otherwise difficult.

Denote $W = \min\{X, Y\}$. Then,

$$\begin{aligned}
E(W) &= \int_0^\infty \int_0^\infty \min\{x,y\} \frac{1}{\lambda\mu} e^{-x/\lambda} e^{-y/\mu} dxdy \\
&= \int_0^\infty \int_0^y x \frac{1}{\lambda\mu} e^{-x/\lambda} e^{-y/\mu} dxdy + \int_0^\infty \int_y^\infty y \frac{1}{\lambda\mu} e^{-x/\lambda} e^{-y/\mu} dxdy \\
&= \int_0^\infty \frac{1}{\mu} e^{-y/\mu} \left[\int_0^y x \frac{1}{\lambda} e^{-x/\lambda} dx \right] dy \\
&\quad + \int_0^\infty \frac{1}{\mu} e^{-y/\mu} \left[\int_y^\infty y \frac{1}{\lambda} e^{-x/\lambda} dx \right] dy \\
&= \int_0^\infty \frac{1}{\mu} e^{-y/\mu} [\lambda - \lambda e^{-y/\lambda} - y e^{-y/\lambda}] dy \\
&\quad + \int_0^\infty \frac{1}{\mu} e^{-y/\mu} y e^{-y/\lambda} dy
\end{aligned}$$

(on integrating the x integral in the first term by parts)

$$= \frac{\lambda\mu^2}{(\lambda+\mu)^2} + \frac{\mu\lambda^2}{(\lambda+\mu)^2}$$

(once again, integrating by parts)

$$= \frac{\lambda\mu}{\lambda+\mu} = \frac{1}{\frac{1}{\lambda}+\frac{1}{\mu}},$$

a very pretty result.

Example 12.13 (Use of Polar Coordinates). Suppose a point (x, y) is picked at random from inside the unit circle. We want to find its expected distance from the center of the circle.

Thus, let (X, Y) have the joint density

$$f(x, y) = \frac{1}{\pi}, x^2 + y^2 \le 1,$$

and zero otherwise.

We will find $E[\sqrt{X^2+Y^2}]$. By definition,

$$E[\sqrt{X^2+Y^2}] = \frac{1}{\pi}\int_{(x,y):x^2+y^2\le 1} \sqrt{x^2+y^2}dxdy.$$

It is now very useful to make a transformation by using the polar coordinates

$$x = r\cos\theta, y = r\sin\theta,$$

with $dxdy = rdrd\theta$. Therefore,

$$\begin{aligned} E[\sqrt{X^2+Y^2}] &= \frac{1}{\pi}\int_{(x,y):x^2+y^2\le 1} \sqrt{x^2+y^2}dxdy \\ &= \frac{1}{\pi}\int_0^1\int_{-\pi}^{\pi} r^2 d\theta dr \\ &= 2\int_0^1 r^2 dr = \frac{2}{3}. \end{aligned}$$

We will later see in various calculations about finding distributions of functions of many continuous variables that transformation to polar and spherical coordinates often simplifies the integrations involved.

Example 12.14 (A Spherically Symmetric Density). Suppose (X, Y) has a joint density function $f(x, y) = \frac{c}{(1+x^2+y^2)^{\frac{3}{2}}}, x, y \geq 0$, where c is a positive normalizing constant. We will prove below that this is a valid joint density and evaluate the normalizing constant c. Note that $f(x, y)$ depends on x, y only through $x^2 + y^2$; such a density function is called *spherically symmetric* because the density $f(x, y)$ takes the same value at *all points* on the perimeter of a circle given by $x^2 + y^2 = k$.

To prove that f is a valid density, first note that it is obviously nonnegative. Next, by making a transformation to polar coordinates, $x = r\cos\theta, y = r\sin\theta$,

$$\int_{x>0,y>0} f(x, y)dxdy = c\int_0^\infty \int_0^{\frac{\pi}{2}} \frac{r}{(1 + r^2)^{\frac{3}{2}}} d\theta dr$$

(here, $0 \leq \theta \leq \frac{\pi}{2}$, as x and y are both positive)

$$= c\frac{\pi}{2}\int_0^\infty \frac{r}{(1 + r^2)^{\frac{3}{2}}} dr = c\frac{\pi}{2} \times 1 = c\frac{\pi}{2} \Rightarrow c = \frac{2}{\pi}.$$

We show that $E(X)$ does not exist. Note that it will then follow that $E(Y)$ also does not exist because $f(x, y) = f(y, x)$ in this example. The expected value of X, again by transforming to polar coordinates, is

$$E(X) = \frac{2}{\pi}\int_0^\infty \int_0^{\frac{\pi}{2}} \frac{r^2}{(1 + r^2)^{\frac{3}{2}}} \cos\theta d\theta dr$$

$$= \frac{2}{\pi}\int_0^\infty \frac{r^2}{(1 + r^2)^{\frac{3}{2}}} dr = \infty,$$

because the final integrand $\frac{r^2}{(1+r^2)^{\frac{3}{2}}}$ behaves like the function $\frac{1}{r}$ for large r and $\int_k^\infty \frac{1}{r} dr$ diverges for any positive k.

12.3 Bivariate Normal

The bivariate normal density is one of the most important densities for two jointly distributed continuous random variables, just like the univariate normal density is for one continuous variable. Many correlated random variables across the applied and social sciences are approximately distributed as bivariate normal. A typical example is the joint distribution of two size variables, such as height and weight.

Definition 12.4. The function $f(x, y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}, -\infty < x, y < \infty$ is called the *bivariate standard normal density*.

Clearly, we see that $f(x, y) = \phi(x)\phi(y)\ \forall\, x, y$. Therefore, the bivariate standard normal distribution corresponds to a pair of independent standard normal variables X, Y. If we make a linear transformation

$$U = \mu_1 + \sigma_1 X,$$
$$V = \mu_2 + \sigma_2[\rho X + \sqrt{1-\rho^2}Y],$$

then we get the general *five-parameter bivariate normal density* with means μ_1, μ_2, standard deviations σ_1, σ_2, and correlation $\rho_{U,V} = \rho$; here, $-1 < \rho < 1$.

Definition 12.5. The density of the five-parameter bivariate normal distribution is

$$f(u, v) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]},$$

$-\infty < u, v < \infty$.

If $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1$, then the bivariate normal density has just the parameter ρ, and it is denoted as $SBVN(\rho)$.

If we sample observations from a general bivariate normal distribution and plot the data points as points in the plane, then they would roughly plot out to an elliptical shape. The reason for this approximate elliptical shape is that the exponent in the formula for the density function is a quadratic form in the variables. In Figure 12.4, plot is given of a simulation of 1000 values from a bivariate normal distribution. The roughly elliptical shape is clear. It is also seen in the plot that the center of the point cloud is quite close to the true means of the variables, which were chosen to be $\mu_1 = 4.5, \mu_2 = 4$.

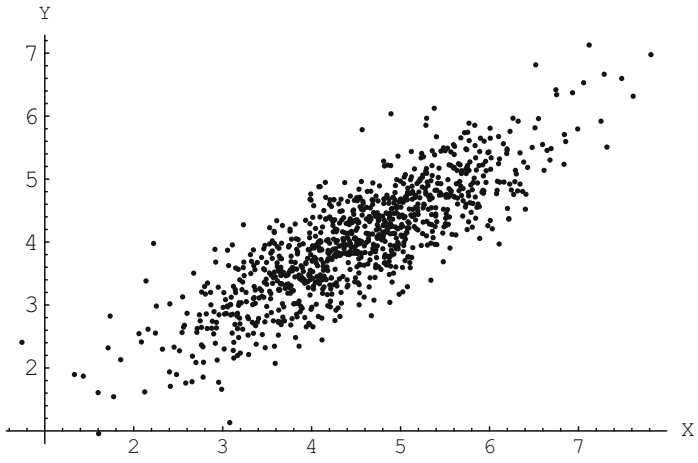


Fig. 12.4 Simulation of a bivariate normal with means 4, 5, 4; variance 1; correlation 75

From the representation we have given above of the general bivariate normal vector (U, V) in terms of independent standard normals X, Y, it follows that

$$E(UV) = \rho\sigma_1\sigma_2 + \mu_1\mu_2 \Rightarrow \text{Cov}(U, V) = \rho\sigma_1\sigma_2.$$

The symmetric matrix with the variances as diagonal entries and the covariance as the off-diagonal entry is called the *variance covariance matrix*, the *dispersion matrix*, or sometimes simply the *covariance matrix* of (U, V). Thus, the covariance matrix of (U, V) is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

A plot of the $SBVN(\rho)$ density is provided in Figure 12.5 for $\rho = 0, .5$; the zero-correlation case corresponds to independence. We see from the plots that the bivariate density has a unique peak at the mean point $(0, 0)$ and falls off from that point like a mound. The higher the correlation, the more the density concentrates near a plane. In the limiting case, when $\rho = \pm 1$, the density becomes fully concentrated on a plane, and we call it a *singular bivariate normal*.

When $\rho = 0$, the bivariate normal density does factorize into the product of the two marginal densities. Therefore, if $\rho = 0$, then U and V are actually independent, so, in that case, $P(U > \mu_1, V > \mu_2) = P$ (Each variable is larger than its mean value) $= \frac{1}{2}\frac{1}{2} = \frac{1}{4}$. When the parameters are general, one has the following classic formula.

Theorem 12.1 (A Classic Bivariate Normal Formula). *Let (U, V) have the five-parameter bivariate normal density with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$. Then,*

$$P(U > \mu_1, V > \mu_2) = P(U < \mu_1, V < \mu_2) = \frac{1}{4} + \frac{\arcsin\rho}{2\pi}.$$

A derivation of this formula can be seen in Tong (1990).

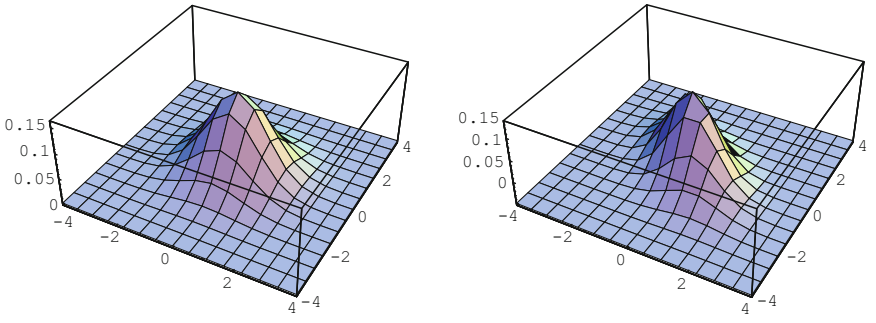


Fig. 12.5 Bivariate normal densities with zero means, unit variances, and rho = 0, .5

Example 12.15. Suppose a bivariate normal vector (U, V) has correlation ρ. Then, by applying the formula above, whatever σ_1, σ_2,

$$P(U > \mu_1, V > \mu_2) = 1/4 + 1/(2\pi)\arcsin\left[\frac{1}{2}\right] = \frac{1}{3}$$

when $\rho = \frac{1}{2}$. When $\rho = .75$, the probability increases to .385. In the limit when $\rho \to 1$, the probability tends to .5. That is, when $\rho \to 1$, all the probability becomes confined to the first and third quadrants $\{U > \mu_1, V > \mu_2\}$ and $\{U < \mu_1, V < \mu_2\}$, with the probability of each of these two quadrants approaching .5.

Another important property of a bivariate normal distribution is the following result.

Theorem 12.2. *Let (U, V) have a general five-parameter bivariate normal distribution. Then, any linear function $aU + bV$ of (U, V) is normally distributed:*

$$aU + bV \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sigma_1\sigma_2).$$

In particular, each of U, V is marginally normally distributed:

$$U \sim N(\mu_1, \sigma_1^2), V \sim N\left(\mu_2, \sigma_2^2\right).$$

If $\rho = 0$, then U and V are independent *with $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ marginal distributions.*

Proof. First note that $E(aU + bV) = a\mu_1 + b\mu_2$ by linearity of expectations, and $\text{Var}(aU + bV) = a^2\text{Var}(U) + b^2\text{Var}(V) + 2ab\text{Cov}(U, V)$ by the general formula for the variance of a linear combination of two jointly distributed random variables (see Chapter 11). But $\text{Var}(U) = \sigma_1^2, \text{Var}(V) = \sigma_2^2$, and $\text{Cov}(U, V) = \rho\sigma_1\sigma_2$. Therefore, $\text{Var}(aU + bV) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sigma_1\sigma_2$.

Therefore, we only have to prove that $aU + bV$ is normally distributed. For this, we use our representation of U, V in terms of a pair of independent standard normal variables X, Y:

$$\begin{aligned} U &= \mu_1 + \sigma_1 X, \\ V &= \mu_2 + \sigma_2[\rho X + \sqrt{1-\rho^2}Y]. \end{aligned}$$

Multiplying the equations by a, b and adding, we get the representation

$$\begin{aligned} aU + bV &= a\mu_1 + b\mu_2 + [a\sigma_1 X + b\sigma_2\rho X + b\sigma_2\sqrt{1-\rho^2}Y] \\ &= a\mu_1 + b\mu_2 + [(a\sigma_1 + b\sigma_2\rho)X + b\sigma_2\sqrt{1-\rho^2}Y]. \end{aligned}$$

That is, $aU + bV$ can be represented as a linear function $cX + dY + k$ of two independent standard normal variables X and Y, so $aU + bV$ is necessarily normally distributed (see Chapter 9).

In fact, a result stronger than the previous theorem holds. What is true is that *any two* linear functions of U, V will again be distributed as a bivariate normal. Here is the stronger result.

Theorem 12.3. *Let (U, V) have a general five-parameter bivariate normal distribution. Let $Z = aU + bV$ and $W = cU + dV$ be two linear functions such that $ad - bc \neq 0$. Then, (Z, W) also has a bivariate normal distribution, with parameters*

$$\begin{aligned} E(Z) &= a\mu_1 + b\mu_2, E(W) = c\mu_1 + d\mu_2; \\ \mathrm{Var}(Z) &= a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sigma_1\sigma_2; \\ \mathrm{Var}(W) &= c^2\sigma_1^2 + d^2\sigma_2^2 + 2cd\rho\sigma_1\sigma_2; \\ \rho_{Z,W} &= \frac{ac\sigma_1^2 + bd\sigma_2^2 + (ad + bc)\rho\sigma_1\sigma_2}{\sqrt{\mathrm{Var}(Z)\mathrm{Var}(W)}}. \end{aligned}$$

The proof of this theorem is similar to the proof of the previous theorem, and the details are omitted.

Example 12.16 (Independence of Mean and Variance). Suppose X_1 and X_2 are two iid $N(\mu, \sigma^2)$ variables. Then, of course, they are also jointly bivariate normal. Now define two linear functions

$$Z = X_1 + X_2, W = X_1 - X_2.$$

Since (X_1, X_2) has a bivariate normal distribution, so does (Z, W). However, plainly,

$$\mathrm{Cov}(Z, W) = \mathrm{Cov}(X_1 + X_2, X_1 - X_2) = \mathrm{Var}(X_1) - \mathrm{Var}(X_2) = 0.$$

Therefore, Z and W must actually be independent. As a consequence, Z and W^2 are also independent. Now note that the sample variance of X_1, X_2 is

$$s^2 = \left(X_1 - \frac{X_1 + X_2}{2}\right)^2 + \left(X_2 - \frac{X_1 + X_2}{2}\right)^2 = \frac{(X_1 - X_2)^2}{2} = \frac{W^2}{2}.$$

And, of course, $\bar{X} = \frac{X_1 + X_2}{2} = \frac{Z}{2}$. Therefore, it follows that $\bar{X}$ and s^2 are independent.

This is true not just for two observations but for any number of iid observations from a normal distribution. Here is the general result, which cannot be proved without introducing additional facts about distribution theory.

Theorem 12.4. *Let $X_1, X_2, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ variables. Then $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are independently distributed.*

Example 12.17 (Normal Marginals Do Not Guarantee Joint Normality). Although joint bivariate normality of two random variables implies that each variable must be marginally univariate normal, the converse in general is not true.

Let $Z \sim N(0,1)$ and let U be a two-valued random variable with the pmf $P(U = \pm 1) = \frac{1}{2}$. Take U and Z to be independent. Now define $X = U|Z|$ and $Y = Z$.

Then, each of X, Y has a standard normal distribution. That X has a standard normal distribution is easily seen in many ways, for example, just by evaluating its CDF. Take $x > 0$. Then,

$$\begin{aligned} P(X \le x) &= P(X \le x|U = -1) \times \frac{1}{2} + P(X \le x|U = 1) \times \frac{1}{2} \\ &= 1 \times \frac{1}{2} + P(|Z| \le x) \times \frac{1}{2} \\ &= \frac{1}{2} + \frac{1}{2} \times [2\Phi(x) - 1] = \Phi(x); \end{aligned}$$

similarly also for $x \le 0$, $P(X \le x) = \Phi(x)$.

But, jointly, X, Y cannot be bivariate normal because $X^2 = U^2 Z^2 = Z^2 = Y^2$ with probability 1. That is, the joint distribution of (X, Y) *lives on just the two lines* $y = \pm x$ and so is certainly not bivariate normal.

12.4 Conditional Densities and Expectations

The conditional distribution for continuous random variables is defined analogously to the discrete case, with pmfs replaced by densities. The formal definitions are as follows.

Definition 12.6 (Conditional Density). Let (X, Y) have a joint density $f(x, y)$. The *conditional density* of X given $Y = y$ is defined as

$$f(x|y) = f(x|Y = y) = \frac{f(x, y)}{f_Y(y)}, \ \forall y \text{ such that } f_Y(y) > 0.$$

The *conditional expectation* of X given $Y = y$ is defined as

$$\begin{aligned} E(X|y) = E(X|Y = y) &= \int_{-\infty}^{\infty} x f(x|y) dx \\ &= \frac{\int_{-\infty}^{\infty} x f(x, y) dx}{\int_{-\infty}^{\infty} f(x, y) dx}, \end{aligned}$$

$\forall y$ such that $f_Y(y) > 0$.

For fixed x, the conditional expectation $E(X|y) = \mu_X(y)$ is a number. As we vary y, we can think of $E(X|y)$ as a function of y. The corresponding function of Y is written as $E(X|Y)$ and is a random variable. It is very important to keep this notational distinction in mind.

The conditional density of Y given $X = x$ and the conditional expectation of Y given $X = x$ are defined analogously. That is, for instance,

$$f(y|x) = \frac{f(x,y)}{f_X(x)}, \ \forall x \text{ such that } f_X(x) > 0.$$

An important relationship connecting the two conditional densities is the following result.

Theorem 12.5 (Bayes' Theorem for Conditional Densities). *Let (X, Y) have a joint density $f(x, y)$. Then, $\forall x, y$ such that $f_X(x) > 0$, $f_Y(y) > 0$,*

$$f(y|x) = \frac{f(x|y)\, f_Y(y)}{f_X(x)}.$$

Proof.

$$\begin{aligned} \frac{f(x|y)\, f_Y(y)}{f_X(x)} &= \frac{\frac{f(x,y)}{f_Y(y)}\, f_Y(y)}{f_X(x)} \\ &= \frac{f(x,y)}{f_X(x)} = f(y|x). \end{aligned}$$

Thus, we can convert one conditional density to the other one by using Bayes' theorem; note the similarity to Bayes' theorem discussed in Chapter 3.

Definition 12.7 (Conditional Variance). Let (X, Y) have a joint density $f(x, y)$. The *conditional variance* of X given $Y = y$ is defined as

$$\text{Var}(X|y) = \text{Var}(X|Y = y) = \frac{\int_{-\infty}^{\infty} (x - \mu_X(y))^2 f(x,y)dx}{\int_{-\infty}^{\infty} f(x,y)dx},$$

$\forall y$ such that $f_Y(y) > 0$, where $\mu_X(y)$ denotes $E(X|y)$.

Remark. All the facts and properties about conditional pmfs and conditional expectations that were presented in the previous chapter for discrete random variables continue to hold verbatim in the continuous case, with densities replacing the pmfs in their statements. In particular, the iterated expectation and variance formula, and all the rules about conditional expectations and variance in Section 11.3, hold in the continuous case.

An important optimizing property of the conditional expectation is that the best predictor of Y based on X among all possible predictors is the conditional expectation of Y given X. Here is the exact result.

Proposition (Best Predictor). Let (X, Y) be jointly distributed random variables (of any kind). Suppose $E(Y^2) < \infty$. Then $E_{X,Y}[(Y - E(Y|X))^2] \leq E_{X,Y}[(Y - g(X))^2]$ for any function $g(X)$. Here, the notation $E_{X,Y}$ stands for expectation with respect to the joint distribution of X, Y.

Proof. Denote $\mu_Y(x) = E(Y|X = x)$. Then, by the property of the mean of any random variable U that $E(U - E(U))^2 \leq E(U - a)^2$ for any a, we get that here

$$E[(Y - \mu_Y(x))^2|X = x] \leq E[(Y - g(x))^2|X = x]$$

for any x.

Since this inequality holds for any x, it will also hold on taking an expectation,

$$E_X[E[(Y - \mu_Y(x))^2|X = x]] \leq E_X[E[(Y - g(x))^2|X = x]]$$
$$\Rightarrow E_{X,Y}[(Y - \mu_Y(X))^2] \leq E_{X,Y}[(Y - g(X))^2],$$

where the final line is a consequence of the iterated expectation formula (see Chapter 11).

We will now see a number of examples.

12.4.1 Examples on Conditional Densities and Expectations

Example 12.18 (Uniform in a Triangle). Consider the joint density

$$f(x, y) = 2 \text{ if } x, y \geq 0, x + y \leq 1.$$

By using the results derived in Example 12.2,

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{1}{1 - y}$$

if $0 \leq x \leq 1 - y$ and is zero otherwise. Thus, we have the interesting conclusion that, given $Y = y$, X is distributed uniformly in $[0, 1 - y]$. Consequently,

$$E(X|y) = \frac{1 - y}{2}, \ \forall y, 0 < y < 1.$$

Also, the conditional variance of X given $Y = y$ is, by the general variance formula for uniform distributions,

$$\text{Var}(X|y) = \frac{(1 - y)^2}{12}.$$

Example 12.19 (Uniform Distribution in a Circle). Let (X, Y) have a uniform density in the unit circle, $f(x, y) = \frac{1}{\pi}$, $x^2 + y^2 \le 1$. We will find the conditional expectation of X given $Y = y$. First, the conditional density is

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{\frac{1}{\pi}}{\frac{2\sqrt{1-y^2}}{\pi}} = \frac{1}{2\sqrt{1-y^2}} - \sqrt{1-y^2} \le x \le \sqrt{1-y^2}.$$

Thus, we have the interesting result that the conditional density of X given $Y = y$ is uniform on $[-\sqrt{1-y^2}, \sqrt{1-y^2}]$. It being an interval symmetric about zero, we have in addition the result that, for any y, $E(X|Y = y) = 0$.

Let us now find the conditional variance. Since the conditional distribution of X given $Y = y$ is uniform on $[-\sqrt{1-y^2}, \sqrt{1-y^2}]$, by the general variance formula for uniform distributions,

$$\text{Var}(X|y) = \frac{(2\sqrt{1-y^2})^2}{12} = \frac{1-y^2}{3}.$$

Thus, the conditional variance decreases as y moves away from zero, which makes sense intuitively because, as y moves away from zero, the line segment in which x varies becomes smaller.

Example 12.20 (A Two-Stage Experiment). Suppose X is a positive random variable with density $f(x)$, and given $X = x$, a number Y is chosen at random between 0 and x. Suppose, however, that you are *only* told the value of Y and the x value is kept hidden from you. What is your guess for x?

The formulation of the problem is

$$X \sim f(x); Y|X = x \sim U[0, x]; \text{ and we want to find } E(X|Y = y).$$

To find $E(X|Y = y)$, our first task would be to find $f(x|y)$, the conditional density of X given $Y = y$. This is, by its definition,

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f(y|x)f(x)}{f_Y(y)} = \frac{\frac{1}{x} I_{\{x \ge y\}} f(x)}{\int_y^\infty \frac{1}{x} f(x)dx}.$$

Therefore,

$$E(X|Y = y) = \int_y^\infty x f(x|y)dx = \frac{\int_y^\infty x \frac{1}{x} f(x)dx}{\int_y^\infty \frac{1}{x} f(x)dx} = \frac{1 - F(y)}{\int_y^\infty \frac{1}{x} f(x)dx},$$

where F denotes the CDF of X.

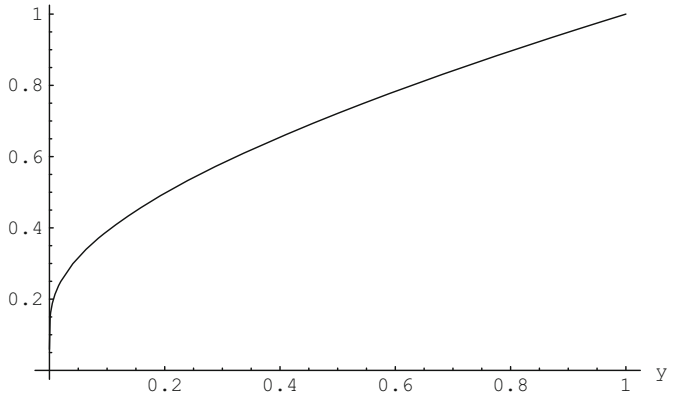


Fig. 12.6 Plot of $E(X|Y = y)$ when x is $U[0, 1]$, $Y|X = x$ *is* $U[0, x]$

Suppose now, in particular, that $f(x)$ is the $U[0, 1]$ density. Then, by plugging into this general formula,

$$E(X|Y = y) = \frac{1 - F(y)}{\int_y^\infty \frac{1}{x} f(x)dx} = \frac{1 - y}{-\log y}, 0 < y < 1.$$

The important thing to note is that although X has marginally a uniform density and expectation $\frac{1}{2}$, given $Y = y$, X is not uniformly distributed and $E(X|Y = y)$ is *not* $\frac{1}{2}$. Indeed, as the plot in Figure 12.6 shows, $E(X|Y = y)$ is an increasing function of y, increasing from zero at $y = 0$ to one at $y = 1$.

Example 12.21 ($E(X|Y = y)$ exists for any y, but $E(X)$ does not). Consider the setup of the preceding example once again, $X \sim f(x)$, and given $X = x, Y \sim U[0, x]$. Suppose $f(x) = \frac{1}{x^2}, x \geq 1$. Then the marginal expectation $E(X)$ does not exist because $\int_1^\infty x \frac{1}{x^2} dx = \int_1^\infty \frac{1}{x} dx$ diverges.

However, from the general formula in the preceding example,

$$E(X|Y = y) = \frac{1 - F(y)}{\int_y^\infty \frac{f(x)}{x} dx} = \frac{\frac{1}{y}}{\frac{1}{2y^2}} = 2y,$$

and thus $E(X|Y = y)$ exists for every y.

Example 12.22 (Using Conditioning to Evaluate Probabilities). We described in the last chapter the iterated expectation technique to calculate expectations. It turns out that it is in fact also a really useful way to calculate probabilities. The reason is that the probability of any event A is also the expectation of $X = I_A$, so, by the iterated expectation technique, we can calculate $P(A)$ as

$$P(A) = E(I_A) = E(X) = E_Y[E(X|Y = y)] = E_Y[P(A|Y = y)]$$

by using a conditioning variable Y judiciously. The choice of the conditioning variable Y is usually clear from the particular context. Here is an example.

Let X and Y be independent $U[0,1]$ random variables. Then $Z = XY$ also takes values in $[0,1]$, and suppose we want to find an expression for $P(Z \le z)$. We can do this by using the iterated expectation technique

$$\begin{aligned} P(XY \le z) &= E[I_{XY \le z}] = E_Y[E(I_{XY \le z}|Y = y)] = E_Y[E(I_{Xy \le z}|Y = y)] \\ &= E_Y[E(I_{X \le \frac{z}{y}}|Y = y)] = E_Y[E(I_{X \le \frac{z}{y}})] \end{aligned}$$

(because X and Y are independent)

$$= E_Y\left[P\left(X \le \frac{z}{y}\right)\right].$$

Now, note that $P(X \le \frac{z}{y})$ is $\frac{z}{y}$ if $\frac{z}{y} \le 1 \Leftrightarrow y \ge z$, and $P(X \le \frac{z}{y}) = 1$ if $y < z$. Therefore,

$$E_Y\left[P\left(X \le \frac{z}{y}\right)\right] = \int_0^z 1dy + \int_z^1 \frac{z}{y}dy = z - z\log z, 0 < z \le 1.$$

So, the final answer to our problem is $P(XY \le z) = z - z\log z, 0 < z \le 1$.

Example 12.23 (Power of the Iterated Expectation Formula). Let X, Y, Z be three independent $U[0,1]$ random variables. We will find the probability that $X^2 \ge YZ$ by once again using the iterated expectation formula.

To do this,

$$\begin{aligned} P(X^2 \ge YZ) &= 1 - P(X^2 < YZ) \\ &= 1 - E[I_{X^2 < YZ}] = 1 - E_{Y,Z}[E(I_{X^2 < YZ}|Y = y, Z = z)] \\ &= 1 - E_{Y,Z}[E(I_{X^2 < yz}|Y = y, Z = z)] \\ &= 1 - E_{Y,Z}[E(I_{X^2 < yz})] \end{aligned}$$

(since X, Y, Z are independent)

$$\begin{aligned} &= 1 - E_{Y,Z}[P(X^2 < yz)] = 1 - E_{Y,Z}[\sqrt{yz}] \\ &= 1 - E_Y[\sqrt{Y}]E_Z[\sqrt{Z}] = 1 - \left(\frac{2}{3}\right)^2 = \frac{5}{9}. \end{aligned}$$

Once again, we see the power of identifying probabilities as expectations of indicator variables and using the iterated expectation formula.

Example 12.24 (Conditional Density Given the Sum). Suppose X and Y are two independent $Exp(1)$ variables. What is the conditional density of X given that $X + Y = t$? Denote $X + Y = T$. Then, we know from Chapter 8 that $T \sim G(2,1)$.

Also, by the definition of probabilities for jointly continuous random variables, by denoting the joint density of (X, Y) as $f(x, y)$,

$$\begin{aligned} P(X \le x, T \le t) &= \int_{u \le x, u+v \le t} f(u, v) du dv \\ &= \int_{0<u \le x, 0<u+v \le t} e^{-u-v} du dv \\ &= \int_0^x e^{-u} \left[\int_0^{t-u} e^{-v} dv \right] du \\ &= \int_0^x e^{-u}(1 - e^{u-t}) du \\ &= \int_0^x e^{-u} du - \int_0^x e^{-t} du \\ &= 1 - e^{-x} - x e^{-t} \end{aligned}$$

for $x > 0, t > x$.

Therefore, the joint density of X and T is

$$\begin{aligned} f_{X,T}(x, t) &= \frac{\partial^2}{\partial x \partial t}[1 - e^{-x} - x e^{-t}] \\ &= e^{-t}, 0 < x < t < \infty. \end{aligned}$$

Now therefore, from the definition of conditional densities,

$$f(x|t) = \frac{f_{X,T}(x, t)}{f_T(t)} = \frac{e^{-t}}{t e^{-t}} = \frac{1}{t},$$

$0 < x < t$.
That is, given that the sum $X + Y = t$, X is distributed uniformly on $[0, t]$. In particular,

$$E(X|X + Y = t) = \frac{t}{2}, \ \mathrm{Var}(X|X + Y = t) = \frac{t^2}{12}.$$

To complete the example, we mention a quick trick to compute the conditional expectation. Note that, by symmetry,

$$\begin{aligned} E(X|X + Y = t) = E(Y|X + Y = t) &\Rightarrow t = E(X + Y|X + Y = t) \\ = 2E(X|X + Y = t) &\Rightarrow E(X|X + Y = t) = \frac{t}{2}. \end{aligned}$$

So, if we wanted just the conditional expectation, then the conditional density calculation was not necessary in this case. This sort of symmetry argument is often very

useful in reducing algebraic calculations. But one needs to be absolutely sure that the symmetry argument will be valid in a given problem.

Example 12.25 (Maximum Given the Minimum). Suppose U, V, W are three independent random variables, each distributed as $U[0, 1]$. Let $X = \max\{U, V, W\}$ and $Y = \min\{U, V, W\}$. We want to find the conditional density and then the conditional expectation of X given $Y = y$.

As a side remark, the ordered values among a set of observations from some general distribution are called *the order statistics of the sample*, and we will treat the joint distribution of the order statistics in greater detail in a later section. What we want to do here is examine the conditional density and expectation of the largest order statistic given the smallest in this specific example. The general treatment is not done in this section.

First, note that, for $0 < y \le x < 1$,

$$P(X \le x, Y \ge y) = P(y \le U \le x, y \le V \le x, y \le W \le x) = (x - y)^3,$$

since U, V, W are iid $U[0, 1]$. Therefore,

$$\begin{aligned} P(X \le x, Y \le y) = P(X \le x) - P(X \le x, Y \ge y) = P(U \le x, V \le x, W \le x) \\ - P(X \le x, Y \ge y) = x^3 - (x - y)^3. \end{aligned}$$

By taking partial derivatives, the joint density of (X, Y) is

$$f(x, y) = \frac{\partial^2}{\partial x \partial y}[x^3 - (x - y)^3] = 6(x - y), 0 < y \le x < 1.$$

Furthermore,

$$\begin{aligned} P(Y \ge y) = P(U \ge y, V \ge y, W \ge y) = (1 - y)^3 \\ \Rightarrow f_Y(y) = 3(1 - y)^2, 0 < y < 1. \end{aligned}$$

Thus, the conditional density of X given $Y = y$ is

$$f(x|y) = \frac{6(x - y)}{3(1 - y)^2} = \frac{2(x - y)}{(1 - y)^2},$$

$0 < y \le x < 1$.

From the conditional density,

$$\begin{aligned} E(X|Y = y) &= \int_y^1 x \frac{2(x - y)}{(1 - y)^2} dx = \frac{2}{(1 - y)^2} \int_y^1 x(x - y) dx \\ &= \frac{2}{(1 - y)^2} \frac{y^3 - 3y + 2}{6} = \frac{y^3 - 3y + 2}{3(1 - y)^2} = \frac{y + 2}{3}. \end{aligned}$$

We see that $E(X|Y = y)$ is increasing in y. From our general discussion of covariance (see Chapter 11), it follows that $\text{Cov}(X, Y) \geq 0$ in this example, and as a result $\rho_{X,Y}$ will also be ≥ 0. Actually, this phenomenon of a positive correlation *between any two order statistics, not just the minimum and the maximum*, is true in general.

12.5 Bivariate Normal Conditional Distributions

Suppose (X, Y) have a joint bivariate normal distribution. A very important property of the bivariate normal is that *each conditional distribution*, the distribution of Y given $X = x$, and that of X given $Y = y$ is a univariate normal for any x and any y. This really helps in easily computing conditional probabilities involving one variable when the other variable is held fixed at some specific value.

Theorem 12.6. *Let (X, Y) have a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$. Then,*

(a) $X|Y = y \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right)$;

(b) $Y|X = x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)\right)$.

In particular, the conditional expectations of X given $Y = y$ and Y given $X = x$ are linear functions of y and x, respectively:

$$E(X|Y = y) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2),$$
$$E(Y|X = x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1),$$

and the variance of each conditional distribution is a constant and does not depend on the conditioning values x or y.

The proof of this theorem involves some tedious integration manipulations and we omit it; the details of the proof are available in Tong (1990).

Remark. We see here that the conditional expectation is linear in the bivariate normal case. Specifically, take $E(Y|X = x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$. Previously, we saw in Chapter 11 that the conditional expectation $E(Y|X)$ is, in general, the best predictor of Y based on X. Now we see that the conditional expectation is a linear predictor in the bivariate normal case, and it is the best predictor and therefore also the best linear predictor. In Chapter 11, we called the best linear predictor the regression line of Y on X. Putting it all together, we have the very special result that, in the bivariate normal case, the regression line of Y on X and the best overall predictor are the same:

For bivariate normal distributions, the conditional expectation of one variable given the other coincides with the regression line of that variable on the other variable.

Example 12.26. Suppose incomes of husbands and wives in a population are bivariate normal with means 75 and 60 (in thousands of dollars), standard deviations 20 each, and a correlation of .75. We want to know in what percentage of those families where the wife earns 80,000 dollars the family income exceeds 175,000 dollars.

Denote the income of the husband and the wife by X and Y, respectively. Then, we want to find $P(X + Y > 175|Y = 80)$. By the theorem above, $X|Y = 80 \sim N(75 + .75(80 - 60), 400(1 - .75^2)) = N(90, 175)$. Therefore,

$$P(X + Y > 175|Y = 80) = P(X > 95|Y = 80)$$
$$= P\left(Z > \frac{95 - 90}{\sqrt{175}}\right) = P(Z > .38) = .3520,$$

where Z denotes a standard normal variable.

Example 12.27 (Galton's Observation: Regression to the Mean). This example is similar to the previous example, but makes a different interesting point. It is often found that students who get a very good grade on the first midterm do not do as well on the second midterm. We can try to explain it by doing a bivariate normal calculation.

Denote the grade on the first midterm by X, that on the second midterm by Y, and suppose X, Y are jointly bivariate normal with means 70, standard deviations 10, and correlation .7. Suppose a student scored 90 on the first midterm. What are the chances that he will get a lower grade on the second midterm?

This is

$$P(Y < X|X = 90) = P(Y < 90|X = 90)$$
$$= P\left(Z < \frac{90 - 84}{\sqrt{51}}\right) = P(Z < .84) = .7995,$$

where Z is a standard normal variable, and we have used the fact that $Y|X = 90 \sim N(70 + .7(90 - 70), 100(1 - .7^2)) = N(84, 51)$.

Thus, with a fairly high probability, the student will not be able to match his first midterm grade on the second midterm. The phenomenon of *regression to mediocrity* was popularized by Galton, who noticed that the offspring of very tall parents tended to be much closer to being of just about average height and the extreme tallness in the parents was not commonly passed on to the children.

12.6 Order Statistics

The ordered values of a sample of observations are called the order statistics of the sample, and the smallest and the largest are called the extremes. Order statistics and extremes are among the most important functions of a set of random variables that we study in probability and statistics. There is natural interest in studying the

highs and lows of a sequence, and the other order statistics help in understanding the concentration of probability in a distribution, or equivalently the diversity in the population represented by the distribution. Order statistics are also useful in statistical inference, where estimates of parameters are often based on some suitable functions of the order statistics. In particular, the median is of very special importance. There is a well-developed theory of the order statistics of a fixed number n of observations from a fixed distribution. Distribution theory for order statistics when the observations are from a discrete distribution is complex, both notationally and algebraically, because of the fact that there could be several observations that are actually equal. These ties among the sample values make the distribution theory cumbersome. We therefore concentrate on the continuous case.

12.6.1 Basic Distribution Theory

Definition 12.8. Let $X_1, X_2, \ldots, X_n$ be any n real-valued random variables. Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ denote the ordered values of $X_1, X_2, \ldots, X_n$. Then, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are called the *order statistics* of $X_1, X_2, \ldots, X_n$.

Remark. Thus, the minimum among $X_1, X_2, \ldots, X_n$ is the first order statistic and the maximum is the nth order statistic. The middle value among $X_1, X_2, \ldots, X_n$ is called the median. But it needs to be defined precisely because there is really no middle value when n is an even integer. Here is our definition.

Definition 12.9. Let $X_1, X_2, \ldots, X_n$ be any n real-valued random variables. Then, the *median* of $X_1, X_2, \ldots, X_n$ is defined to be $M_n = X_{(m+1)}$ if $n = 2m+1$ (an odd integer) and $M_n = X_{(m)}$ if $n = 2m$ (an even integer). That is, in either case, the median is the order statistic $X_{(k)}$, where k is the smallest integer $\ge \frac{n}{2}$.

Example 12.28. Suppose .3, .53, .68, .06, .73, .48, .87, .42, .89, .44 are ten independent observations from the $U[0,1]$ distribution. Then, the order statistics are .06, .3, .42, .44, .48, .53, .68, .73, .87, .89. Thus, $X_{(1)} = .06, X_{(n)} = .89$, and since $\frac{n}{2} = 5, M_n = X_{(5)} = .48$.

We now specialize to the case where $X_1, X_2, \ldots, X_n$ are independent random variables with a common density function $f(x)$ and CDF $F(x)$, and work out the fundamental distribution theory of the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$.

Theorem 12.7 (Joint Density of All the Order Statistics). *Let $X_1, X_2, \ldots, X_n$ be independent random variables with a common density function $f(x)$. Then, the joint density function of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is given by*

$$f_{1,2,\ldots,n}(y_1, y_2, \ldots, y_n) = n! f(y_1) f(y_2) \cdots f(y_n) I_{\{y_1 < y_2 < \cdots < y_n\}}.$$

Proof. A verbal heuristic argument is easy to understand. If $X_{(1)} = y_1, X_{(2)} = y_2, \ldots, X_{(n)} = y_n$, then exactly one of the sample values $X_1, X_2, \ldots, X_n$ is y_1, exactly one is y_2, etc., but we can put any of the n observations at y_1, any of

the other $n-1$ observations at y_2, etc., so the density of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is $f(y_1)f(y_2)\ldots f(y_n) \times n(n-1)\ldots 1 = n!f(y_1)f(y_2)\ldots f(y_n)$, and obviously, if the inequality $y_1 < y_2 < \cdots < y_n$ is not satisfied, then at such a point the joint density of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ must be zero.

Here is a formal proof. The multivariate transformation $(X_1, X_2, \ldots, X_n) \to (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ is not one-to-one, as any permutation of a fixed $(X_1, X_2, \ldots, X_n)$ vector has exactly the same set of order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. However, fix a specific permutation $\{\pi(1), \pi(2), \ldots, \pi(n)\}$ of $\{1, 2, \ldots, n\}$ and consider the subset $A_\pi = \{(x_1, x_2, \ldots, x_n) : x_{\pi(1)} < x_{\pi(2)} < \cdots < x_{\pi(n)}\}$. Then, the transformation $(x_1, x_2, \ldots, x_n) \to (x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ is one-to-one on each such A_π, and indeed, then $x_{(i)} = x_{\pi(i)}, i = 1, 2, \ldots, n$. The Jacobian matrix of the transformation is 1, for each such A_π. A particular vector $(x_1, x_2, \ldots, x_n)$ falls in exactly one A_π, and there are $n!$ such regions A_π, as we exhaust all the $n!$ permutations $\{\pi(1), \pi(2), \ldots, \pi(n)\}$ of $\{1, 2, \ldots, n\}$. By a modification of the Jacobian density theorem, we then get

$$\begin{aligned} f_{1,2,\ldots,n}(y_1, y_2, \ldots, y_n) &= \sum_\pi f(x_1)f(x_2)\ldots f(x_n) \\ &= \sum_\pi f(x_{\pi(1)})f(x_{\pi(2)})\ldots f(x_{\pi(n)}) \\ &= \sum_\pi f(y_1)f(y_2)\ldots f(y_n) \\ &= n!f(y_1)f(y_2)\ldots f(y_n). \end{aligned}$$

Example 12.29 (Uniform Order Statistics). Let $U_1, U_2, \ldots, U_n$ be independent $U[0, 1]$ variables and $U_{(i)}, 1 \le i \le n$, their order statistics. Then, by our theorem above, the joint density of $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ is

$$f_{1,2,\ldots,n}(u_1, u_2, \ldots, u_n) = n!I_{0<u_1<u_2<\cdots<u_n<1}.$$

Once we know the joint density of all the order statistics, we can find the marginal density of any subset of them by simply integrating out the rest of the coordinates but *being extremely careful in using the correct domain over which to integrate the rest of the coordinates*. For example, if we want the marginal density of just $U_{(1)}$, (that is, of the sample minimum), then we will want to integrate out $u_2, \ldots, u_n$, and the correct domain of integration would be, for a given u_1 in (0,1),

$$u_1 < u_2 < u_3 < \cdots < u_n < 1.$$

So, we will integrate down in the order $u_n, u_{n-1}, \ldots, u_2$, to obtain

$$\begin{aligned} f_1(u_1) &= n! \int_{u_1}^1 \int_{u_2}^1 \cdots \int_{u_{n-1}}^1 du_n du_{n-1} \cdots du_3 du_2 \\ &= n(1-u_1)^{n-1}, 0 < u_1 < 1. \end{aligned}$$

Likewise, if we want the marginal density of just $U_{(n)}$ (that is, of the sample maximum), then we will want to integrate out $u_1, u_2, \ldots, u_{n-1}$, and now the answer will be

$$\begin{aligned} f_n(u_n) &= n! \int_0^{u_n} \int_0^{u_{n-1}} \ldots \int_0^{u_2} du_1 du_2 \ldots du_{n-2} du_{n-1} \\ &= n u_n^{n-1}, 0 < u_n < 1. \end{aligned}$$

However, it is useful to note that for the special case of the minimum and the maximum, we could have obtained the densities much more easily and directly. Here is why. First take the maximum. Consider its CDF for $0 < u < 1$,

$$P(U_{(n)} \le u) = P(\cap_{i=1}^n \{X_i \le u\}) = \prod_{i=1}^n P(X_i \le u) = u^n,$$

and hence the density of $U_{(n)}$ is $f_n(u) = \frac{d}{du}[u^n] = nu^{n-1}, 0 < u < 1$.

Likewise, for the minimum, for $0 < u < 1$, the tail CDF is

$$P(U_{(1)} > u) = P(\cap_{i=1}^n \{X_i > u\}) = (1-u)^n,$$

so the density of $U_{(1)}$ is

$$f_1(u) = \frac{d}{du}[1 - (1-u)^n] = n(1-u)^{n-1}, 0 < u < 1.$$

For a general $r, 1 \le r \le n$, the density of $U_{(r)}$ works out to a Beta density,

$$f_r(u) = \frac{n!}{(r-1)!(n-r)!} u^{r-1}(1-u)^{n-r}, 0 < u < 1,$$

which is the $Be(r, n-r+1)$ density.

In Figure 12.7, we provide a plot of the density of the minimum, the median, and the maximum in the $U[0,1]$ case when $n = 15$. The minimum and the maximum clearly have skewed densities, while the density of the median is symmetric about .5.

12.6.2 ∗ More Advanced Distribution Theory

Somewhat more advanced calculations, such as derivation of the joint density of two different order statistics, are presented here. An example will help us understand the line of mathematical reasoning that we use in these calculations.

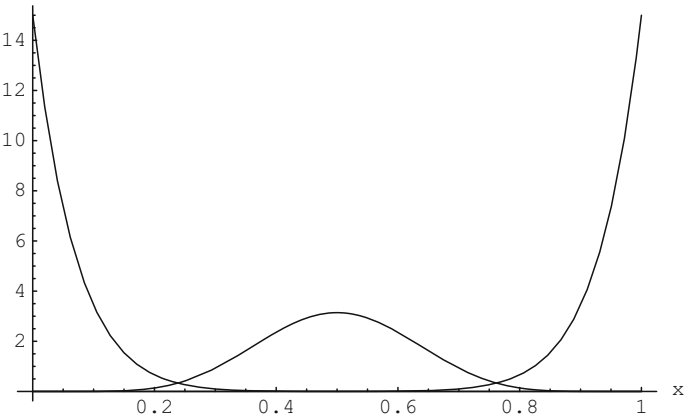


Fig. 12.7 Density of minimum, median, and maximum of $U[0, 1]$ variables; $n = 15$

Example 12.30 (Density of One and Two Order Statistics). The joint density of any subset of the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ can be worked out from the joint density of all the order statistics, which we have already worked out. The most important case in applications is the joint density of two specific order statistics, say $X_{(r)}$ and $X_{(s)}$, $1 \le r < s \le n$, or the density of a specific one, say $X_{(r)}$. A heuristic argument is the most helpful in understanding the formula for the joint density of $X_{(r)}$ and $X_{(s)}$ and also the density of a specific one $X_{(r)}$.

First consider the density of just $X_{(r)}$. Fix u. To have $X_{(r)} = u$, we must have exactly one observation at u, another $r-1$ below u, and another $n-r$ above u. This will *suggest* that the density of $X_{(r)}$ is

$$\begin{aligned} f_r(u) &= nf(u)\binom{n-1}{r-1}(F(u))^{r-1}(1-F(u))^{n-r} \\ &= \frac{n!}{(r-1)!(n-r)!}(F(u))^{r-1}(1-F(u))^{n-r}f(u), \end{aligned}$$

$-\infty < u < \infty$. This is in fact the correct formula for the density of $X_{(r)}$.

Next, consider the case of the joint density of two order statistics, $X_{(r)}$ and $X_{(s)}$. Fix $0 < u < v < 1$. Then, to have $X_{(r)} = u$, and $X_{(s)} = v$, we must have exactly one observation at u, another $r-1$ below u, one at v, another $n-s$ above v, and $s-r-1$ between u and v. This will *suggest* that the joint density of $X_{(r)}$ and $X_{(s)}$ is

$$\begin{aligned} f_{r,s}(u,v) = nf(u)\binom{n-1}{r-1}(F(u))^{r-1}(n-r)f(v)\binom{n-r-1}{n-s} \\ (1-F(v))^{n-s}(F(v)-F(u))^{s-r-1} \end{aligned}$$

$$= \frac{n!}{(r-1)!(n-s)!(s-r-1)!}(F(u))^{r-1}$$
$$(1-F(v))^{n-s}(F(v)-F(u))^{s-r-1}f(u)f(v),$$
$$-\infty < u < v < \infty.$$

Again, this is indeed the joint density of two specific order statistics, $X_{(r)}$ and $X_{(s)}$.

The heuristic argument is made rigorous by using the binomial and multinomial distributions. It is quite clear where the binomial and multinomial distributions come from in a rigorous proof; for example, to find the probability that $X_{(r)} \le x$, we can compute the binomial probability that at least r of the n observations among $X_1, X_2, \ldots, X_n$ are less than or equal to x. We will omit the formal details.

Theorem 12.8 (Density of One and Two Order Statistics and the Sample Range). *Let $X_1, X_2, \ldots, X_n$ be independent observations from a continuous CDF $F(x)$ with density function $f(x)$. Then:*

(a) $X_{(n)}$ has the density $f_n(u) = nF^{n-1}(u)f(u), -\infty < u < \infty$.
(b) $X_{(1)}$ has the density $f_1(u) = n(1-F(u))^{n-1}f(u), -\infty < u < \infty$.
(c) For a general $r, 1 \le r \le n$, $X_{(r)}$ has the density

$$f_r(u) = \frac{n!}{(r-1)!(n-r)!}F^{r-1}(u)(1-F(u))^{n-r}f(u), -\infty < u < \infty.$$

(d) For general $1 \le r < s \le n$, $(X_{(r)}, X_{(s)})$ have the joint density

$$= \frac{n!}{(r-1)!(n-s)!(s-r-1)!}(F(u))^{r-1}(1-F(v))^{n-s}$$
$$(F(v)-F(u))^{s-r-1}f(u)f(v), -\infty < u < v < \infty.$$

(e) The minimum and the maximum, $X_{(1)}$ and$X_{(n)}$, have the joint density

$$f_{1,n}(u,v) = n(n-1)(F(v)-F(u))^{n-2}f(u)f(v), -\infty < u < v < \infty.$$

(f) **(CDF of Range).** $W = W_n = X_{(n)} - X_{(1)}$ has the CDF

$$F_W(w) = n\int_{-\infty}^{\infty}[F(x+w)-F(x)]^{n-1}f(x)dx, w > 0.$$

(g) **(Density of Range).** $W = W_n = X_{(n)} - X_{(1)}$ has the density

$$f_W(w) = n(n-1)\int_{-\infty}^{\infty}[F(x+w)-F(x)]^{n-2}f(x)f(x+w)dx, w > 0.$$

Example 12.31 (Moments of Uniform Order Statistics). The general formulas in the theorem above lead to the following moment formulas in the uniform case.

In the $U[0,1]$ case,

$$E(U_{(1)}) = \frac{1}{n+1}, E(U_{(n)}) = \frac{n}{n+1},$$
$$\text{Var}(U_{(1)}) = \text{Var}(U_{(n)}) = \frac{n}{(n+1)^2(n+2)}; 1 - U_{(n)} \stackrel{\mathcal{L}}{=} U_{(1)};$$
$$\text{Cov}(U_{(1)}, U_{(n)}) = \frac{1}{(n+1)^2(n+2)}, E(W_n) = \frac{n-1}{n+1}, \text{Var}(W_n) = \frac{2(n-1)}{(n+1)^2(n+2)}.$$

For a general order statistic, from the fact that $U_{(r)} \sim Be(r, n-r+1)$, we get

$$E(U_{(r)}) = \frac{r}{n+1}; \text{Var}(U_{(r)}) = \frac{r(n-r+1)}{(n+1)^2(n+2)}.$$

Furthermore, it follows from the formula for the joint density of two order statistics that

$$\text{Cov}(U_{(r)}, U_{(s)}) = \frac{r(n-s+1)}{(n+1)^2(n+2)}.$$

What are the important lessons to learn from these formulas? The variance of the two extreme order statistics, namely $U_{(1)}$ and $U_{(n)}$, converge to zero at the rate $\frac{1}{n^2}$, but the variance of the median converges to zero at the rate $\frac{1}{n}$. The covariance between any two order statistics is strictly positive. The expected value of any specific order statistic, say $U_{(r)}$, is just a little smaller than $\frac{r}{n}$, as $\frac{r}{n+1}$ is a little smaller than $\frac{r}{n}$.

Example 12.32 (Exponential Order Statistics). A second distribution of special importance in the theory of order statistics is the exponential distribution. The mean λ essentially arises as just a multiplier in the calculations. So, we will treat only the standard exponential case.

Let $X_1, X_2, \ldots, X_n$ be independent standard exponential variables. Then, by the general theorem on the joint density of the order statistics, in this case the joint density of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is

$$f_{1,2,\ldots,n}(u_1, u_2, \ldots, u_n) = n! e^{-\sum_{i=1}^n u_i},$$

$0 < u_1 < u_2 < \cdots < u_n < \infty$. Also, in particular, the minimum $X_{(1)}$ has the density

$$f_1(u) = n(1 - F(u))^{n-1} f(u) = n e^{-(n-1)u} e^{-u} = n e^{-nu},$$

$0 < u < \infty$. In other words, we have the quite remarkable result that the *minimum of n independent standard exponentials is itself an exponential with mean $\frac{1}{n}$*. Also, from the general formula, the maximum $X_{(n)}$ has the density

$$f_n(u) = n(1 - e^{-u})^{n-1} e^{-u} = n \sum_{i=0}^{n-1} (-1)^i \binom{n-1}{i} e^{-(i+1)u}, 0 < u < \infty.$$

As a result,

$$E(X_{(n)}) = n\sum_{i=0}^{n-1}(-1)^i\binom{n-1}{i}\frac{1}{(i+1)^2} = \sum_{i=1}^{n}(-1)^{i-1}\binom{n}{i}\frac{1}{i},$$

which also can be shown to be equal to $1+\frac{1}{2}+\cdots+\frac{1}{n}$.

To summarize then, the minimum of iid exponentials is another exponential, the maximum is *not* an exponential, and the maximum has the remarkable expected value $1+\frac{1}{2}+\cdots+\frac{1}{n}$.

Example 12.33 (Normal Order Statistics). Another clearly important case is that of the order statistics of a normal distribution. Because the general $N(\mu,\sigma^2)$ random variable is a location-scale transformation of a standard normal variable, we have the distributional equivalence that $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ have the same joint distribution as $(\mu+\sigma Z_{(1)}, \mu+\sigma Z_{(2)}, \ldots, \mu+\sigma Z_{(n)})$. So, we consider just the standard normal case.

Because of the symmetry of the standard normal distribution around zero, for any r, $Z_{(r)}$ has the same distribution as $-Z_{(n-r+1)}$. In particular, $Z_{(1)}$ has the same distribution as $-Z_{(n)}$. From our general formula, the density of $Z_{(n)}$ is

$$f_n(x) = n\Phi^{n-1}(x)\phi(x), -\infty < x < \infty.$$

Again, this is a skewed density. It can be shown, either directly or by making use of general theorems on existence of moments of order statistics, that every moment, and in particular the mean and the variance of $Z_{(n)}$, exists. Except for very small n, closed-form formulas for the mean or variance are not possible. For small n, integration tricks do produce exact formulas. For example,

$$E(Z_{(n)}) = \frac{1}{\sqrt{\pi}} \text{ if } n=2; E(Z_{(n)}) = \frac{3}{2\sqrt{\pi}} \text{ if } n=3.$$

Such formulas are available for $n \le 5$; see David (1980).

We tabulate the expected value of the maximum for some values of n to illustrate the slow growth.

n	$E(Z_{(n)})$
2	.56
5	1.16
10	1.54
20	1.87
30	2.04
50	2.25
100	2.51
500	3.04
1000	3.24
10000	3.85

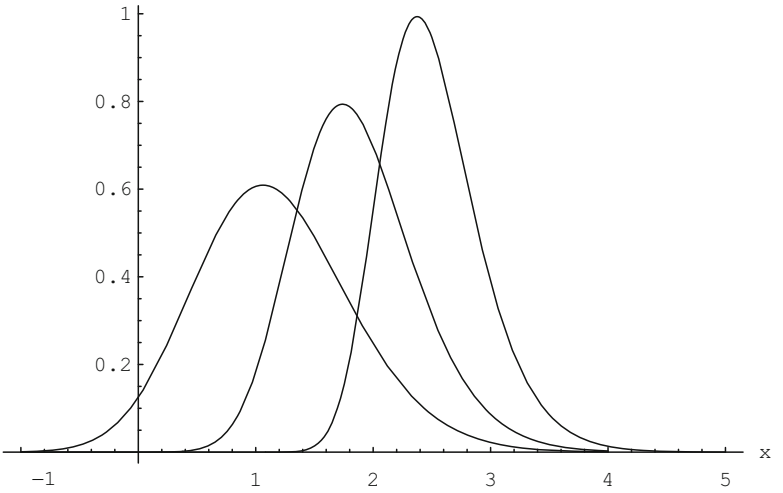


Fig. 12.8 Density of maximum of standard normals; $n = 5, 20, 100$

The density of $Z_{(n)}$ is plotted in Figure 12.8 for three values of n. We can see from the plot that the density is shifting to the right and at the same time getting more peaked.

12.7 Synopsis

(a) The joint density of n random variables $X_1, X_2, \ldots, X_n$ is a nonnegative function $f(x_1, x_2, \ldots, x_n)$ such that, for all $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n, -\infty < a_i \le b_i < \infty$,

$$P(a_1 \le X_1 \le b_1, a_2 \le X_2 \le b_2, \ldots, a_n \le X_n \le b_n)$$

$$= \int_{a_n}^{b_n} \ldots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n.$$

In particular, the joint density $f(x_1, x_2, \ldots, x_n)$ must satisfy

$$\int_{\mathcal{R}^n} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n = 1.$$

(b) The joint CDF of $X_1, X_2, \ldots, X_n$ is defined as

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} f(t_1, \ldots, t_n) dt_1 \ldots dt_n.$$

(c) The joint density can be obtained from the joint CDF by iterated partial differentiation:

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_1, x_2, \ldots, x_n).$$

(d) The marginal densities of any subset of $X_1, X_2, \ldots, X_n$ can be found by simply integrating out the other variables. For example, if $n = 2$, then the marginal density of X_1 is $f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$.

More generally, in the case of n variables, the marginal joint density of $X_1, X_2, \ldots, X_p$ is obtained from the joint density of $X_1, X_2, \ldots, X_n$ as

$$f_{1,2,\ldots,p}(x_1, x_2, \ldots, x_p) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_n) dx_{p+1} \cdots dx_n.$$

(e) $X_1, X_2, \ldots, X_n$ are independent if and only if the joint density factorizes at any $(x_1, x_2, \ldots, x_n)$ as

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i),$$

where $f_i(x_i)$ is the marginal density function of X_i.

(f) The expected value of a function $g(X_1, X_2, \ldots, X_n)$ is computed as

$$E[g(X_1, X_2, \ldots, X_n)] = \int_{\mathcal{R}^n} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n.$$

If $X_1, X_2, \ldots, X_n$ are independent, then $E(X_1 X_2 \ldots X_n) = E(X_1) E(X_2) \ldots E(X_n)$ and, more generally, $E[\prod_{i=1}^{n} g_i(X_i)] = \prod_{i=1}^{n} E[g_i(X_i)]$.

(g) Conditional densities and conditional expectations are defined analogously to the discrete case, except that integrals replace the sums.

The conditional density of X given $Y = y$ is defined as

$$f(x|y) = f(x|Y = y) = \frac{f(x, y)}{f_Y(y)}, \ \forall y \text{ such that } f_Y(y) > 0.$$

The conditional expectation of X given $Y = y$ is defined as

$$E(X|y) = E(X|Y = y) = \int_{-\infty}^{\infty} x f(x|y) dx = \frac{\int_{-\infty}^{\infty} x f(x, y) dx}{\int_{-\infty}^{\infty} f(x, y) dx},$$

$\forall y$ such that $f_Y(y) > 0$.

(h) It is important to understand that given two variables X and Y, $E(X|Y)$ will be a function of Y; there will not be any X left in the formula. If X and Y are independent, then $E(X|Y) = E(X)$, the marginal expectation of X.

(i) Bayes' theorem for conditional densities says that

$$f(y|x) = \frac{f(x|y)\, f_Y(y)}{f_X(x)}.$$

Thus, we can convert one conditional density to the other one by using Bayes' theorem.

(j) Conditional variance is also defined analogously to the discrete case, with integrals replacing the sums. The conditional variance of X given $Y = y$ is defined as

$$\text{Var}(X|y) = \text{Var}(X|Y = y) = \frac{\int_{-\infty}^{\infty}(x - \mu_X(y))^2 f(x, y)dx}{\int_{-\infty}^{\infty} f(x, y)dx},$$

where $\mu_X(y)$ denotes $E(X|y)$.

(k) Two special bivariate distributions are the bivariate uniform and the bivariate normal. The bivariate uniform corresponds to just a pair of iid $U[0, 1]$ variables. We can put a uniform distribution on any bounded set; e.g., we can put a uniform distribution in a circle or in a sphere. However, then the different variables are no longer independent.

(l) The density of the general bivariate normal distribution is

$$f(u, v) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]},$$

$-\infty < u, v < \infty$. If (U, V) have a general five-parameter bivariate normal distribution, then any linear function $aU + bV$ of (U, V) is normally distributed:

$$aU + bV \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sigma_1\sigma_2).$$

In particular, each of U, V is marginally normally distributed:

$$U \sim N(\mu_1, \sigma_1^2), V \sim N(\mu_2, \sigma_2^2).$$

If $\rho = 0$, then U and V are *independent* with $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ marginal distributions.

(m) Conditional distributions in a bivariate normal distribution are univariate normal. This is an important property. If (X, Y) have a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$, then

$$X|Y = y \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right);$$
$$Y|X = x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)\right).$$

In particular, the conditional expectations of X given $Y = y$ and of Y given $X = x$ are linear functions of y and x, respectively:

$$E(X|Y = y) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2);$$
$$E(Y|X = x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

(n) If $X_1, X_2, \ldots, X_n$ are independent random variables with a common density function $f(x)$, then the joint density function of the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is given by

$$f_{1,2,\ldots,n}(y_1, y_2, \ldots, y_n) = n!f(y_1)f(y_2)\ldots f(y_n)I_{\{y_1<y_2<\cdots<y_n\}}.$$

The density of one order statistic and the joint density of two order statistics are derived from here.

For a general $r, 1 \le r \le n, X_{(r)}$ has the density

$$f_r(u) = \frac{n!}{(r-1)!(n-r)!}F^{r-1}(u)(1 - F(u))^{n-r}f(u), -\infty < u < \infty.$$

For general $1 \le r < s \le n, (X_{(r)}, X_{(s)})$ have the joint density

$$= \frac{n!}{(r-1)!(n-s)!(s-r-1)!}(F(u))^{r-1}(1 - F(v))^{n-s}$$
$$(F(v) - F(u))^{s-r-1}f(u)f(v), -\infty < u < v < \infty.$$

(o) Any two order statistics are, in general, positively correlated.
(p) Closed-form formulas can be found for the means, variances, and covariances of uniform order statistics. They are given in the text.
The minimum of n iid exponential variables is again an exponential, but the maximum is not. Formulas for their expectations are given in the text.
Closed-form formulas for moments of normal order statistics cannot be found, unless n is very small. The maximum order statistic in the normal case grows at a very slow rate. Tables of the expected values are given in the text. The slow growth is due to the thin tails of the normal density.

12.8 Exercises

Exercise 12.1. Suppose (X, Y) have the joint density $f(x, y) = cxy, x, y \in [0, 1]$.

(a) Find the normalizing constant c.
(b) Are X and Y independent?

(c) Find the marginal densities and expectations of X, Y.
(d) Find the expectation of XY.

Exercise 12.2. Suppose (X, Y) have the joint density $f(x, y) = cxy, x, y \geq 0; x + y \leq 1$.

(a) Find the normalizing constant c.
(b) Are X and Y independent?
(c) Find the marginal densities and expectations of X, Y.
(d) Find the expectation of XY.

Exercise 12.3. Suppose (X, Y) have the joint density $f(x, y) = ce^{-y}, 0 \leq x \leq y < \infty$.

(a) Find the normalizing constant c.
(b) Are X and Y independent?
(c) Find the marginal densities and expectations of X, Y.
(d) Find the conditional expectation of X given $Y = y$.
(e) Find the conditional expectation of Y given $X = x$.
(f) Find the correlation between X and Y.

Exercise 12.4 (Uniform in a Triangle). Suppose X and Y are uniformly distributed in the triangle bounded by $-1 \leq x \leq 1, y \geq 0$, and the two lines $y = 1 + x$ and $y = 1 - x$.

(a) Find $P(X \geq -.5)$.
(b) Find $P(Y \geq .5)$.
(c) Find the marginal densities and expectations of X, Y.

Exercise 12.5 (Uniform Distribution in a Sphere). Suppose (X, Y, Z) have the density $f(x, y, z) = c$ if $x^2 + y^2 + z^2 \leq 1$.

(a) Find the constant c.
(b) Are any of X, Y or Y, Z or X, Z pairwise independent?
(c) Find the marginal densities and expectations of X, Y, Z.
(d) Find the conditional expectation of X given $Y = y$ and find the conditional expectation of X given $Z = z$.
(e) Find the conditional expectation of X given $Y = y$ and $Z = z$.
(f) Find the correlation between any pair, say X and Y.

Exercise 12.6. Suppose X and Y are independent $U[0, 1]$ variables. Find the conditional expectation $E(|X - Y||Y = y)$.

Exercise 12.7 (Uniform in a Triangle). Suppose X and Y are uniformly distributed in the triangle $x, y \geq 0, x + y \leq 1$. Find the conditional expectation $E(|X - Y||Y = y)$.

Exercise 12.8. Suppose X, Y, Z are independent $U[0,1]$ variables. Find $P(|X - Y| > |Y - Z|)$.

Exercise 12.9. * **(Iterated Expectation).** Suppose X and Y are independent standard exponential variables. Find $E(X\sqrt{X+Y})$.

Exercise 12.10 (Expectation of a Quotient). Suppose X and Y are independent and $X \sim Be(2,2), Y \sim Be(3,3)$. Find $E(\frac{X^2}{Y^2})$.

Exercise 12.11. Suppose X, Y, Z are three independent standard exponential variables. Find $P(X < 2Y < 3Z)$.

Exercise 12.12. * **(Conceptual).** Suppose $X \sim U[0,1]$ and $Y = 2X$. What is the joint distribution of (X, Y)? Does the joint distribution have a density?

Exercise 12.13. * **(Breaking a Stick).** Suppose $X \sim U[0,1]$, and given that $X = x$, $Y \sim U[0,x]$. Let $U = 1 - X, V = Y, W = X - Y$. Find the expectation of the maximum of U, V, W. This amounts to breaking a stick and then breaking the left piece again.

Exercise 12.14 (Iterated Expectation). Suppose $X_1 \sim U[0,1]$ and, for $n \geq 2$, X_n given that $X_{n-1} = x$ is distributed as $U[0,x]$. What is $E(X_n)$ and its limit as $n \to \infty$?

Exercise 12.15 (Bivariate Normal Probability). Suppose X and Y are jointly bivariate normal with zero means, unit standard deviations, and correlation ρ. Find all values of ρ for which $\frac{1}{4} \leq P(X > 0, Y > 0) \leq \frac{5}{12}$.

Exercise 12.16. Suppose X and Y are jointly bivariate normal with zero means, unit standard deviations, and correlation $\rho = .75$. Find $P(Y > 2|X = 1)$.

Exercise 12.17. Suppose X and Y are jointly bivariate normal with general parameters. Characterize all constants a, b such that $X + Y$ and $aX + bY$ are independent.

Exercise 12.18. * **(Probability of a Diamond).** Suppose X, Y, Z are independent $U[-1,1]$ variables. Find the probability that $|X| + |Y| + |Z| \leq 1$.

Exercise 12.19 (Missing the Bus). A bus arrives at a stop at a random time between 9:00 AM and 9:15 AM. Tim will arrive at that stop at a random time between 9:00 AM and 9:15 AM, independently of the bus, and will wait for (at most) five minutes at the stop. Find the probability that Tim will meet the bus.

Exercise 12.20. Cathy and Jen plan to meet at a cafe, and each will arrive at the cafe at a random time between 11:00 AM and 11:30 AM, independently of each other. Find the probability that the first to arrive has to wait between five and ten minutes for the other to arrive.

Exercise 12.21 (Bivariate Normal Probability). Suppose the amounts of oil (in barrels) lifted on a given day from two wells are jointly bivariate normal, with means 150 and 200, variances 100 and 25, and correlation .5. What is the probability that the total amount lifted is larger than 400 barrels on any given day? What is he probability that the amounts lifted from the two wells on any day differ by more than 50 barrels?

Exercise 12.22. * **(Conceptual).** Suppose (X, Y) have a bivariate normal distribution with zero means, unit standard deviations, and correlation $\rho, -1 < \rho < 1$. What is the joint distribution of $(X + Y, X - Y, Y)$? Does this joint distribution have a density?

Exercise 12.23. Suppose $X \sim N(\mu, \sigma^2)$. Find the correlation between X and Y, where $Y = X^2$. Find all values of (μ, σ) for which the correlation is zero.

Exercise 12.24. * **(Maximum Correlation).** Suppose (X, Y) has a general bivariate normal distribution with a positive correlation ρ. Show that among all functions $g(X), h(Y)$ with finite variances, the correlation between $g(X)$ and $h(Y)$ is maximized when $g(X) = X, h(Y) = Y$.

Exercise 12.25 (Bivariate Normal Calculation). Suppose $X \sim N(0, 1)$ and, given $X = x$, $Y \sim N(x + 1, 1)$.

(a) What is the marginal distribution of Y?
(b) What is the correlation between X and Y?
(c) What is the conditional distribution of X given $Y = y$?

Exercise 12.26. * **(Uniform Distribution in a Sphere).** Suppose X, Y, Z are uniformly distributed in the unit sphere. Find the mean and the variance of the distance of the point (X, Y, Z) from the origin.

Exercise 12.27. * **(Uniform Distribution in a Sphere).** Suppose X, Y, Z are uniformly distributed in the unit sphere.

(a) Find the marginal density of (X, Y).
(b) Find the marginal density of X.

Exercise 12.28. Suppose X, Y, Z are independent exponentials with means $\lambda, 2\lambda, 3\lambda$. Find $P(X < Y < Z)$.

Exercise 12.29. * **(Mean Residual Life).** Suppose $X \sim N(\mu, \sigma^2)$. Derive a formula for the mean residual life and investigate its monotonicity behavior with respect to each of σ, c, μ, each time holding the other two fixed.

Exercise 12.30. * **(Bivariate Normal Conditional Calculation).** Suppose the systolic blood pressure X and fasting blood sugar Y are jointly distributed as bivariate normal in some population with means 120 and 105, standard deviations 10 and 20, and correlation 0.7. Find the average fasting blood sugar of those with a systolic blood pressure greater than 140.

Exercise 12.31 (Uniform Order Statistics). Suppose X, Y, Z are three independent $U[0, 1]$ variables. Let U, V, W denote the minimum, median, and the maximum of X, Y, Z.

(a) Find the densities of U, V, W.
(b) Find the densities of $\frac{U}{V}$ and $\frac{V}{W}$ and their joint density.
(c) Find $E(\frac{U}{V})$ and $E(\frac{V}{W})$.

Exercise 12.32 (Uniform Order Statistics). Suppose $X_1, \ldots, X_5$ are independent $U[0, 1]$ variables. Find the joint density of $X_{(2)}, X_{(3)}, X_{(4)}$, and $E(X_{(4)} + X_{(2)} - 2X_{(3)})$.

Exercise 12.33 (Exponential Order Statistics). Suppose X, Y, Z are three independent standard exponential variables, and let U, V, W be their minimum, median, and maximum. Find the densities of $U, V, W, W - U$.

Exercise 12.34. * **(Waiting Time).** Peter, Paul, and Mary went to a bank to do some business. Two counters were open, and Peter and Paul went first. Peter, Paul, and Mary will each take, independently, an $Exp(\lambda)$ amount of time to finish their business from the moment they arrive at the counter.

(a) What is the density of the epoch of the last departure?
(b) What is the probability that Mary will be the last to finish?
(c) What is the density of the total time taken by Mary from arrival to finishing her business?

Exercise 12.35 (Density of the Median). Let $X_1, \ldots, X_n$ be independent observations from a continuous CDF F with a density symmetric about some μ. Show that, for all odd sample sizes $n = 2m + 1$, the median $X_{(m+1)}$ has a density symmetric about μ.

Exercise 12.36. * Suppose $X_1, \ldots, X_n$ are independent $U[0, 1]$ variables.

(a) Find the probability that all n observations fall within some interval of length at most .9.
(b) Find the smallest n such that $P(X_{(n)} \geq .99, X_{(1)} \leq .01) \geq .99$.

Exercise 12.37 (Use Your Computer). Simulate 500 independent standard normal values. Use these to generate 250 pairs of bivariate normal vectors (X_i, Y_i) with zero means, unit standard deviations, and correlation 0.5. Find the average value of $(X_i - Y_i)^2$ over the simulations. Does it equal the theoretical average approximately?

Exercise 12.38 (Use Your Computer). Simulate 500 independent standard normal values. Use these to generate 250 pairs of bivariate normal vectors (X_i, Y_i), with zero means, unit standard deviations, and correlation $\rho = .25, .5, .75$. Use your simulated pairs to estimate $P(X > 0, Y > 0)$. Does your estimate equal the theoretical value approximately?

Exercise 12.39. * **(Buffon's Needle).** Suppose a plane is gridded by a series of parallel lines drawn h units apart. A needle of length l is dropped at random on the plane. Let $p(l,h)$ be the probability that the needle intersects one of the parallel lines. Show that:

(a) $p(l,h) = \frac{2l}{h\pi}$, if $l \leq h$.
(b) $p(l,h) = \frac{2l}{h\pi} - \frac{2}{h\pi}[\sqrt{l^2 - h^2} + h \arcsin(\frac{h}{l})] + 1$ if $l > h$.

References

David, H.A. (1980). *Order Statistics*, Wiley, New York.
Tong, Y.L. (1990). *The Multivariate Normal Distribution*, Springer-Verlag, New York.

Chapter 13
Convolutions and Transformations

Very naturally, in applications we often want to study suitable functions or transformations of an original collection of variables $X_1, X_2, \ldots, X_n$. For example, the original variables $X_1, X_2, \ldots, X_n$ could be the inputs into some process or system, and we may be interested in the output, which is some suitable function of these input variables. We dealt with the problem of finding distributions of functions of one continuous variable in Chapter 7. Similar, but technically more involved, techniques for studying distributions of functions of many continuous variables are presented with illustrations in this chapter. Sums, products, and quotients are special functions that arise quite naturally in applications. These will be discussed with special emphasis, although the general theory is also presented. Specifically, we present in this chapter the classic theory of polar transformations, the Helmert transformation in arbitrary dimensions, and the development of the t, and F distributions.

13.1 Convolutions and Examples

Definition 13.1. Let X and Y be independent random variables. The distribution of their sum $X + Y$ is called the *convolution* of the distributions of X and Y.

Remark. We also sometimes refer to the convolution of the distributions of X and Y as simply the convolution of X and Y. Usually, X and Y will both be discrete or both continuous in applications, but they do not have to be; for example, X could be normal and Y Poisson.

Example 13.1. Suppose X and Y have a joint density function $f(x, y)$, and suppose we want to find the density of their sum, namely $X + Y$. Denote the conditional density of X given $Y = y$ by $f_{X|y}(x|y)$ and the conditional CDF, namely $P(X \le u|Y = y)$, by $F_{X|Y}(u)$. Then, by the iterated expectation formula,

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_13,

$$\begin{aligned} P(X+Y \le z) &= E[I_{X+Y\le z}] \\ &= E_Y[E(I_{X+Y\le z}|Y=y)] = E_Y[E(I_{X+y\le z}|Y=y)] \\ &= E_Y[P(X \le z-y|Y=y)] \\ &= E_Y[F_{X|Y}(z-y)] = \int_{-\infty}^{\infty} F_{X|Y}(z-y)\, f_Y(y)dy. \end{aligned}$$

In particular, if X and Y are independent, then the conditional CDF $F_{X|Y}(u)$ will be the same as the marginal CDF $F_X(u)$ of X. In this case, the expression above simplifies to

$$P(X+Y \le z) = \int_{-\infty}^{\infty} F_X(z-y)\, f_Y(y)dy.$$

The density of $X+Y$ can be obtained by differentiating the CDF of $X+Y$:

$$\begin{aligned} f_{X+Y}(z) &= \frac{d}{dz}P(X+Y \le z) = \frac{d}{dz}\int_{-\infty}^{\infty} F_X(z-y)\, f_Y(y)dy \\ &= \int_{-\infty}^{\infty} \left[\frac{d}{dz}F_X(z-y)\, f_Y(y)\right] dy \end{aligned}$$

(assuming that the derivative can be carried inside the integral)

$$= \int_{-\infty}^{\infty} f_X(z-y)\, f_Y(y)dy.$$

Indeed, this is the general formula for the density of the sum of two real-valued independent continuous random variables.

Theorem 13.1. *Let X and Y be independent real valued random variables with densities $f_X(x)$ and $f_Y(y)$, respectively. Let $Z = X+Y$ be the sum of X and Y. Then, the density of the convolution is*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)\, f_Y(y)dy.$$

More generally, if X and Y are not necessarily independent and have joint density $f(x,y)$, then $Z = X+Y$ has the density

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X|Y}(z-y)\, f_Y(y)dy.$$

Definition 13.2. If X and Y are independent continuous random variables with a common density $f(x)$, then the density of the convolution is denoted as $f * f$. In general, if $X_1, X_2, \ldots, X_n$ are n independent continuous random variables with a common density $f(x)$, then the density of their sum $X_1 + X_2 + \cdots + X_n$ is called the *n-fold convolution of f* and is denoted as $f^{*^{(n)}}$.

Example 13.2 (Sum of Exponentials). Suppose X and Y are independent $Exp(\lambda)$ variables and we want to find the density of $Z = X + Y$. By the convolution formula, for $z > 0$,

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\lambda} e^{-\frac{z-y}{\lambda}} I_{y<z} \frac{1}{\lambda} e^{-\frac{y}{\lambda}} I_{y>0} dy$$
$$= \frac{1}{\lambda^2} \int_0^z e^{-\frac{z}{\lambda}} dy = \frac{z e^{-\frac{z}{\lambda}}}{\lambda^2},$$

which is the density of a Gamma distribution with parameters 2 and λ. Recall that we had proved this earlier in Chapter 8 by using mgfs.

Example 13.3 (Difference of Exponentials). Let U and V be independent standard exponentials. We want to find the density of $Z = U - V$. Writing $X = U$ and $Y = -V$, we notice that $Z = X + Y$ and X and Y are still independent. However, now Y is a *negative exponential* and so has density $f_Y(y) = e^y I\, y < 0$. It is also important to note that Z can now take any real value, positive or negative. Substituting into the formula for the convolution density,

$$f_Z(z) = \int_{-\infty}^{\infty} e^{-(z-y)} (I\, y < z) e^y (I\, y < 0) dy.$$

Now, first consider $z > 0$. Then this last expression becomes

$$f_Z(z) = \int_{-\infty}^{0} e^{-(z-y)} e^y dy = e^{-z} \int_{-\infty}^{0} e^{2y} dy = \frac{1}{2} e^{-z}.$$

On the other hand, for $z < 0$, the convolution formula becomes

$$f_Z(z) = \int_{-\infty}^{z} e^{-(z-y)} e^y dy = e^{-z} \int_{-\infty}^{z} e^{2y} dy$$
$$= e^{-z} \frac{1}{2} e^{2z} = \frac{1}{2} e^z.$$

Combining the two cases, we can write the single formula

$$f_Z(z) = \frac{1}{2} e^{-|z|}, -\infty < z < \infty,$$

i.e., if X and Y are independent standard exponentials, then the difference $X - Y$ has a standard double exponential density. This representation of the double exponential distribution is often useful. Also note that although the standard exponential distribution is obviously not symmetric, the distribution of the difference of two independent exponentials is symmetric. This is a useful technique for symmetrizing a random variable.

Definition 13.3 (Symmetrization of a Random Variable). Let X_1 and X_2 be independent random variables with a common distribution F. Then $X_s = X_1 - X_2$ is called *the symmetrization of F or symmetrization of X_1*.

If X_1 is a continuous random variable with density $f(x)$, then its symmetrization has the density

$$f_s(z) = \int_{-\infty}^{\infty} f(z+y)f(y)dy.$$

Example 13.4 (A Neat General Formula). Suppose X and Y are positive random variables with a joint density of the form $f(x,y) = g(x+y)$. What is the density of the convolution?

Note that now X and Y are in general not independent because a joint density of the form $g(x+y)$ does not in general factorize into the product form necessary for independence. First, the conditional density

$$f_{X|Y}(x) = \frac{g(x+y)}{\int_0^{\infty} g(x+y)dx} = \frac{g(x+y)}{\int_y^{\infty} g(x)dx} = \frac{g(x+y)}{\bar{G}(y)},$$

writing $\bar{G}(y)$ for $\int_y^{\infty} g(x)dx$. Also, the marginal density of Y is

$$f_Y(y) = \int_0^{\infty} g(x+y)dx = \int_y^{\infty} g(x)dx = \bar{G}(y).$$

Substituting into the general case formula for the density of a sum,

$$f_Z(z) = \int_0^z \frac{g(z)}{\bar{G}(y)}\bar{G}(y)dy = zg(z),$$

a very neat formula.

As an application, consider the example of (X,Y) being uniformly distributed in a triangle with the joint density $f(x,y) = 2, x, y \geq 0, x+y \leq 1$. Identifying the function g as $g(z) = 2I_{0\leq z\leq 1}$, we have, from our general formula above, that in this case $Z = X+Y$ has the density $f_Z(z) = 2z, 0 \leq z \leq 1$.

Example 13.5 (Sums of Cauchy Variables). Let X and Y be independent standard Cauchy random variables with the common density function $f(x) = \frac{1}{\pi(1+x^2)}$, $-\infty < x < \infty$. Then, the density of the convolution is

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy = \int_{-\infty}^{\infty} f(z-y)f(y)dy \\ &= \frac{1}{\pi^2}\int_{-\infty}^{\infty} \frac{1}{(1+(z-y)^2)(1+y^2)}dy \\ &= \frac{1}{\pi^2}\frac{2\pi}{4+z^2} \end{aligned}$$

$$\left(\text{on a partial fraction expansion of } \frac{1}{(1+(z-y)^2)(1+y^2)}\right)$$

$$= \frac{2}{\pi(4+z^2)}.$$

Therefore, the density of $W = \frac{Z}{2} = \frac{X+Y}{2}$ would be $\frac{1}{\pi(1+w^2)}$, which is, remarkably, the same standard Cauchy density that we had started with.

By using characteristic functions, which we have not discussed in this text, it can be shown that if $X_1, X_2, \ldots, X_n$ are independent standard Cauchy variables, then, for any $n \geq 2$, their average $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ also has the standard Cauchy distribution.

Example 13.6 (Normal-Poisson Convolution). Here is an example of the convolution of one continuous and one discrete random variable. Let $X \sim N(0,1)$ and $Y \sim Poi(\lambda)$. Then their sum $Z = X + Y$ is still continuous and has the density

$$f_Z(z) = \sum_{y=0}^{\infty} \phi(z-y)\frac{e^{-\lambda}\lambda^y}{y!}.$$

More generally, if $X \sim N(0,\sigma^2)$ and $Y \sim Poi(\lambda)$, then the density of the sum is

$$f_Z(z) = \frac{1}{\sigma}\sum_{y=0}^{\infty} \phi\left(\frac{z-y}{\sigma}\right)\frac{e^{-\lambda}\lambda^y}{y!}.$$

This is not expressible in terms of the elementary functions. However, it is interesting to plot the density. The plot in Figure 13.1 shows an unconventional and strange density function for the convolution with multiple local maxima and shoulders.

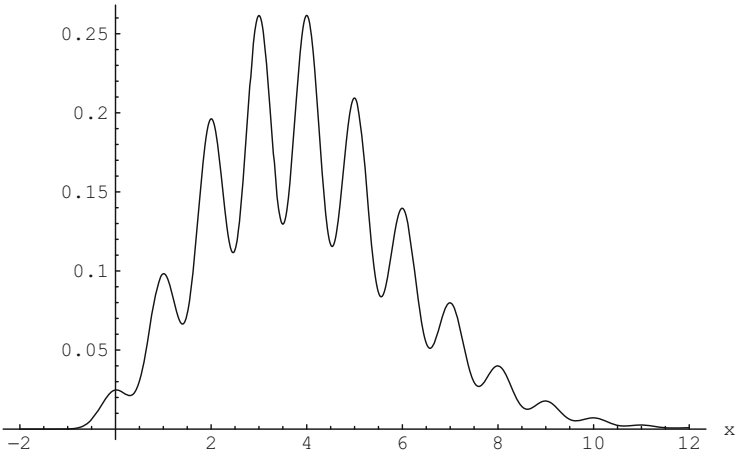


Fig. 13.1 Convolution of $N(0, .09)$ and *Poi* (4)

For purposes of summary and easy reference, we list some convolutions of common types below.

Distribution of Summands	Distribution of Sum
$X_i \sim Bin(n_i, p)$	$Bin(\sum n_i, p)$
$X_i \sim Poi(\lambda_i)$	$Poi(\sum \lambda_i)$
$X_i \sim NB(r_i, p)$	$NB(\sum r_i, p)$
$X_i \sim Exp(\lambda)$	$Gamma(n, \lambda)$
$X_i \sim N(\mu_i, \sigma_i^2)$	$N(\sum \mu_i, \sum \sigma_i^2)$
$X_i \sim C(\mu_i, \sigma_i^2)$	$C(\sum \mu_i, (\sum \sigma_i)^2)$
$X_i \sim U[a, b]$	See Chapter 10

13.2 Products and Quotients and the t and F Distributions

Suppose X and Y are two random variables. Then two other functions that arise naturally in many applications are the product XY and the quotient $\frac{X}{Y}$. Following exactly the same technique as for convolutions, one can find the density of each of XY and $\frac{X}{Y}$. More precisely, one first finds the CDF by using the iterated expectation technique, exactly as we did for convolutions, and then differentiates the CDF to obtain the density. The density formulas are given below; they are extremely important and useful. They are proved in exactly the same way that the formula for the density of the convolution was obtained above; you would condition and then take an iterated expectation. Therefore, the formal details are omitted.

Theorem 13.2. *Let X and Y be continuous random variables with a joint density $f(x, y)$. Let $U = XY, V = \frac{X}{Y}$. Then the densities of U, V are given by*

$$f_U(u) = \int_{-\infty}^{\infty} \frac{1}{|x|} f\left(x, \frac{u}{x}\right) dx,$$

$$f_V(v) = \int_{-\infty}^{\infty} |y| f(vy, y) dy.$$

Example 13.7 (Product and Quotient of Uniforms). Suppose X and Y are independent $U[0, 1]$ random variables. Then, by the theorem above, the density of the product $U = XY$ is

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} \frac{1}{|x|} f\left(x, \frac{u}{x}\right) dx \\ &= \int_u^1 \frac{1}{x} \times 1 dx = -\log u, 0 < u < 1. \end{aligned}$$

Next, again by the theorem above, the density of the quotient $V = \frac{X}{Y}$ is

$$f_V(v) = \int_{-\infty}^{\infty} |y| f(vy, y) dy = \int_0^{\min\{\frac{1}{v}, 1\}} y dy$$
$$= \frac{(\min\{\frac{1}{v}, 1\})^2}{2}, 0 < v < \infty;$$

thus, the density of the quotient V is

$$f_V(v) = \frac{1}{2} \text{ if } 0 < v \leq 1;$$
$$= \frac{1}{2v^2} \text{ if } v > 1.$$

The density of the quotient is plotted in Figure 13.2; we see that it is continuous, but not differentiable at $v = 1$.

Example 13.8 (Ratio of Standard Normals). The distribution of the ratio of two independent standard normal variables is an interesting one; we now show that it is in fact a standard Cauchy distribution. Indeed, by applying the general formula, the density of the quotient $V = \frac{X}{Y}$ is

$$f_V(v) = \int_{-\infty}^{\infty} |y| f(vy, y) dy$$
$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} |y| e^{-\frac{y^2}{2}(1+v^2)} dy$$

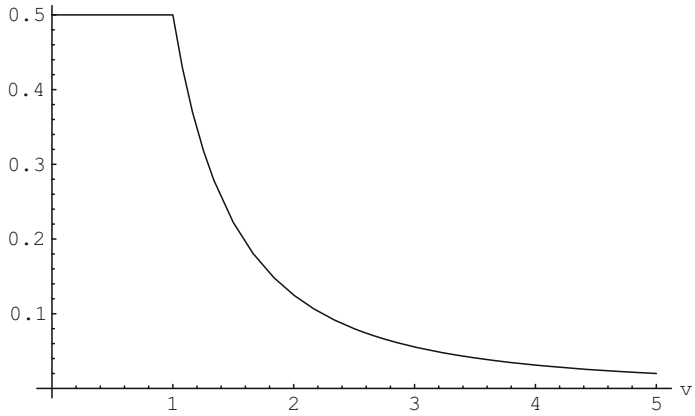


Fig. 13.2 Density of quotient of two uniforms

$$= \frac{1}{\pi}\int_0^\infty y e^{-\frac{y^2}{2}(1+v^2)}dy$$
$$= \frac{1}{\pi(1+v^2)}, -\infty < v < \infty,$$

by making the substitution $t = \sqrt{1+v^2}y$ in the integral on the last line. This proves that the quotient has a standard Cauchy distribution.

It is important to note that zero means for the normal variables are essential for this result. If either X or Y has a nonzero mean, the quotient has a complicated distribution and is definitely not Cauchy. The distribution is worked out in Hinkley (1969). It is also highly interesting that there are many other distributions F such that if X and Y are independent with the common distribution F, then the quotient $\frac{X}{Y}$ is distributed as a standard Cauchy. One example of such a distribution F is a continuous distribution with the density $f(x) = \frac{\sqrt{2}}{\pi}\frac{1}{1+x^4}, -\infty < x < \infty$.

Example 13.9 (The F Distribution). Let $X \sim G(\alpha, 1), Y \sim G(\beta, 1)$, and suppose X and Y are independent. The distribution of the ratio $R = \frac{X/\alpha}{Y/\beta}$ arises in statistics in many contexts and is called *an F distribution*. We derive the explicit form of the density here.

First, we will find the density of $\frac{X}{Y}$, from which the density of $R = \frac{X/\alpha}{Y/\beta}$ will follow easily. Again, by applying the general formula for the density of a quotient, the density of the quotient $V = \frac{X}{Y}$ is

$$f_V(v) = \int_{-\infty}^\infty |y| f(vy, y)dy = \frac{1}{\Gamma(\alpha)\Gamma(\beta)}\int_0^\infty y e^{-y(1+v)}(vy)^{\alpha-1}y^{\beta-1}dy$$
$$= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}v^{\alpha-1}\int_0^\infty e^{-y(1+v)}y^{\alpha+\beta-1}dy$$
$$= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}v^{\alpha-1}\frac{\Gamma(\alpha+\beta)}{(1+v)^{\alpha+\beta}}$$
$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{v^{\alpha-1}}{(1+v)^{\alpha+\beta}}, 0 < v < \infty.$$

To complete the example, notice now that $R = cV$, where $c = \frac{\beta}{\alpha}$. Therefore, the density of R is immediately obtained from the density of V. Indeed,

$$f_R(r) = \frac{1}{c}f_V\left(\frac{r}{c}\right),$$

where f_V is the function we just derived above. If we simplify $f_R(r)$, we get the final expression

$$f_R(r) = \frac{\left(\frac{\beta}{\alpha}\right)^\beta r^{\alpha-1}}{B(\alpha,\beta)\left(r+\frac{\beta}{\alpha}\right)^{\alpha+\beta}}, r > 0.$$

This is the F- density with parameters α, β; it is common in statistics to refer to 2α and 2β as the degrees of freedom of the distribution.

Example 13.10 (The Student t Distribution). Once again, the t distribution is one that arises frequently in statistics. Suppose $X \sim N(0,1), Z \sim \chi^2_m$, and that X and Z are independent. Let $Y = \sqrt{\frac{Z}{m}}$. Then the distribution of the quotient $V = \frac{X}{Y}$ is called *the t distribution with m degrees of freedom*. We derive its density in this example.

Recall that Z has the density $\frac{e^{-z/2}z^{m/2-1}}{2^{m/2}\Gamma(\frac{m}{2})}, z > 0$. Therefore, $Y = \sqrt{\frac{Z}{m}}$ has the density

$$f_Y(y) = \frac{m^{m/2}e^{-my^2/2}y^{m-1}}{2^{m/2-1}\Gamma\left(\frac{m}{2}\right)}, y > 0.$$

Since, by hypothesis, X and Z are independent, it follows that X and Y are also independent, so their joint density $f(x, y)$ is just the product of the marginal densities of X and Y.

Once again, by applying our general formula for the density of a quotient,

$$\begin{aligned}
f_V(v) &= \int_{-\infty}^{\infty} |y| f(vy, y) dy \\
&= \frac{m^{m/2}}{\sqrt{2\pi}2^{m/2-1}\Gamma\left(\frac{m}{2}\right)} \int_0^{\infty} y e^{-v^2y^2/2} e^{-my^2/2} y^{m-1} dy \\
&= \frac{m^{m/2}}{\sqrt{2\pi}2^{m/2-1}\Gamma\left(\frac{m}{2}\right)} \int_0^{\infty} e^{-(v^2+m)y^2/2} y^m dy \\
&= \frac{m^{m/2}}{\sqrt{2\pi}2^{m/2-1}\Gamma\left(\frac{m}{2}\right)} \times \frac{\Gamma\left(\frac{m+1}{2}\right)2^{(m-1)/2}}{(m+v^2)^{(m+1)/2}} \\
&= \frac{m^{m/2}\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{m}{2}\right)} \frac{1}{(m+v^2)^{(m+1)/2}} \\
&= \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\Gamma\left(\frac{m}{2}\right)\left(1+\frac{v^2}{m}\right)^{(m+1)/2}}, -\infty < v < \infty.
\end{aligned}$$

This is the density of *the Student t distribution with m degrees of freedom*.

Note that when the degree of freedom $m = 1$, this becomes just the standard Cauchy density. The t distribution was first derived in 1908 by William Gossett writing under the pseudonym Student. The distribution was later named the *Student t distribution* by Ronald Fisher.

The t density, just like the standard normal, is symmetric and unimodal around zero, although with tails much heavier than that of the standard normal for small

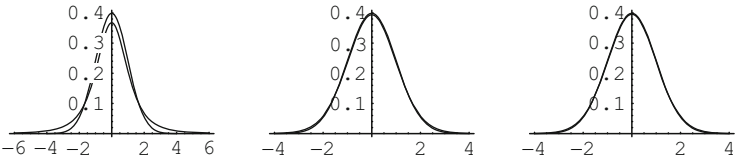


Fig. 13.3 t density for $m = 3, 20, 30$ degrees of freedom with $N(0, 1)$ density superimposed

values of m. However, as $m \to \infty$, the density converges pointwise to the standard normal density, and then the t and the standard normal densities look almost the same. We give in Figure 13.3 a plot of the t density for a few degrees of freedom and that of the standard normal density for a visual comparison.

Example 13.11 (An Interesting Gaussian Factorization). We will exhibit independent random variables X and Y in this example such that XY has a standard normal distribution. Note that if we allow Y to be a constant random variable, then we can always write such a factorization. After all, we can take X to be standard normal and Y to be 1! So, we will exhibit *nonconstant* X, Y such that they are independent and XY has a standard normal distribution.

For this, let X have the density $xe^{-x^2/2}, x > 0$, and let Y have the density $\frac{1}{\pi\sqrt{1-y^2}}, -1 < y < 1$. Then, by our general formula for the density of a product, the product $U = XY$ has the density

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} \frac{1}{|x|} f\left(x, \frac{u}{x}\right) dx \\ &= \frac{1}{\pi} \int_{|u|}^{\infty} \frac{1}{x} x e^{-x^2/2} \frac{1}{\sqrt{1 - \frac{u^2}{x^2}}} dx \\ &= \frac{1}{\pi} \int_{|u|}^{\infty} \frac{x e^{-x^2/2}}{\sqrt{x^2 - u^2}} dx \\ &= \frac{1}{\pi} \sqrt{\frac{\pi}{2}} e^{-u^2/2} = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \end{aligned}$$

where the final integral is obtained by the substitution $x^2 = u^2 + z^2$.

13.3 Transformations

The simple technique that we used in the previous section to derive the density of a sum or a product does not extend to functions of a more complex nature. Consider the simple case of just two continuous variables X and Y with some joint density $f(x, y)$, and suppose we want to find the density of some function $U = g(X, Y)$.

Then, the general technique is to pair up U with *another* function $V = h(X, Y)$ and first obtain the joint CDF of (U, V) from the joint CDF of (X, Y). The pairing up has to be done carefully; i.e., only some judicious choices of V will work in a given example. Having found the joint CDF of (U, V), by differentiation one finds the joint density of (U, V) and then finally integrates out v to obtain the density of just U. Fortunately, this agenda does work out because the transformation from (X, Y) to (U, V) can be treated as just a change of variable in manipulation with double integrals, and calculus tells us how to find double integrals by making suitable changes of variables (i.e., substitutions). Indeed, the method works out for *any* number of jointly distributed variables, $X_1, X_2, \ldots, X_n$, and a function $U = g(X_1, X_2, \ldots, X_n)$, and the reason it works out is that the whole method is just a change of variables in manipulating a multivariate integral.

The following theorem is on density of a multivariate transformation and is a major theorem in multivariate distribution theory. It is really nothing but the change of variable theorem of multivariate calculus. After all, probabilities in the continuous case are integrals, and an integral can be evaluated by changing variables to a new set of coordinates. If we do that, then we have to put in the Jacobian term coming from making the change of variable. Translated into densities, the theorem is the following.

Theorem 13.3 (Multivariate Jacobian Formula). *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *have the joint density function* $f(x_1, x_2, \ldots, x_n)$*, such that there is an open set* $S \subseteq \mathcal{R}^n$ *with* $P(\mathbf{X} \in S) = 1$*. Suppose* $u_i = g_i(x_1, x_2, \ldots, x_n), 1 \leq i \leq n$ *are* n *real-valued functions of* $x_1, x_2, \ldots, x_n$ *such that:*

(a) $(x_1, x_2, \ldots, x_n) \to (g_1(x_1, x_2, \ldots, x_n), \ldots, g_n(x_1, x_2, \ldots, x_n))$ *is a one-to-one function of* $(x_1, x_2, \ldots, x_n)$ *on* S *with range space* T;

(b) *the inverse functions* $x_i = h_i(u_1, u_2, \ldots, u_n), 1 \leq i \leq n$, *are continuously differentiable on* T *with respect to each* u_j*; and*

(c) *the Jacobian determinant*

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial u_1} & \dfrac{\partial x_1}{\partial u_2} & \cdots & \dfrac{\partial x_1}{\partial u_n} \\ \dfrac{\partial x_2}{\partial u_1} & \dfrac{\partial x_2}{\partial u_2} & \cdots & \dfrac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial x_n}{\partial u_1} & \dfrac{\partial x_n}{\partial u_2} & \cdots & \dfrac{\partial x_n}{\partial u_n} \end{vmatrix}$$

is nonzero.

Then the joint density of $(U_1, U_2, \ldots, U_n)$ *is given by*

$$f_{U_1, U_2, \ldots, U_n}(u_1, u_2, \ldots, u_n) = f(h_1(u_1, u_2, \ldots, u_n), \ldots, h_n(u_1, u_2, \ldots, u_n)) \times |J|,$$

where $|J|$ *denotes the absolute value of the Jacobian determinant* J *and the notation* f *on the right-hand side means the original joint density of* $(X_1, X_2, \ldots, X_n)$.

13.4 Applications of the Jacobian Formula

We will now see a number of examples applying the Jacobian formula to finding the density of interesting transformations. We emphasize that quite often *only one of the functions $u_i = g_i(x_1, x_2, \ldots, x_n)$ will be provided for which we want to find the density function. But unless that function is a really simple one, its density cannot be found directly without invoking the Jacobian theorem given here. It is necessaary to make up the remaining $(n-1)$ functions and then obtain their joint density by using this Jacobian theorem. Finally, one would integrate out all these other coordinates to get the density function of just u_i. The other $(n-1)$ functions need to be found judiciously.*

Example 13.12 (A Relation Between Exponential and Uniform). Let X and Y be independent standard exponentials, and define $U = \frac{X}{X+Y}$. We want to find the density of U. We have to pair it up with another function V in order to use the Jacobian theorem. We choose $V = X + Y$. We have here a one-to-one function for $x > 0, y > 0$. Indeed, the inverse functions are

$$x = x(u, v) = uv; y = y(u, v) = v - uv = v(1 - u).$$

The partial derivatives of the inverse functions are

$$\frac{\partial x}{\partial u} = v; \frac{\partial x}{\partial v} = u; \frac{\partial y}{\partial u} = -v; \frac{\partial y}{\partial v} = 1 - u.$$

Thus, the Jacobian determinant equals $J = v(1-u) + uv = v$. By invoking the Jacobian theorem, the joint density of U, V is

$$f_{U,V}(u, v) = e^{-uv}e^{-v(1-u)}|v| = ve^{-v},$$

$0 < u < 1, v > 0$.

Thus, the joint density of U, V has factorized into a product form on a rectangle; the marginals are

$$f_U(u) = 1, 0 < u < 1; f_V(v) = ve^{-v}, v > 0,$$

and the rectangle is $(0, 1) \times (0, \infty)$. Therefore, we have proved that if X and Y are independent standard exponentials, then $\frac{X}{X+Y}$ and $X + Y$ are independent and are respectively uniform and Gamma. Of course, we already knew that $X+Y \sim G(2, 1)$ from our mgf proof in Chapter 8.

Example 13.13 (A Relation Between Gamma and Beta). The previous example generalizes in a nice way. Let X and Y be independent variables, distributed respectively as $G(\alpha, 1)$ and $G(\beta, 1)$. Again let $U = \frac{X}{X+Y}, V = X + Y$. Then, from our previous example, the Jacobian determinant is still $J = v$. Therefore, the joint density of U, V is

$$f_{U,V}(u,v) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} e^{-v}(uv)^{\alpha-1}(v(1-u))^{\beta-1}v$$

$$= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}e^{-v}v^{\alpha+\beta-1},$$

$0 < u < 1, v > 0$.

Once again, we have factorized the joint density of U and V as the product of the marginal densities, with (U, V) varying in the rectangle $(0, 1) \times (0, \infty)$, the marginal densities being

$$f_V(v) = \frac{e^{-v}v^{\alpha+\beta-1}}{\Gamma(\alpha+\beta)}, v > 0,$$

$$f_U(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}, 0 < u < 1.$$

That is, if X and Y are independent Gamma variables, then $\frac{X}{X+Y}$ and $X + Y$ are independent and are respectively distributed as Beta and Gamma. Of course, we already knew from an mgf argument that $X + Y$ is Gamma.

This relationship between the Gamma and the Beta distributions is useful in simulating values from a Beta distribution.

13.5 Polar Coordinates in Two Dimensions

Example 13.14 (Transformation to Polar Coordinates). We have already worked out a few examples where we transformed two variables to their polar coordinates in order to calculate expectations of suitable functions when the variables have a spherically symmetric density. We now use a transformation to polar coordinates to do distributional calculations. *In any spherically symmetric situation, transformation to polar coordinates is a technically useful device and gets the answers out quickly for many problems.*

Let (X, Y) have a spherically symmetric joint density given by $f(x, y) = g(\sqrt{x^2 + y^2})$. Consider the polar transformation $r = \sqrt{X^2 + Y^2}, \theta = \arctan(\frac{Y}{X})$. This is a one-to-one transformation, with the inverse functions given by

$$x = r\cos\theta, y = r\sin\theta.$$

The partial derivatives of the inverse functions are

$$\frac{\partial x}{\partial r} = \cos\theta, \frac{\partial x}{\partial \theta} = -r\sin\theta, \frac{\partial y}{\partial r} = \sin\theta, \frac{\partial y}{\partial \theta} = r\cos\theta.$$

Therefore, the Jacobian determinant is $J = r\cos^2\theta + r\sin^2\theta = r$. By the Jacobian theorem, the density of (r, θ) is

$$f_{r,\theta}(r, \theta) = rg(r),$$

with r, θ belonging to a suitable rectangle that will depend on the exact set of values (x, y) on which the original joint density $f(x, y)$ is strictly positive. But, in any case, we have established that the joint density of (r, θ) factorizes into the product form on a rectangle, so *in any spherically symmetric situation, the polar coordinates r and θ are independent, a very convenient fact.* In a spherically symmetric case, r will always have the density *crg(r)* on some interval and for some suitable normalizing constant c, and θ will have a uniform density on some interval.

Now consider three specific choices of the original density function. First consider the uniform case:

$$f(x, y) = \frac{1}{\pi}, 0 < x^2 + y^2 < 1.$$

Then $g(r) = \frac{1}{\pi}, 0 < r < 1$. So, in this case, r has the density $2r, 0 < r < 1$, and θ has the uniform density $\frac{1}{2\pi}, -\pi < \theta < \pi$.

Next consider the case of two independent standard normals. Indeed, in this case, the joint density is spherically symmetric;

$$f(x, y) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}, -\infty < x, y < \infty.$$

Thus, $g(r) = \frac{1}{2\pi}e^{-r^2/2}, r > 0$. Therefore, in this case, r has the *Weibull density* $re^{-r^2/2}, r > 0$, and θ again is uniform on $(-\pi, \pi)$.

Finally, consider the case of two independent *folded standard normals*; i.e., each of X, Y has the density $\sqrt{\frac{2}{\pi}}e^{-x^2/2}, x > 0$. In this case, r varies on $(0, \infty)$ but θ varies on $(0, \frac{\pi}{2})$. Thus, r and θ are still independent, but this time θ is uniform on $(0, \frac{\pi}{2})$, while r still has the same Weibull density $re^{-r^2/2}, r > 0$.

Example 13.15 (Usefulness of the Polar Transformation). Suppose (X, Y) are jointly uniform in the unit circle. We will use the joint density of (r, θ) to find the answers to a number of questions.

First, by using the polar transformation,

$$\begin{aligned} E(X + Y) &= E[r(\cos\theta + \sin\theta)] \\ &= E(r)E(\cos\theta + \sin\theta). \end{aligned}$$

Now, $E(r) < \infty$ and

$$\begin{aligned} E(\cos\theta + \sin\theta) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}(\cos\theta + \sin\theta)d\theta \\ &= \frac{1}{2\pi}(0 + 0) = 0. \end{aligned}$$

Therefore, $E(X+Y)=0$. It should be noted that actually each of X, Y has mean zero in this case, so we could have proved that $E(X+Y)=0$ directly, too.

Next, suppose we want to find the probability that (X, Y) lies in the intersection of the spherical shell $\frac{1}{4} \le \sqrt{X^2+Y^2} \le \frac{3}{4}$ and the cone $X, Y > 0$, $\frac{1}{\sqrt{3}} \le \frac{X}{Y} \le \sqrt{3}$. This looks like a hard problem! But polar coordinates will save us. Transforming to the polar coordinates, this probability is

$$P\left(\frac{1}{4} \le r \le \frac{3}{4}, \frac{\pi}{6} \le \theta \le \frac{\pi}{3}\right) = P\left(\frac{1}{4} \le r \le \frac{3}{4}\right) P\left(\frac{\pi}{6} \le \theta \le \frac{\pi}{3}\right)$$

$$= \int_{\frac{1}{4}}^{\frac{3}{4}} 2r dr \times \frac{1}{12} = \frac{1}{24}.$$

It would have been a much more tedious calculation to do this using the original rectangular coordinates.

Example 13.16 (Product of n Uniforms). Let $X_1, X_2, \ldots, X_n$ be independent $U[0,1]$ variables, and suppose we want to find the density of the product $U = U_n = \prod_{i=1}^{n} X_i$. According to our general discussion, we have to choose $n-1$ other functions and then apply the Jacobian theorem. Define

$$u_1 = x_1, u_2 = x_1 x_2, u_3 = x_1 x_2 x_3, \ldots, u_n = x_1 x_2 \ldots x_n.$$

This is a one-to-one transformation, and the inverse functions are $x_i = \frac{u_i}{u_{i-1}}, 2 \le i \le n; x_1 = u_1$. Thus, the Jacobian matrix of the partial derivatives is lower triangular, and therefore the Jacobian determinant equals the product of the diagonal elements

$$J = \prod_{i=1}^{n} \frac{\partial x_i}{\partial u_i} = \frac{1}{\prod_{i=1}^{n-1} u_i}.$$

Now, applying the Jacobian density theorem, the joint density of $U_1, U_2, \ldots, U_n$ is

$$f_{U_1, U_2, \ldots, U_n}(u_1, u_2, \ldots, u_n) = \frac{1}{\prod_{i=1}^{n-1} u_i},$$

$0 < u_n < u_{n-1} < \cdots < u_1 < 1$.

On integrating out $u_1, u_2, \ldots, u_{n-1}$, we get the density of U_n:

$$f_{U_n}(u) = \int_u^1 \int_{u_{n-1}}^1 \cdots \int_{u_2}^1 \frac{1}{u_1 u_2 \cdots u_{n-1}} du_1 du_2 \cdots du_{n-2} du_{n-1}$$

$$= \frac{|(\log u)^{n-1}|}{(n-1)!},$$

$0 < u < 1$. This example illustrates that applying the Jacobian theorem needs careful manipulation with multiple integrals and that skills in using the Jacobian technique are very important in deriving distributions of functions of many variables.

13.6 Synopsis

(a) If X and Y are continuous random variables, then the CDF of $Z = X + Y$ is given by

$$P(Z \leq z) = \int_{-\infty}^{\infty} F_{X|Y}(z-y)\, f_Y(y) dy,$$

where $F_{X|Y}(u)$ denotes the CDF of the conditional distribution of X given Y and $f_Y(y)$ denotes the marginal density of Y. If X and Y are independent, then this simplifies to

$$P(X + Y \leq z) = \int_{-\infty}^{\infty} F_X(z-y)\, f_Y(y) dy,$$

and the density of $X + Y$ is given by

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y)\, f_Y(y) dy.$$

This is called the density of the convolution of the distributions of X, Y, or simply the density of the convolution.

(b) If X and Y have a joint density $f(x, y)$, then the density of their product $U = XY$ is given by

$$f_U(u) = \int_{-\infty}^{\infty} \frac{1}{|x|} f\left(x, \frac{u}{x}\right) dx$$

and the density of the quotient $V = \frac{X}{Y}$ is given by

$$f_V(v) = \int_{-\infty}^{\infty} |y| f(vy, y) dy.$$

(c) (i) The ratio of two iid standard normal variables has a standard Cauchy distribution.

(ii) If $X \sim N(0,1)$, $Z \sim \chi^2_m$, X and Z are independent, and $Y = \sqrt{\frac{Z}{m}}$, then the ratio $V = \frac{X}{Y}$ has the t distribution with m degrees of freedom, with density given by

$$f_V(v) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})(1 + \frac{v^2}{m})^{(m+1)/2}}, -\infty < v < \infty.$$

(iii) If $X \sim G(\alpha, 1)$, $Y \sim G(\beta, 1)$, and X and Y are independent, then the ratio $V = \frac{X/\alpha}{Y/\beta}$ has an F distribution with density given by

$$f_V(v) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{v^{\alpha-1}}{(1+v)^{\alpha+\beta}}, 0 < v < \infty.$$

(d) The multivariate Jacobian formula gives the density function of a set of n one-to-one functions of an original set of n continuous variables. It says that if $X_1, X_2, \ldots, X_n$ have the joint density function $f(x_1, x_2, \ldots, x_n)$ and $u_i = g_i(x_1, x_2, \ldots, x_n), 1 \leq i \leq n$ are n real-valued functions of $x_1, x_2, \ldots, x_n$ such that $(x_1, x_2, \ldots, x_n) \rightarrow (u_1, u_2, \ldots, u_n)$ is a one-to-one function of $(x_1, x_2, \ldots, x_n)$ with the inverse functions $x_i = h_i(u_1, u_2, \ldots, u_n), i = 1, 2, \ldots, n$, then the joint density of $(U_1, U_2, \ldots, U_n)$ is given by

$$f_{U_1,U_2,\ldots,U_n}(u_1, u_2, \ldots, u_n) = f(h_1(u_1, u_2, \ldots, u_n), \ldots, h_n(u_1, u_2, \ldots, u_n)) \times |J|,$$

where $|J|$ denotes the absolute value of the Jacobian determinant J, given by

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \cdots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \cdots & \frac{\partial x_n}{\partial u_n} \end{vmatrix}.$$

This is one of the major theorems in multidimensional probabilities, and it is essential to become skilled in using this theorem correctly.

(e) Transformation to polar coordinates is a useful technique in dealing with spherically symmetric densities. A joint density of the form $f(x, y) = g(\sqrt{x^2 + y^2})$ is called spherically symmetric in two dimensions. If (X, Y) have a spherically symmetric joint density $g(\sqrt{x^2 + y^2})$, then the polar coordinates, defined as

$$r = \sqrt{X^2 + Y^2}, \theta = \arctan\left(\frac{Y}{X}\right),$$

are independently distributed. In particular, the density of r is $crg(r)$ for some constant c and θ has a uniform density on some suitable interval. See the text for additional details.

(f) In the special case where X and Y are iid standard normal, the density of r is $re^{-r^2/2}, 0 < r < \infty$, and θ is uniform on $(-\pi, \pi)$.

13.7 Exercises

Exercise 13.1. Suppose $X \sim U[0, 1]$, that Y has the density $2y, 0 < y < 1$, and that X and Y are independent. Find the densities of XY and $\frac{X}{Y}$.

Exercise 13.2. Suppose $X \sim U[0,1]$, that Y has the density $2y, 0 < y < 1$, and that X and Y are independent. Find the densities of $X+Y, X-Y, |X-Y|$.

Exercise 13.3. Suppose (X,Y) have the joint pdf $f(x,y) = c(x+y)e^{-x-y}$, $x, y > 0$.

(a) Are X and Y independent?
(b) Find the normalizing constant c.
(c) Find the density of $X+Y$.

Exercise 13.4. Suppose X and Y have the joint density $cxy, 0 < x < y < 1$.

(a) Are X and Y independent?
(b) Find the normalizing constant c.
(c) Find the density of XY.

Exercise 13.5. * **(A Conditioning Argument).** Suppose a fair coin is tossed twice and the number of heads obtained is N. Let X and Y be independent $U[0,1]$ variables and independent of N. Find the density of NXY.

Exercise 13.6. Suppose $X \sim U[0,a], Y \sim U[0,b], Z \sim U[0,c], 0 < a < b < c$, and that X, Y, Z are independent. Let $m = \min\{X,Y,Z\}$. Find expressions for $P(m = X), P(m = Y), P(m = Z)$.

Exercise 13.7. Suppose X and Y are independent standard exponential random variables. Find the densities of XY and $\frac{XY}{(X+Y)^2}$.

Hint: Use $\frac{Y}{X+Y} = 1 - \frac{X}{X+Y}$, and see the examples in the text.

Exercise 13.8. * **(Uniform in a Circle).** Suppose (X,Y) are jointly uniform in the unit circle. By transforming to polar coordinates, find the expectations of $\frac{XY}{X^2+Y^2}$ and of $\frac{XY}{\sqrt{X^2+Y^2}}$.

Exercise 13.9. * **(Length of Bivariate Uniform).** Suppose X and Y are independent $U[0,1]$ variables.

(a) Find the densities of X^2+Y^2 and $P(X^2+Y^2 \le 1)$.
(b) Show that $E(\sqrt{X^2+Y^2}) \approx .765$.

Hint: It is best to do this directly and not try polar coordinates.

Exercise 13.10. Suppose (X,Y) have the joint CDF $F(x,y) = x^3y^2, 0 \le x, y \le 1$. Find the densities of XY and $\frac{X}{Y}$.

Exercise 13.11. * **(Distance Between Two Random Points).** Suppose $P = (X,Y)$ and $Q = (Z,W)$ are two picked points independently from the unit circle, each according to a uniform distribution in the circle. What is the average distance between P and Q?

Exercise 13.12. * **(Distance from the Boundary).** A point is picked uniformly from the unit square. What is the expected value of the distance of the point from the boundary of the unit square?

Exercise 13.13. Suppose X and Y are independent standard normal variables. Find the values $P(\frac{X}{Y} < 1)$ and of $P(X < Y)$. Why are they not the same?

Exercise 13.14 (A Normal Calculation). A marksman is going to take two shots at a bull's eye. The distances from the bull's eye of the first and second shots are distributed as $(|X|, |Y|)$, where $X \sim N(0, \sigma^2), Y \sim N(0, \tau^2)$, and X and Y are independent. Find a formula for the probability that the second shot is closer to the target.

Exercise 13.15. * **(Quotient in Bivariate Normal).** Suppose (X, Y) have a bivariate normal distribution with zero means, unit standard deviations, and a correlation ρ. Show that $\frac{X}{Y}$ still has a Cauchy distribution.

Exercise 13.16. * **(Product of Beta).** Suppose X and Y are independent $Be(\alpha, \beta)$, $Be(\gamma, \delta)$ random variables. Find the density of XY. Do you recognize the form?

Exercise 13.17. * **(Product of Normals).** Suppose X and Y are independent standard normal variables. Find the density of XY.

Hint: The answer will involve a Bessel function K_0.

Exercise 13.18. * **(Product of Cauchy).** Suppose X and Y are independent standard Cauchy variables. Derive a formula for the density of XY.

Exercise 13.19. Prove that the square of a t random variable has an F- distribution.

Exercise 13.20. * **(Box-Mueller Transformation).** Suppose X and Y are independent $U[0, 1]$ variables. Let $U = \sqrt{-2 \log X} \cos(2\pi Y), V = \sqrt{-2 \log X} \sin(2\pi Y)$. Show that U and V are independent and that each is standard normal.

Remark. This is a convenient method for generating standard normal values by using only uniform random numbers.

Exercise 13.21. * **(Deriving a General Formula).** Suppose (X, Y, Z) have a joint density of the form $f(x, y, z) = g(x + y + z), x, y, z > 0$. Find a formula for the density of $X + Y + Z$.

Exercise 13.22. Suppose (X, Y, Z) have a joint density $f(x, y, z) = \frac{6}{(1+x+y+z)^4}$, $x, y, z > 0$. Find the density of $X + Y + Z$.

Exercise 13.23 (Deriving a General Formula). Suppose $X \sim U[0, 1]$ and Y is an arbitrary continuous random variable. Derive a general formula for the density of $X + Y$.

Exercise 13.24 (Convolution of Uniform and Exponential). Let $X \sim U[0, 1]$, $Y \sim Exp(\lambda)$, and X and Y are independent. Find the density of $X + Y$.

Exercise 13.25 (Convolution of Uniform and Normal). Let $X \sim U[0,1], Y \sim N(\mu, \sigma^2)$, and X and Y are independent. Find the density of $X + Y$.

Exercise 13.26 (Convolution of Uniform and Cauchy). Let $X \sim U[0,1], Y \sim C(0,1)$, and X and Y are independent. Find the density of $X + Y$.

Exercise 13.27 (Convolution of Uniform and Poisson). Let $X \sim U[0,1], Y \sim Poi(\lambda)$, and let X and Y be independent. Find the density of $X + Y$.

Exercise 13.28. * **(Bivariate Cauchy).** Suppose (X, Y) has the joint pdf $f(x, y) = \frac{c}{(1+x^2+y^2)^{3/2}}, -\infty < x, y < \infty$.

(a) Find the normalizing constant c.
(b) Are X and Y independent?
(c) Find the densities of the polar coordinates r, θ.
(d) Find $P(X^2 + Y^2 \leq 1)$.

Exercise 13.29. * Suppose X, Y, Z are independent standard exponentials. Find the joint density of $\frac{X}{X+Y+Z}, \frac{X+Y}{X+Y+Z}, X + Y + Z$.

Exercise 13.30 (Correlation). Suppose X and Y are independent $U[0,1]$ variables. Find the correlation between $X + Y$ and $\sqrt{XY}$.

Exercise 13.31 (Correlation). Suppose X and Y are jointly uniform in the unit circle. Find the correlation between XY and $X^2 + Y^2$.

Exercise 13.32. * **(Sum and Difference of General Exponentials).** Suppose $X \sim Exp(\lambda), Y \sim Exp(\mu)$, and that X and Y are independent. Find the densities of $X + Y$ and $X - Y$.

Exercise 13.33. * **(Double Exponential Convolution).** Suppose X and Y are independent standard double exponentials, each with the density $\frac{1}{2}e^{-|x|}, -\infty < x < \infty$. Find the density of $X + Y$.

Exercise 13.34. * Let X, Y, Z be independent standard normals. Show that $\frac{X+YZ}{\sqrt{1+Z^2}}$ has a normal distribution.

Exercise 13.35. * **(Decimal Expansion of a Uniform).** Let $X \sim N(0,1)$ and suppose $X = .n_1 n_2 n_3 \ldots$ is the decimal expansion of X. Find the marginal and joint distributions of $n_1, n_2, \ldots, n_k$, for $k \geq 1$.

Exercise 13.36. * **(Integer Part and Fractional Part).** Let X be a standard exponential variable. Find the joint distribution of the integer part and the fractional part of X. Note that they do not have a joint density.

Exercise 13.37. * **(Factorization of Chi-square).** Suppose X has a chi-square distribution with one degree of freedom. Find nonconstant independent random variables Y and Z such that YZ has the same distribution as X.

Hint: Look at the text.

Exercise 13.38. * **(Multivariate Cauchy).** Suppose $X_1, X_2, \ldots, X_n$ have the joint density $f(x_1, \ldots, x_n) = \frac{c}{(1+x_1^2+\cdots+x_n^2)^{\frac{n+3}{2}}}$, where c is a normalizing constant.

Find the density of $X_1^2 + X_2^2 + \cdots + X_n^2$.

Exercise 13.39 (Ratio of Independent Chi-squares). Suppose $X_1, X_2, \ldots, X_m$ are independent $N(\mu, \sigma^2)$ variables and $Y_1, Y_2, \ldots, Y_n$ are independent $N(\theta, \tau^2)$ variables. Assume also that all $m + n$ variables are independent. Show that $\frac{(n-1)\tau^2 \sum_{i=1}^{m}(X_i-\bar{X})^2}{(m-1)\sigma^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}$ has an F distribution.

Exercise 13.40 (Use Your Computer). Use the definition of a t distribution in terms of a normal and a chi-square variable to simulate 100 values from a t distribution with n degrees of freedom, $n = 10, 20, 40$. Find the variance of your simulated values. Do you see a trend in the variance as you increase n? Would you expect to?

References

Hinkley, D. (1969). On the ratio of two correlated normal random variables, Biometrika, 56(3), 635–639.

Chapter 14
Markov Chains and Applications

In many applications, successive observations of a process, say $X_1, X_2, \ldots$, have an inherent time component associated with them. For example, the X_i could be the state of the weather at a particular location on the ith day, counting from some fixed day. In a simplistic model, the state of the weather could be "dry" or "wet," quantified as, say, 0 and 1. It is hard to believe that in such an example, the sequence $X_1, X_2, \ldots$ could be mutually independent. The question then arises how to model the dependence among the $X_i's$. Probabilists have numerous dependency models. A particular model that has earned a very special status is called the *Markov chain*. In a Markov chain model, we assume that the future, given the entire past and the present state of a process, depends only on the present state. In the weather example, suppose we want to assign a probability that tomorrow, say March 10, will be dry, and suppose that we have available to us the precipitation history for each day from January 1 to March 9. In a Markov chain model, our probability that March 10 will be dry will depend only on the state of the weather on March 9, even though the entire past precipitation history was available to us. As simple as it sounds, Markov chains are enormously useful in applications, perhaps more than any other specific dependency model. They also are independently relevant to statistical computing in very important ways. The topic has an incredibly rich and well-developed theory, with links to many other topics in probability theory. Familiarity with basic Markov chain terminology and theory is often considered essential for anyone interested in studying statistics and probability. We present an introduction to basic Markov chain theory in this chapter.

Feller (1968), Freedman (1975), and Isaacson and Madsen (1976) are classic references on Markov chains. Modern treatments are available in Bhattacharya and Waymire (2009), Brémaud (1999), Meyn and Tweedie (1993), Norris (1997), Seneta (1981), and Stirzaker (1994). Classic treatments of the problem of gambler's ruin are available in Feller (1968) and Kemperman (1950). Numerous interesting examples at more advanced levels are available in Diaconis (1988); sophisticated applications at an advanced level are also available in Bhattacharya and Waymire (2009).

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_14,

14.1 Notation and Basic Definitions

Definition 14.1. A sequence of random variables $\{X_n\}, n \geq 0$, is said to be a *Markov chain* if, for some countable set $S \subseteq \mathcal{R}$ and any $n \geq 1, x_{n+1}, x_n, \ldots, x_0 \in S$,

$$P(X_{n+1} = x_{n+1} \,|X_0 = x_0, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1} \,|X_n = x_n).$$

The set S is called the *state space of the chain*. If S is a finite set, the chain is called a *finite-state Markov chain*. X_0 is called the *initial state*.

Without loss of generality, we can denote the elements of S as $1, 2, \ldots$, although in some examples we may use the original labeling of the states to avoid confusion.

Definition 14.2. The distribution of the initial state X_0 is called the *initial distribution* of the chain. We will denote the pmf of the initial distribution as $\lambda_i = P(X_0 = i)$.

Definition 14.3. A Markov chain $\{X_n\}$ is called *homogeneous* or *stationary* if $P(X_{n+1} = y|X_n = x)$ is independent of n for any x, y.

Definition 14.4. Let $\{X_n\}$ be a stationary Markov chain. Then the probabilities $p_{ij} = P(X_{n+1} = j \,|X_n = i)$ are called the *one-step transition probabilities*, or simply transition probabilities. The matrix $P = ((p_{ij}))$ is called the *transition probability matrix*.

Definition 14.5. Let $\{X_n\}$ be a stationary Markov chain. Then the probabilities $p_{ij}(n) = P(X_{n+m} = j \,|X_m = i) = P(X_n = j \,|X_0 = i)$ are called the *n-step transition probabilities*, and the matrix $P^{(n)} = ((p_{ij}(n)))$ is called the *n-step transition probability matrix*.

Remark. If the state space of the chain is finite and has, say, t elements, then the transition probability matrix P is a $t \times t$ matrix. Note that $\sum_{j \in S} p_{ij}$ is always one. A matrix with this property is called a *stochastic matrix*.

Definition 14.6. A $t \times t$ square matrix P is called a stochastic matrix if, for each $i, \sum_{j=1}^{t} p_{ij} = 1$. It is called doubly stochastic or bi-stochastic if, in addition, for every $j, \sum_{i=1}^{t} p_{ij} = 1$. Thus, a transition probability matrix is always a stochastic matrix.

Example 14.1 (Weather Pattern). Suppose that, in some particular city, any day is either dry or wet. If it is dry on some day, it remains dry the next day with probability α and will be wet with the residual probability $1 - \alpha$. On the other hand, if it is wet on some day, it remains wet the next day with probability β and becomes dry with probability $1 - \beta$. Let $X_0, X_1, \ldots$ be the sequence of states of the weather, with X_0 being the state on the initial day (on which observation starts). Then $\{X_n\}$ is a two-state stationary Markov chain with the transition probability matrix

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}.$$

Example 14.2 (Voting Preferences). Suppose that in a presidential election voters can vote for either the Labor Party, the Conservative Party, or the Independent Party. Someone who has voted for the Labor candidate in this election will vote Labor again with 80% probability, will switch to Conservative with 5% probability, and vote Independent with 15% probability. Someone who has voted for the Conservative candidate in this election will vote Conservative again with 90% probability, switch to Labor with 3% probability, and vote Independent with 7% probability. Someone who has voted for the Independent candidate in this election will vote Independent again with 80% probability or switch to one of the other parties with 10% probability each. This is a three-state stationary Markov chain with state space $S = \{1, 2, 3\} \equiv \{\text{Labor, Conservative, Independent}\}$ and the transition matrix

$$P = \begin{pmatrix} .8 & .05 & .15 \\ .03 & .9 & .07 \\ .1 & .1 & .8 \end{pmatrix}.$$

Example 14.3 (An Urn Model Example). Two balls, say A and B, are initially in urn 1, and two others, say C and D, are in urn 2. In each successive trial, one ball is chosen at random from urn 1 and one independently and also at random from urn 2, and these balls switch urns. We let X_n denote the vector of locations of the four balls A, B, C, D, in that order of the balls, after the nth trial. Thus, $X_{10} = 1122$ means that, after the tenth trial, A and B are located in urn 1 and C and D in urn 2, etc. Note that $X_0 = 1122$. Two of the four balls are always in urn 1 and two in urn 2. Thus, the possible number of states is $\binom{4}{2} = 6$. They are $1122, 1212, 1221, 2112, 2121, 2211$. Then $\{X_n\}$ is a six-state stationary Markov chain. What are the transition probabilities?

For notational convenience, denote the six states above as $1, 2, \ldots, 6$, respectively. For the state of the chain to move from state 1 to state 2 in one trial, B and C have to switch their urns. This will happen with probability $.5 \times .5 = .25$. As another example, for the state of the chain to move from state 1 to state 6, all of the four balls must switch their urns. This is not possible. Therefore, this transition probability is zero. Also, note that if the chain is in some state now, it *cannot* remain in that same state after the next trial. Thus, all diagonal elements in the transition probability matrix must be zero. Indeed, the transition probability matrix is

$$P = \begin{pmatrix} 0 & .25 & .25 & .25 & .25 & 0 \\ .25 & 0 & .25 & .25 & 0 & .25 \\ .25 & .25 & 0 & 0 & .25 & .25 \\ .25 & .25 & 0 & 0 & .25 & .25 \\ .25 & 0 & .25 & .25 & 0 & .25 \\ 0 & .25 & .25 & .25 & .25 & 0 \end{pmatrix}.$$

Notice that there are really three distinct rows in P, each occurring twice. It is easy to argue that is how it should be in this particular urn experiment. Also note the very interesting fact that in each row and each column there are two zeros, and the nonzero entries obviously add to 1. This is an example of a transition probability matrix that is *doubly stochastic*. Markov chains with a doubly stochastic transition probability matrix show a unified long-run behavior. By definition, initially the chain is in state 1, so $P(X_0 = 1) = 1, P(X_0 = i) = 0\ \forall i \neq 0$. However, after many trials, the state of the chain would be any of the six possible states with essentially an equal probability; i.e., $P(X_n = i) \approx \frac{1}{6}$ for each possible state i for large n. This unifying long-run behavior of Markov chains with a doubly stochastic transition probability matrix is a significant result with wide applications in Markov chain theory.

Example 14.4 (Urn Model II: Ehrenfest Model). This example has wide applications in the theory of heat transfer. The mathematical model is that we initially have m balls, some in one urn, say urn I, and the rest in another urn, say urn II. At each subsequent time $n = 1, 2, \ldots$, one ball among the m balls is selected at random. If the ball is in urn I, with probability α it is transferred to urn II and with probability $1 - \alpha$ it continues to stay in urn I. If the ball is in urn II, with probability β it is transferred to urn I and with probability $1 - \beta$ it continues to stay in urn II.

Let X_0 be the number of balls initially in urn I and X_n the number of balls in urn I after time n. Then $\{X_n\}$ is a stationary Markov chain with state space $S = \{0, 1, \ldots, m\}$. If there are, say, i balls in urn I at a particular time, then at the next instant urn I could lose one ball, gain one ball, or neither lose nor gain any ball. It loses a ball if one of its i balls gets selected for possible transfer and then the transfer actually happens. So $p_{i,i-1} = \frac{i}{m}\alpha$. Using this simple argument, we get as the one-step transition probabilities

$$p_{i,i-1} = \frac{i}{m}\alpha;\ p_{i,i+1} = \frac{m-i}{m}\beta;\ p_{ii} = 1 - \frac{i}{m}\alpha - \frac{m-i}{m}\beta,$$

and $p_{ij} = 0$ if $j \neq i - 1, i, i + 1$.

As a specific example, suppose $m = 7$ and $\alpha = \beta = \frac{1}{2}$. Then the transition matrix on the state space $S = \{0, 1, \ldots, 7\}$ can be worked out by using the formulas given just above, and it is

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{14} & \frac{1}{2} & \frac{3}{7} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{7} & \frac{1}{2} & \frac{5}{14} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{14} & \frac{1}{2} & \frac{2}{7} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{2}{7} & \frac{1}{2} & \frac{3}{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{5}{14} & \frac{1}{2} & \frac{1}{7} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{3}{7} & \frac{1}{2} & \frac{1}{14} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Example 14.5 (Machine Maintenance). Of the machines in a factory, a certain number break down or are identified to be in need of maintenance on a given day. They are sent to a maintenance shop the next morning. The maintenance shop is capable of finishing its maintenance work on some k machines on any given day. We are interested in the sequence $\{X_n\}$, where X_n denotes the number of machines in the maintenance shop on the nth day, $n \geq 0$. We may take $X_0 = 0$.

Let Z_0 machines break down on day zero. Then, $X_1 = Z_0$. Of these, *up to* k machines can be fixed by the shop on that day, and these are returned. But now, on day 1, some Z_1 machines break down at the factory, so that $X_2 = \max\{X_1 - k, 0\} + Z_1 = \max\{Z_0 - k, 0\} + Z_1$, of which up to k machines can be fixed by the shop on the second day itself, and those are returned to the factory. We then have $X_3 = \max\{X_2 - k, 0\} + Z_2$

$$\begin{aligned} &= Z_0 + Z_1 + Z_2 - 2k \text{ if } Z_0 \geq k, Z_0 + Z_1 \geq 2k, \\ &= Z_2 \text{ if } Z_0 \geq k, Z_0 + Z_1 < 2k, \\ &= Z_1 + Z_2 - k \text{ if } Z_0 < k, Z_1 \geq k, \\ &= Z_2 \text{ if } Z_0 < k, Z_1 < k, \end{aligned}$$

and so on.

If $Z_i, i \geq 0$ are iid, then $\{X_n\}$ forms a stationary Markov chain. The state space of this chain is $\{0, 1, 2, \ldots\}$. What is the transition probability matrix? For simplicity, take $k = 1$. For example, $P(X_2 = 1 \,|X_1 = 0) = P(Z_1 = 1 \,|Z_0 = 0) = P(Z_1 = 1) = p_1$ (say). On the other hand, as another example, $P(X_2 = 2 \,|X_1 = 4) = 0$. If we denote the common mass function of the Z_i by $P(Z_i = j) = p_j, j \geq 0$, then the transition probability matrix is

$$P = \begin{pmatrix} p_0 & p_1 & p_2 & p_3 & \cdots \\ p_0 & p_1 & p_2 & p_3 & \cdots \\ 0 & p_0 & p_1 & p_2 & \cdots \\ 0 & 0 & p_0 & p_1 & \cdots \\ 0 & 0 & 0 & p_0 & \cdots \\ & & \vdots & & \end{pmatrix}.$$

Example 14.6 (Hopping Mosquito). Suppose a mosquito makes movements between the forehead, the left cheek, and the right cheek of an individual, which we designate as states $1, 2, 3$, according to the following rules. If at some time n the mosquito is sitting on the forehead, then it will definitely move to the left cheek at the next time $n + 1$; if it is sitting on the left cheek, it will stay there or move to the right cheek with probability .5 each and if it is on the right cheek, it will stay there or move to the forehead with probability .5 each.

Then the sequence of locations of the mosquito forms a three-state Markov chain with the one-step transition probability matrix

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & .5 & .5 \\ .5 & 0 & .5 \end{pmatrix}.$$

Example 14.7 (An Example from Genetics). Many traits in organisms, for example humans, are determined by genes. For example, eye color in humans is determined by a pair of genes. Genes can come in various forms or versions, which are called *alleles*. An offspring receives one allele from each parent. A parent contributes one of his or her alleles to an offspring with equal probability, and the parents make their contributions independently. Certain alleles dominate over others. For example, the allele for blue eye color is dominated by the allele for brown eye color. The allele for blue color would be called *recessive*, and the allele for brown eye color would be called *dominant*. If we denote these as b and B, respectively, then a person may have the pair of alleles BB, Bb, or bb. They are called the dominant, hybrid, and recessive *genotypes*, respectively. We denote them as d, h, and r, respectively. Consider now the sequence of genotypes of descendants of an initial individual, and denote the sequence as $\{X_n\}$; for any n, X_n must be one of d, h, or r (we may call them 1, 2, 3).

Consider now a person with an unknown genotype (X_0) mating with a known hybrid. Suppose he has genotype d. He will necessarily contribute B to the offspring. Therefore, the offspring can only have genotype d or h, and not r. It will be d if the offspring also gets the B allele from the mother, and it will be h if the offspring gets b from the mother. The chance of each is $\frac{1}{2}$. Therefore, the transition probability $P(X_1 = d \mid X_0 = d) = P(X_1 = h \mid X_0 = d) = \frac{1}{2}$, and $P(X_1 = r \mid X_0 = d) = 0$.

Suppose $X_0 = h$. Then the father contributes B or b with probability $\frac{1}{2}$ each, and so does the mother, who was assumed to be a hybrid. So the probabilities that $X_1 = d, h, r$ are respectively $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$.

If $X_0 = r$, then X_1 can only be h or r, with probability $\frac{1}{2}$ each. So, if we assume this same mating scheme over generations, then $\{X_n\}$ forms a three-state stationary Markov chain with the transition probability matrix

$$P = \begin{pmatrix} .5 & .5 & 0 \\ .25 & .5 & .25 \\ 0 & .5 & .5 \end{pmatrix}.$$

Example 14.8 (Simple Random Walk). Consider a particle starting at the origin at time zero and making independent movements of one step to the right or one step to the left at each successive time instant $1, 2, \ldots$. Assume that the particle moves to the right at any particular time with probability p and to the left with probability $q = 1 - p$. The mathematical formulation is that the successive movements are iid random variables $X_1, X_2, \ldots$ with common pmf $P(X_i = 1) = p, P(X_i = -1) = q$. The particle's location after the nth step has been taken is denoted as $S_n = X_0 + X_1 + \cdots + X_n = X_1 + \cdots + X_n$, assuming that $X_0 = 0$ with probability 1. Since at each time the particle can move by just one unit, $\{S_n\}$ is

a stationary Markov chain with state space $S = \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ and with the transition probabilities

$$\begin{aligned} p_{ij} &= P(X_{n+1} = j \mid X_n = i) \\ &= p \text{ if } j = i + 1, \\ &= q \text{ if } j = i - 1, \\ &= 0 \text{ if } |j - i| > 1, \end{aligned}$$

$i, j \in \mathbb{Z}$.

By virtue of the importance of random walks in theory and applications of probability, this is an important example of a stationary Markov chain. Note that the chain is stationary because the individual steps X_i are iid. This is also an example of a Markov chain with an infinite state space.

14.2 Chapman-Kolmogorov Equation

The Chapman-Kolmogorov equation provides a simple method for obtaining the higher-order transition probabilities of a Markov chain in terms of lower-order transition probabilities. Carried to its most convenient form, the equation describes how to calculate by a simple and explicit method all higher-order transition probabilities in terms of the one-step transition probabilities. Because we always start analyzing a chain with the one-step probabilities, it is evidently very useful to know how to calculate all higher-order transition probabilities using just the knowledge of the one-step transition probabilities.

Theorem 14.1 (Chapman-Kolmogorov Equation). *Let $\{X_n\}$ be a stationary Markov chain with the state space S. Let $n, m \geq 1$. Then,*

$$p_{ij}(m + n) = P(X_{m+n} = j \mid X_0 = i) = \sum_{k \in S} p_{ik}(m) p_{kj}(n).$$

Proof. A verbal proof is actually the most easily understood. In order to get to state j from state i in $m + n$ steps, the chain *must* go to *some state* $k \in S$ in m steps and then travel from that k to the state j in the next n steps. By adding over all possible $k \in S$, we get the Chapman-Kolmogorov equation.

An extremely important corollary is the following result.

Corollary. Let $P^{(n)}$ denote the n-step transition probability matrix. Then, for all $n \geq 2$, $P^{(n)} = P^n$, where P^n denotes the usual nth power of P.

Proof. From the Chapman-Kolmogorov equation, by using the definition of a matrix product, for all $m, n \geq 1$, $P^{(m+n)} = P^{(m)} P^{(n)} \Rightarrow P^{(2)} = PP = P^2$. We now

finish the proof by induction. Suppose $P^{(n)} = P^n \ \forall n \leq k$. Then, $P^{(k+1)} = P^{(k)}P = P^k P = P^{k+1}$, which finishes the proof.

A further important consequence is that we can now write an explicit formula for the pmf of the state of the chain at a given time n.

Proposition. Let $\{X_n\}$ be a stationary Markov chain with the state space S and one-step transition probability matrix P. Fix $n \geq 1$. Then, $\lambda_n(i) = P(X_n = i) = \sum_{k \in S} p_{ki}(n) P(X_0 = k)$. In matrix notation, if $\lambda = (\lambda_1, \lambda_2, \ldots)'$ denotes the vector of the initial probabilities $P(X_0 = k), k = 1, 2, \ldots$, and if λ_n denotes the row vector of probabilities $P(X_n = i), i = 1, 2, \ldots$, then $\lambda_n = \lambda P^n$.

This is an important formula because it lays out how to explicitly find the distribution of X_n from the *initial distribution* λ and the one-step transition matrix P.

Example 14.9 (Weather Pattern). Consider once again the weather pattern example with the one-step transition probability matrix

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}.$$

We let the states be 1, 2 (1 = dry; 2 = wet). We use the Chapman-Kolmogorov equation to answer two questions. First, suppose it is Wednesday today and it is dry. We want to know the probability that Saturday will be dry. In notation, we want to find $p_{11}^{(3)} = P(X_3 = 1 \,|\, X_0 = 1)$. In order to get a concrete numerical answer at the end, let us take $\alpha = \beta = .8$. Now, by direct matrix multiplication,

$$P^3 = \begin{pmatrix} .608 & .392 \\ .392 & .608 \end{pmatrix}.$$

Therefore, the probability that Saturday will be dry if Wednesday is dry is $p_{11}^{(3)} = .608$.

Next, suppose that we want to know the probability that Saturday and Sunday will both be dry if Wednesday is dry. In notation, we now want to find

$$\begin{aligned} &P(X_3 = 1, X_4 = 1 \,|\, X_0 = 1) \\ &\quad = P(X_3 = 1 \,|\, X_0 = 1) P(X_4 = 1 \,|\, X_3 = 1, X_0 = 1) \\ &\quad = P(X_3 = 1 \,|\, X_0 = 1) P(X_4 = 1 \,|\, X_3 = 1) = .608 \times .8 \\ &\quad = .4864. \end{aligned}$$

Coming now to evaluating the pmf of X_n itself, we calculate it as $P(X_n = i) = \sum_{k \in S} p_{ki}(n) P(X_0 = k)$. Denote $P(\text{The initial day was dry}) = \lambda_1$, $P(\text{The initial day was wet}) = \lambda_2, \lambda_1 + \lambda_2 = 1$. Let us evaluate the probabilities that it will be dry one week, two weeks, three weeks, or four weeks from the initial day. This requires calculation of, respectively, $P^7, P^{14}, P^{21}, P^{28}$. For example,

$$P^7 = \begin{pmatrix} .513997 & .486003 \\ .486003 & .513997 \end{pmatrix}.$$

Therefore,

$$\begin{aligned} P(\text{It will be dry one week from the initial day}) &= .513997\lambda_1 + .486003\lambda_2 \\ &= .5 + .013997(\lambda_1 - \lambda_2). \end{aligned}$$

Similarly, we can compute P^{14} and show that

$$\begin{aligned} P(\text{It will be dry two weeks from the initial day}) &= .500392\lambda_1 + .499608\lambda_2 \\ &= .5 + .000392(\lambda_1 - \lambda_2). \\ P(\text{It will be dry three weeks from the initial day}) &= .5 + .000011(\lambda_1 - \lambda_2). \\ P(\text{It will be dry four weeks from the initial day}) &= .5. \end{aligned}$$

We see that convergence to .5 has occurred, regardless of λ_1, λ_2. That is, regardless of the initial distribution, eventually you will put a 50/50 probability that a day far into the future will be dry or wet. Is this always the case? The answer is no. In this case, convergence to .5 occurred because the one-step transition matrix P has the doubly stochastic characteristic: each row as well as each column of P adds to 1. We will see more about this later.

Example 14.10 (Voting Preferences). Consider the previous example on voting preferences. Suppose we want to know the probabilities that a Labor voter in this election will vote respectively Labor, Conservative, or Independent two elections from now. Denoting the states as 1, 2, 3 in notation, we want to find $P(X_2 = i \,|X_0 = 1), i = 1, 2, 3$. We can answer this by simply computing P^2. Since

$$P = \begin{pmatrix} .80 & .05 & .15 \\ .03 & .90 & .07 \\ .1 & .1 & .8 \end{pmatrix},$$

by direct computation,

$$P^2 = \begin{pmatrix} .66 & .1 & .24 \\ .06 & .82 & .12 \\ .16 & .18 & .66 \end{pmatrix}.$$

Hence, the probabilities that a Labor voter in this election will vote Labor, Conservative, or Independent two elections from now are 66%, 10%, and 24%. We also see from P^2 that a Conservative voter will vote Conservative two elections from now with 82% probability and has a chance of just 6% of switching to Labor, etc.

Example 14.11 (Hopping Mosquito). Consider again the *hopping mosquito* example previously introduced. The goal of this example is to find the n-step transition probability matrix P^n for a general n. We describe a general method for finding P^n using a linear algebra technique known as *diagonalization of a matrix*. If a square matrix P (not necessarily symmetric) of order $t \times t$ has t distinct eigenvalues, say $\delta_1, \ldots, \delta_t$, which are complex numbers in general, and if $\mathbf{u}_1, \ldots, \mathbf{u}_t$ are a set of t

eigenvectors of P corresponding to the eigenvalues $\delta_1, \ldots, \delta_t$, then define a matrix U as $U = (\mathbf{u}_1, \ldots, \mathbf{u}_t)$; i.e., U has $\mathbf{u}_1, \ldots, \mathbf{u}_t$ as its t columns. Then, U has the property that $U^{-1}PU = L$, where L is the diagonal matrix with the diagonal elements $\delta_1, \ldots, \delta_t$. Now, just note that

$$U^{-1}PU = L \Rightarrow P = ULU^{-1} \Rightarrow P^n = UL^nU^{-1},$$

$\forall n \geq 2$.

Therefore, we only need to compute the eigenvalues of P and the matrix U of a set of eigenvectors. As long as the eigenvalues are distinct, the n-step transition matrix will be provided by the unified formula $P^n = UL^nU^{-1}$.

The eigenvalues of our P are

$$\delta_1 = -\frac{i}{2}, \delta_2 = \frac{i}{2}, \delta_3 = 1;$$

note that they are distinct. The eigenvectors (one set) turn out to be

$$\mathbf{u}_1 = \left(-1-i, \frac{i-1}{2}, 1\right)', \mathbf{u}_2 = \left(i-1, -\frac{i+1}{2}, 1\right)', \mathbf{u}_3 = (1,1,1)'.$$

Therefore,

$$U = \begin{pmatrix} -i-1 & i-1 & 1 \\ \frac{i-1}{2} & -\frac{i+1}{2} & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

$$U^{-1} = \begin{pmatrix} \frac{3i-1}{10} & -\frac{2i+1}{5} & \frac{i+3}{10} \\ -\frac{3i+1}{10} & \frac{2i-1}{5} & \frac{3-i}{10} \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{pmatrix}.$$

This leads to

$$P^n = U \begin{pmatrix} \left(-\frac{i}{2}\right)^n & 0 & 0 \\ 0 & \left(\frac{i}{2}\right)^n & 0 \\ 0 & 0 & 1 \end{pmatrix} U^{-1},$$

with U, U^{-1} as above.

For example,

$$\begin{aligned} p_{11}(n) &= (-i-1)\frac{3i-1}{10}\left(-\frac{i}{2}\right)^n + (i-1)\left(-\frac{3i+1}{10}\right)\left(\frac{i}{2}\right)^n + 1\left(\frac{1}{5}\right)1 \\ &= \frac{1}{5} + \frac{2-i}{5}\left(-\frac{i}{2}\right)^n + \frac{2+i}{5}\left(\frac{i}{2}\right)^n; \end{aligned}$$

this is the probability that the mosquito will be back on the forehead after n moves if it started at the forehead. If we take $n = 2$, we get, on doing the complex multiplication, $p_{11}(2) = 0$. We can logically verify that $p_{11}(2)$ must be zero by just looking at the one-step transition matrix P. However, if we take $n = 3$, then the formula will give $p_{11}(3) = \frac{1}{4} > 0$. Indeed, if we take $n = 3$, we get

$$P^3 = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \\ \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \end{pmatrix}.$$

We notice that every element in P^3 is strictly positive. That is, no matter where the mosquito was initially seated, by the time it has made three moves, we cannot rule out any location for where it will be: it can now be *anywhere*. In fact, this property of a transition probability matrix is so important in Markov chain theory that it has a name. It is the first definition in our next section.

14.3 Communicating Classes

Definition 14.7. Let $\{X_n\}$ be a stationary Markov chain with transition probability matrix P. It is called a *regular chain* if there exists a universal $n_0 > 0$ such that $p_{ij}(n_0) > 0 \ \forall i, j \in S$.

So, what we just saw in the last example is that the mosquito is engaged in movements according to a regular Markov chain.

A weaker property is that of *irreducibility*.

Definition 14.8. Let $\{X_n\}$ be a stationary Markov chain with transition probability matrix P. It is called an *irreducible chain* if, for any $i, j \in S, i \neq j$, there exists $n_0 > 0$, possibly depending on i, j such that $p_{ij}(n_0) > 0$.

Irreducibility means that it is possible to travel from any state to any other state, however many steps it might take, depending on which two states are involved.

Another terminology also commonly used is that of *communicating*.

Definition 14.9. Let $\{X_n\}$ be a stationary Markov chain with transition probability matrix P. Let i and j be two specific states. We say that i *communicates with* j $(i \leftrightarrow j)$ if there exists $n_0 > 0$, possibly depending on i, j such that $p_{ij}(n_0) > 0$, and there also exists $n_1 > 0$, possibly depending on i, j such that $p_{ji}(n_1) > 0$.

In words, a pair of specific states i and j are *communicating states* if it is possible to travel back and forth between i and j, however many steps it might take, depending on i, j, and possibly even depending on the direction of the journey; i.e., whether the direction is from i to j or from j to i.

By convention, we say that $i \leftrightarrow i$. Thus, $\leftrightarrow$ defines an *equivalence relation* on the state space S:

$$i \leftrightarrow i; i \leftrightarrow j \Rightarrow j \leftrightarrow i; i \leftrightarrow j, j \leftrightarrow k \Rightarrow i \leftrightarrow k.$$

Therefore, like all equivalence relations, $\leftrightarrow$ partitions the state space S into mutually exclusive subsets of S, say $C_1, C_2, \ldots$. These partitioning subsets $C_1, C_2, \ldots$ are called the *communicating classes* of the chain.

Here is an example to help illustrate the notion.

Example 14.12 (Identifying Communicating Classes). Consider the one-step transition matrix

$$P = \begin{pmatrix} .75 & .25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ .25 & 0 & 0 & .25 & .5 & 0 \\ 0 & 0 & 0 & .75 & .25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Inspecting the transition matrix, we see that $1 \leftrightarrow 2$ because it is possible to go from 1 to 2 in just one step, and conversely, starting at 2, one will always go to 3, and it is then possible to go from 3 to 1. Likewise, $2 \leftrightarrow 3$ because if we are at 2, we will always go to 3, and conversely, if we are at 3, then we can first go to 1 and then from 1 to 2. Next, state 5 and state 6 are communicative, but they are clearly not communicative with any other state because once at 5 we can only go to 6, and once at 6 we can only go to 5. Finally, if we are at 4, then we can go to 5 and from 5 to 6, but 6 does not communicate with 4. So, the communicating classes in this example are

$$C_1 = \{1, 2, 3\}, C_2 = \{4\}, C_3 = \{5, 6\}.$$

Note that they are disjoint and that $C_1 \cup C_2 \cup C_3 = \{1, 2, 3, 4, 5, 6\} = S$. As a further interesting observation, if we are in C_3, then we cannot make even one-way trips to any state outside of C_3. Such a communicating class is called *closed*. In this example, C_3 is the only closed communicating class. For example, $C_1 = \{1, 2, 3\}$ is not a closed class because one can make one-way trips from 1 to 4 or 5. The reader can verify trivially that $C_2 = \{4\}$ is also not a closed class.

We can observe more interesting things about the chain from the transition matrix. Consider, for example, state 5. If you are in state 5, then your transitions would have to be $565656\ldots$. So, starting at 5, you can return to 5 only at times $n = 2k, k \geq 1$. In such a case, we call the state *periodic* with period equal to 2. Likewise, state 6 is also periodic with period equal to 2. An exercise asks you to show that all states within the same communicating class always have the same period. It is useful to have a formal definition because there is an element of subtlety about the exact meaning of the period of a state.

Definition 14.10. A state $i \in S$ is said to have the period $d(> 1)$ if the greatest common divisor of all positive integers n for which $p_{ii}(n) > 0$ is the given number d. If a state i has no period $d > 1$, it is called an *aperiodic state*. If every state of a chain is aperiodic, the chain itself is called an *aperiodic chain*.

Example 14.13 (Computing the Period). Consider the hopping mosquito example again. First, let us look at state 1. Evidently, we can go to 1 from 1 in any number of steps; i.e., $p_{11}(n) > 0 \ \forall n \geq 1$. So the set of integers n for which $p_{11}(n) > 0$ is $\{1, 2, 3, 4, \ldots\}$, and the gcd (greatest common divisor) of these integers is 1. So 1 is an aperiodic state. Since $\{1, 2, 3\}$ is a communicating class, we then must have that 3 is also an aperiodic state. To see it, note that in fact we cannot go to 3 from 3 in one step. But we can go from 3 to 1, then from 1 to 2, and then from 2 to 3. That takes three steps. But we can also go from 3 to 3 in n steps for any $n > 3$ because once we go from 3 to 1 we can stay there with a positive probability for any number of times and then go from 1 to 2 and from 2 to 3. So the set of integers n for which $p_{33}(n) > 0$ is $\{3, 4, 5, 6, \ldots\}$, and we now see that 3 is an aperiodic state. Similarly, one can verify that 2 is also an aperiodic state.

Remark. It is important to note the subtle point that just because a state i has period d, it does not mean that $p_{ii}(d) > 0$. Suppose, for example, that we can travel from i back to i in steps $6, 9, 12, 15, 18, \ldots$, which have gcd equal to 3, and yet $p_{ii}(3)$ is not greater than zero.

A final definition for now is that of an *absorbing state*. Absorption means that once you have gotten there, you will remain there forever. The formal definition is as follows.

Definition 14.11. A state $i \in S$ is called an *absorbing state* if $p_{ij}(n) = 0$ for all n and for all $j \neq i$. Equivalently, $i \in S$ is an absorbing state if $p_{ii} = 1$; that is, the singleton set $\{i\}$ is a closed class.

Remark. Plainly, if a chain has an absorbing state, then it cannot be regular and cannot even be irreducible. Absorption is fundamentally interesting in gambling scenarios. A gambler may decide to quit the game as soon as his net fortune becomes zero. If we let X_n denote the gambler's net fortune after the nth play, then zero will be an absorbing state for the chain $\{X_n\}$. For chains that have absorbing states, the time taken to get absorbed is considered to be of basic interest.

14.4 ∗ Gambler's Ruin

The problem of the *gambler's ruin* is a classic and entertaining example in the theory of probability. It is an example of a Markov chain with absorbing states. Answers to numerous interesting questions about the problem of the gambler's ruin have been worked out; this is all very classic. We provide an introductory exposition to this interesting problem.

Imagine a gambler who goes to a casino with a dollars in his pocket. He will play a game that pays him one dollar if he wins the game or has him pay one dollar to the house if he loses the game. He will play repeatedly until he either goes broke or his total fortune increases from his initial amount a to a prespecified larger amount $b(b > a)$. The idea is that he is forced to quit if he goes broke and he leaves of his own choice if he wins handsomely and is happy to quit. We can ask numerous interesting questions. But let us just ask what the probability is that he will leave because he goes broke.

This is really a simple random walk problem again. Let the gambler's initial fortune be $S_0 = a$. Then, the gambler's net fortune after n plays is $S_n = S_0 + X_1 + X_2 + \cdots + X_n$, where the X_i are iid with the distribution $P(X_i = 1) = p, P(X_i = -1) = q = 1 - p$. We make the realistic assumption that $p < q \Leftrightarrow p < \frac{1}{2}$, i.e., the game is favorable to the house and unfavorable to the player. Let p_a denote the probability that the player will leave broke if he started with a dollars as his initial fortune. In the following argument, we hold b fixed and consider p_a as a function of a, with a varying between 0 and the fixed $b; 0 \le a \le b$. Note that $p_0 = 1$ and $p_b = 0$. Then, p_a satisfies the recurrence relation

$$p_a = pp_{a+1} + (1-p)p_{a-1}, 1 \le a < b.$$

The argument is that if the player wins the very first time, which would happen with probability p, then he can eventually go broke with probability p_{a+1} because the first win increases his fortune by one dollar from a to $a+1$, but if the player loses the very first time, which would happen with probability $1 - p$, then he can eventually go broke with probability p_{a-1} because the first loss will decrease his fortune by one dollar from a to $a - 1$.

Rewrite the equation above in the form

$$p_{a+1} - p_a = \frac{1-p}{p}(p_a - p_{a-1}).$$

If we iterate this identity, we get

$$p_{a+1} - p_a = \left(\frac{1-p}{p}\right)^a (p_1 - 1);$$

here, we have used the fact that $p_0 = 1$.

Now use this to find an expression for p_{a+1} as follows:

$$\begin{aligned} p_{a+1} - 1 &= [p_{a+1} - p_a] + [p_a - p_{a-1}] + \cdots + [p_1 - p_0] \\ &= (p_1 - 1)\left[\left(\frac{1-p}{p}\right)^a + \left(\frac{1-p}{p}\right)^{a-1} + \cdots + 1\right] \end{aligned}$$

$$= (p_1 - 1)\frac{\left(\frac{1-p}{p}\right)^a - 1}{\frac{1-p}{p} - 1}$$

$$\Rightarrow p_{a+1} = 1 + (p_1 - 1)\frac{\left(\frac{1-p}{p}\right)^a - 1}{\frac{1-p}{p} - 1}.$$

However, we can find p_1 explicitly by using the last equation with the choice $a = b - 1$, which will give

$$0 = p_b = 1 + (p_1 - 1)\frac{\left(\frac{1-p}{p}\right)^{b-1} - 1}{\frac{1-p}{p} - 1}.$$

Substituting the expression we get for p_1 from here into the formula for p_{a+1}, we will have

$$p_{a+1} = \frac{(q/p)^b - (q/p)^{a+1}}{(q/p)^b - 1}.$$

This last formula actually gives an expression for p_x for a general $x \leq b$; we can use it with $x = a$ in order to write the final formula,

$$p_a = \frac{(q/p)^b - (q/p)^a}{(q/p)^b - 1}.$$

Note that this formula does give $p_0 = 1$, $p_b = 0$, and that $\lim_{b\to\infty} p_a = 1$ on using the important fact that $\frac{q}{p} > 1$. The practical meaning of $\lim_{b\to\infty} p_a = 1$ is that if the gambler is targeting too high, then actually he will certainly go broke before he reaches that high target.

To summarize, this is an example of a stationary Markov chain with two distinct absorbing states, and we have worked out here the probability that the chain reaches one absorbing state (the gambler going broke) before it reaches the other absorbing state (the gambler leaving as a winner on his terms).

14.5 ∗ First Passage, Recurrence, and Transience

Recurrence, transience, and first-passage times are fundamental to understanding the long-run behavior of a Markov chain. Recurrence is also linked to the stationary distribution of a chain, one of the most important things to study in analyzing and using a Markov chain.

Definition 14.12. Let $\{X_n\}, n \geq 0$ be a stationary Markov chain. Let D be a given subset of the state space S. Suppose the initial state of the chain is state i. The *first-passage time* to the set D, denoted as T_{iD}, is defined to be the first time that the chain enters the set D; formally,

$$T_{iD} = \inf\{n > 0 : X_n \in D\},$$

with $T_{iD} = \infty$ if $X_n \in D^c$, the complement of D, for every $n > 0$. If D is a singleton set $\{j\}$, then we denote the first-passage time to j as just T_{ij}. If $j = i$, then the first-passage time T_{ii} is just the first time the chain returns to its initial state i. We use the simpler notation T_i to denote T_{ii}.

Example 14.14 (Simple Random Walk). Let $X_i, i \geq 1$ be iid random variables with $P(X_i = \pm 1) = \frac{1}{2}$, and let $S_n = X_0 + \sum_{i=1}^{n} X_i, n \geq 0$, with the understanding that $X_0 = 0$. Then $\{S_n\}, n \geq 0$ is a stationary Markov chain with initial state zero and state space $S = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

A graph of the first 50 steps of a simulated random walk is given in Figure 14.1. By carefully reading the plot, we see that the first passage to zero, the initial state, occurs at $T_0 = 4$. We can also see from the graph that the walk returns to zero a total of nine times within these first 50 steps. The first passage to $j = 5$ occurs at $T_{05} = 9$. The first passage to the set $D = \{\cdots, -9, -6, -3, 3, 6, 9, \ldots\}$ occurs at $T_{0D} = 7$. The walk goes up to a maximum of 6 at the tenth step. So, we can say that $T_{07} > 50$; in fact, we can make a stronger statement about T_{07} by looking at where the walk is at time $n = 50$. The reader is asked to find the best statement we can make about T_{07} based on the graph.

Example 14.15 (Infinite Expected First-Passage Times). Consider the three-state Markov chain with state space $S = \{1, 2, 3\}$ and transition probability matrix

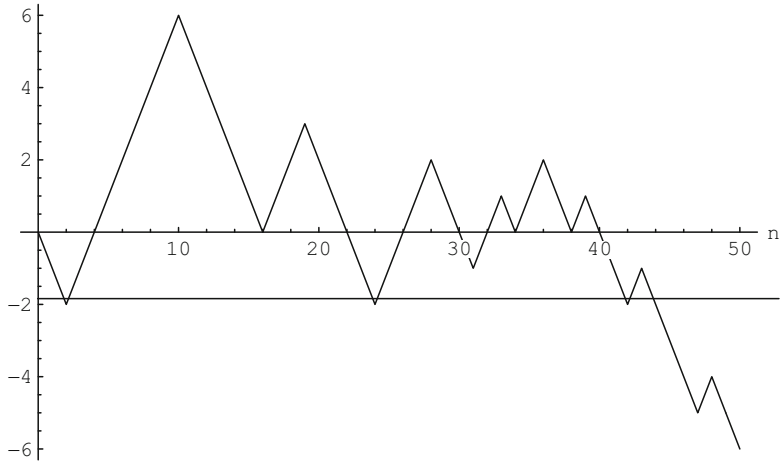


Fig. 14.1 First 50 steps of a simple symmetric random walk

$$P = \begin{pmatrix} x & y & z \\ p & q & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $x + y + z = p + q = 1$.

First consider the recurrence time T_1. Note that for the chain to return at all to state 1 having started at 1, it can never land in state 3 because 3 is an absorbing state. So, if $T_1 = t$, then the chain spends $t - 1$ time instants in state 2 and then returns to 1. In other words, $P(T_1 = 1) = x$, and for $t > 1$, $P(T_1 = t) = yq^{t-2}p$. From here, we can compute $P(T_1 < \infty)$. Indeed,

$$\begin{aligned} P(T_1 < \infty) &= x + \frac{py}{q^2} \sum_{t=2}^{\infty} q^t \\ &= x + \frac{py}{q^2} \frac{q^2}{p} = x + y = 1 - z. \end{aligned}$$

Therefore, $P(T_1 = \infty) = z$, and if $z > 0$, then obviously $E(T_1) = \infty$ because T_1 itself can be ∞ with a positive probability. If $z = 0$, then

$$\begin{aligned} E(T_1) &= x + \frac{py}{q^2} \sum_{t=2}^{\infty} tq^t \\ &= x + \frac{py}{q^2} \frac{2q^2 - q^3}{p^2 q} = \frac{1 + p - x(1 + p^2)}{p(1 - p)}. \end{aligned}$$

We now define the properties of recurrence and transience of a state. At first glance, it would appear that there could be something in between recurrence and transience, but in fact a state is either recurrent or transient. The mathematical meanings of recurrence and transience would really correspond to what their dictionary meanings are. A recurrent state is one that you keep coming back to over and over again with certainty; a transient state is one that you will ultimately leave behind forever with certainty. Below, we are going to use the simpler notation $P_i(A)$ to denote the conditional probability $P(A|X_0 = i)$, where A is a generic event. Here are the formal definitions of recurrence and transience.

Definition 14.13. A state $i \in S$ is called *recurrent* if $P_i(X_n = i \text{ for infinitely many } n \geq 1) = 1$. The state $i \in S$ is called *transient* if $P_i(X_n = i \text{ for infinitely many } n \geq 1) = 0$.

Remark. Note that if a stationary chain returns to its original state i (at least) once with probability 1, then it will also return infinitely often with probability 1. So, we could also think of recurrence and transience of a state in terms of the following questions:

(a) Is $P_i(X_n = i \text{ for some } n \geq 1) = 1$?
(b) Is $P_i(X_n = i \text{ for some } n \geq 1) < 1$?

Here is another way to think about it. Consider our previously defined recurrence time T_i (still with the understanding that the initial state is i). We can think of recurrence in terms of whether $P_i(T_i < \infty) = 1$ or not.

Needless to say, just because $P_i(T_i < \infty) = 1$, it does not follow that its expectation $E_i(T_i) < \infty$. It is a key question in Markov chain theory whether $E_i(T_i) < \infty$ for every state i or not. Not only is it of practical value to compute $E_i(T_i)$, but the finiteness of $E_i(T_i)$ for every state i crucially affects the long-run behavior of the chain. If we want to predict where the chain will be after it has run for a long time, our answers will depend on these expected values $E_i(T_i)$, provided they are all finite. The relationship of $E_i(T_i)$ to the limiting value of $P(X_n = i)$ will be made clear in the next section. Because of the importance of the issue of finiteness of $E_i(T_i)$, the following are important definitions.

Definition 14.14. A state i is called *null recurrent* if $P_i(T_i < \infty) = 1$, but $E_i(T_i) = \infty$. The state i is called *positive recurrent* if $E_i(T_i) < \infty$. The Markov chain $\{X_n\}$ is called positive recurrent if every state i is positive recurrent.

Recurrence and transience can be discussed at various levels of sophistication, and the treatment and ramifications can be confusing, so a preview is going to be useful.

Preview

(a) You can verify recurrence or transience of a given state i by verifying whether $\sum_{i=0}^{n} p_{ii}(n) = \infty$ or $< \infty$.
(b) You can also try to verify directly whether $P_i(T_i < \infty) = 1$ or < 1.
(c) Chains with a finite state space are more easily handled with regard to settling recurrence or transience issues. For finite chains, there must be at least one recurrent state; i.e., not all states can be transient if the chain has a finite state space.
(d) Recurrence is a *class property*; i.e., states within the same communicating class have the same recurrence status. If one of them is recurrent, so are all the others.
(e) In identifying exactly which communicating classes have the recurrence property, you can identify which of the communicating classes are *closed*.
(f) Even if a state i is recurrent, $E_i(T_i)$ can be infinite; i.e., the state i can be *null recurrent*. However, if the state space is finite and if the chain is regular, then you do not have to worry about it. As a matter of fact, for *any set* D, T_{iD} will be finite with probability 1, and even $E_i(T_{iD})$ will be finite. So, for a finite regular chain, you have a very simple recurrence story; every state is not just recurrent but even *positive recurrent*.
(g) For chains with an infinite state space, it is possible that every state is transient, and it is also possible that every state is recurrent or something in between. Whether or not the chain is *irreducible* is going to be a key factor in sorting out the exact recurrence structure.

Some of the major results on recurrence and transience are now given.

Theorem 14.2. *Let $\{X_n\}$ be a stationary Markov chain. If $\sum_{n=0}^{\infty} p_{ii}(n) = \infty$, then i is a recurrent state, and if $\sum_{n=0}^{\infty} p_{ii}(n) < \infty$, then i is a transient state.*

Proof. Introduce the variable $V_i = \sum_{n=0}^{\infty} I_{\{X_n=i\}}$; thus, V_i is the total number of visits of the chain to state i. Also let $p_i = P_i(T_i < \infty)$. By using the Markov property of $\{X_n\}$, it follows that $P_i(V_i > m) = p_i^m$ for any $m \geq 0$. Suppose now that $p_i < 1$. Then, by the tailsum formula for expectations,

$$\begin{aligned} E_i(V_i) &= \sum_{m=0}^{\infty} P_i(V_i > m) \\ &= \sum_{m=0}^{\infty} p_i^m = \frac{1}{1-p_i} < \infty. \end{aligned}$$

But also

$$\begin{aligned} E_i(V_i) &= E_i\left[\sum_{n=0}^{\infty} I_{\{X_n=i\}}\right] \\ &= \sum_{n=0}^{\infty} E[I_{\{X_n=i\}}] = \sum_{n=0}^{\infty} P_i(X_n = i) \\ &= \sum_{n=0}^{\infty} p_{ii}(n). \end{aligned}$$

So, if $p_i < 1$, then we must have $\sum_{n=0}^{\infty} p_{ii}(n) < \infty$, which is the same as saying that if $\sum_{n=0}^{\infty} p_{ii}(n) = \infty$, then p_i must be equal to 1, so i must be a recurrent state.

Suppose, on the other hand, that $p_i = 1$. Then, for any m, $P_i(V_i > m) = 1$, so, with probability 1, $V_i = \infty$. So, $E_i(V_i) = \infty$, which implies that $\sum_{n=0}^{\infty} p_{ii}(n) = E_i(V_i) = \infty$. So, if $p_i = 1$, then $\sum_{n=0}^{\infty} p_{ii}(n)$ must be ∞, which is the same as saying that if $\sum_{n=0}^{\infty} p_{ii}(n) < \infty$, then $p_i < 1$, which would mean that i is a transient state. The next theorem formalizes the intuition that if you keep coming back to some state over and over again and that state communicates with some other state, then you will be visiting that state over and over again as well. That is, recurrence is a class property, and that implies that transience is also a class property.

Theorem 14.3. *Let C be any communicating class of states of a stationary Markov chain $\{X_n\}$. Then, either all states in C are recurrent or all states in C are transient.*

Proof. The theorem will be proved if we can show that if i and j both belong to a common communicating class and i is transient, then j must also be transient. If we can prove this, it follows that if j is recurrent, then i must also be recurrent; otherwise it would be transient, which would make j transient, a contradiction.

So, suppose $i \in C$, and assume that i is transient. By virtue of the transience of i, we know that $\sum_{r=0}^{\infty} p_{ii}(r) < \infty$, so $\sum_{r=R}^{\infty} p_{ii}(r) < \infty$ for any fixed R. This will be useful to us in the proof.

Now consider another state $j \in C$. Because C is a communicating class, there exist k, n such that $p_{ij}(k) > 0$, $p_{ji}(n) > 0$. Take such k, n and hold them fixed.

Now observe that, for any m, we have the inequality

$$\begin{aligned} p_{ii}(k+m+n) &\geq p_{ij}(k)p_{jj}(m)p_{ji}(n) \\ &\Rightarrow p_{jj}(m) \leq \frac{1}{p_{ij}(k)p_{ji}(n)} p_{ii}(k+m+n) \\ &\Rightarrow \sum_{m=0}^{\infty} p_{jj}(m) \leq \frac{1}{p_{ij}(k)p_{ji}(n)} \sum_{m=0}^{\infty} p_{ii}(k+m+n) < \infty \end{aligned}$$

because $p_{ij}(k)$ and $p_{ji}(n)$ are two fixed positive numbers and $\sum_{m=0}^{\infty} p_{ii}(k+m+n) = \sum_{r=k+n}^{\infty} p_{ii}(r) < \infty$. But, if $\sum_{m=0}^{\infty} p_{jj}(m) < \infty$, then we already know that j must be transient, which is what we want to prove.

If a particular communicating class C consists of (only) recurrent states, we will call C a *recurrent class*. The following are two important consequences of the theorem above.

Theorem 14.4.

(a) *Let $\{X_n\}$ be a stationary irreducible Markov chain with a finite state space. Then every state of $\{X_n\}$ must be recurrent.*

(b) *For any stationary Markov chain with a finite state space, a communicating class is recurrent if and only if it is closed.*

Example 14.16 (Various Illustrations). We will revisit some of the chains in our previous examples and examine their recurrence structure.

In the weather pattern example,

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}.$$

If $0 < \alpha < 1$ and also $0 < \beta < 1$, then clearly the chain is irreducible, and it obviously has a finite state space. So, each of the two states is recurrent. If $\alpha = \beta = 1$, then each state is an absorbing state, and clearly $\sum_{n=0}^{\infty} p_{ii}(n) = \infty$ for both $i = 1, 2$. So, each state is recurrent. If $\alpha = \beta = 0$, then the chain evolves either as $121212\ldots$ or $212121\ldots$. Each state is periodic and recurrent.

In the hopping mosquito example,

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & .5 & .5 \\ .5 & 0 & .5 \end{pmatrix}.$$

In this case, some elements of P are zero. However, we have previously seen that every element in P^3 is strictly positive. Hence, the chain is again irreducible. Once again, each of the three states is recurrent.

Next consider the chain with the transition matrix

$$P = \begin{pmatrix} .75 & .25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ .25 & 0 & 0 & .25 & .5 & 0 \\ 0 & 0 & 0 & .75 & .25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

We have previously proved that the communicating classes of this chain are $\{1, 2, 3\}, \{4\}, \{5, 6\}$, of which $\{5, 6\}$ is the only closed class. Therefore, 5 and 6 are the only recurrent states of this chain.

14.6 Long-Run Evolution and Stationary Distributions

A natural human instinct is to want to predict the future. It is not surprising that we often want to know exactly where a Markov chain will be after it has evolved for a fairly long time. Of course, we cannot say with certainty where it will be. But perhaps we can make probabilistic statements. In notation, suppose a stationary Markov chain $\{X_n\}$ started at some initial state $i \in S$. A natural question is, what can we say about $P(X_n = j | X_0 = i)$ for arbitrary $j \in S$ if n is large? Again, a short preview might be useful.

Preview. For chains with a finite state space, the answers are concrete and extremely structured, and furthermore, convergence occurs rapidly. That is, under some reasonable conditions on the chain, regardless of what the initial state i is, $P(X_n = j \mid X_0 = i)$ has a limit π_j and $P(X_n = j \mid X_0 = i) \approx \pi_j$ for quite moderate values of n. In addition, the marginal probabilities $P(X_n = j)$ are also well approximated by the same π_j, and there is an explicit formula for determining the limiting probability π_j for each $j \in S$. Somewhat different versions of these results are often presented in different texts under different sets of conditions on the chain. Our version balances the ease of understanding the results with the applicability of the conditions assumed. But first let us see two illustrative examples.

Example 14.17. Consider first the weather pattern example, and, for concreteness, take the one-step transition probability matrix to be

$$P = \begin{pmatrix} .8 & .2 \\ .2 & .8 \end{pmatrix}.$$

Then, by direct computation,

$$P^{10} = \begin{pmatrix} .50302 & .49698 \\ .49698 & .50302 \end{pmatrix}; \quad P^{15} = \begin{pmatrix} .50024 & .49976 \\ .49976 & .50024 \end{pmatrix};$$

$$P^{20} = \begin{pmatrix} .50018 & .49982 \\ .49982 & .50018 \end{pmatrix}; \quad P^{25} = \begin{pmatrix} .50000 & .50000 \\ .50000 & .50000 \end{pmatrix}.$$

We notice that P^n appears to converge to a limiting matrix, with each row being the same, namely $(.5, .5)$. That is, regardless of the initial state i, $P(X_n = j \mid X_0 = i)$ appears to converge to $\pi_j = .5$. Thus, if indeed $\alpha = \beta = .8$ in the weather pattern example, then in the long run the chances of a dry or wet day would both be just $50 - 50$, and the effect of the weather on the initial day is going to wash out.

On the other hand, consider a chain with the one-step transition matrix

$$P = \begin{pmatrix} x & y & z \\ p & q & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Notice that this chain has an absorbing state; once you are in state 3, you can never leave. To be concrete, take $x = .25, y = .75, p = q = .5$. Then, by direct computation,

$$P^{10} = \begin{pmatrix} .400001 & .599999 & 0 \\ .4 & .6 & 0 \\ 0 & 0 & 1 \end{pmatrix}; P^{20} = \begin{pmatrix} .4 & .6 & 0 \\ .4 & .6 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This time it appears that P^n converges to a limiting matrix whose first two rows are the same but the third row is different. Specifically, the first two rows of P^n seem to be converging to $(.4, .6, 0)$, while the third row is $(0, 0, 1)$, the same as the third row in P itself. Thus, the limiting behavior of $P(X_n = j \mid X_0 = i)$ seems to depend on the initial state i.

The difference between the two chains in this example is that the first chain is regular, while the second chain has an absorbing state and cannot be regular. *Indeed, regularity of the chain is going to have a decisive effect on the limiting behavior of* $P(X_n = j \mid X_0 = i)$. An important theorem is the following.

Theorem 14.5 (Fundamental Theorem for Finite Markov Chains). *Let $\{X_n\}$ be a stationary Markov chain with a finite state space S consisting of t elements. Assume furthermore that $\{X_n\}$ is regular. Then, there exist $\pi_j, j = 1, 2, \ldots, t$ such that:*

(a) *For any initial state i, $P(X_n = j \mid X_0 = i) \to \pi_j, j = 1, 2, \ldots, t$.*
(b) *$\pi_1, \pi_2, \ldots, \pi_t$ are the unique solutions of the system of equations $\pi_j = \sum_{i=1}^{t} \pi_i p_{ij}, j = 1, 2, \ldots, t$, $\sum_{j=1}^{t} \pi_j = 1$, where p_{ij} denotes the (i, j)th element in the one-step transition matrix P. Equivalently, the row vector $\pi = (\pi_1, \pi_2, \ldots, \pi_t)$ is the unique solution of the equations $\pi P = \pi$, $\pi \mathbf{1}' = 1$, where $\mathbf{1}$ is a row vector with each coordinate equal to 1.*
(c) *The chain $\{X_n\}$ is positive recurrent; i.e., for any state i, the mean recurrence time $\mu_i = E_i(T_i) < \infty$, and furthermore $\mu_i = \frac{1}{\pi_i}$.*

The vector $\pi = (\pi_1, \pi_2, \ldots, \pi_t)$ *is called the* stationary distribution *of the regular finite chain* $\{X_n\}$. *It is also sometimes called the* equilibrium distribution *or the* invariant distribution *of the chain. The difference in terminology can be confusing. Suppose now that a stationary chain has a stationary distribution* π. *If we use this* π *as the initial distribution of the chain, then we observe that*

$$P(X_1 = j) = \sum_{k \in S} P(X_1 = j \mid X_0 = k)\pi_k = \pi_j$$

by the fact that π *is a stationary distribution of the chain. Indeed, it now follows easily by induction that* for any n, $P(X_n = j) = \pi_j, j \in S$. Thus, if a chain has a stationary distribution and starts out with that distribution, then at all subsequent times the distribution of the state of the chain remains exactly the same; that is, it is a stationary distribution. This is why a chain that starts out with its stationary distribution is sometimes described to be in steady-state.

We now give a proof of part (a) and part (b) of the fundamental theorem of Markov chains. For this, we will use a famous result in linear algebra, which we state as a lemma.

Lemma (Perron-Frobenius Theorem). *Let* P *be a real* $t \times t$ *square matrix with all elements* p_{ij} *strictly positive. Then:*

(a) P *has a positive real eigenvalue* λ_1 *such that for any other eigenvalue* λ_j *of* $P, |\lambda_j| < \lambda_1, j = 2, \ldots, t$.

(b) λ_1 *satisfies*

$$\min_i \sum_j p_{ij} \leq \lambda_1 \leq \max_i \sum_j p_{ij}.$$

(c) *There exist left and right eigenvectors of* P, *each having only strictly positive elements, corresponding to the eigenvalue* λ_1; *that is, there exist vectors* π, ω, *with both* π *and* ω *having only strictly positive elements, such that* $\pi P = \lambda_1 \pi$; $P\omega = \lambda_1 \omega$.

(d) *The algebraic multiplicity of* λ_1 *is* 1 *and the dimension of the set of both left and right eigenvectors corresponding to* λ_1 *equals* 1.

Proof of fundamental theorem. Because for a transition probability matrix of a Markov chain the row sums are all equal to 1, it follows immediately from the Perron-Frobenius theorem that if every element of P is strictly positive, then $\lambda_1 = 1$ is an eigenvalue of P and that there is a left eigenvector π with only strictly positive elements such that $\pi P = \pi$. We can always normalize π so that its elements add to exactly 1, so the renormalized π is a stationary distribution for the chain by the definition of a stationary distribution. If the chain is regular, then in general we can only assert that every element of P^n is strictly positive for some n. Then the Perron-Frobenius theorem applies to P^n and we have a left eigenvector π satisfying $\pi P^n = \pi$. It can be proved from this that the same vector π satisfies $\pi P = \pi$, so the chain has a stationary distribution. The uniqueness of the stationary distribution is a consequence of part (d) of the Perron-Frobenius theorem.

Coming to part (a), note that it asserts that every row of P^n converges to the vector π; i.e.,

$$P^n \to \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix}.$$

We prove this by the diagonalization argument we previously used in working out a closed-form formula for P^n in the *hopping mosquito example*. Thus, consider the case where the eigenvalues of P are distinct, remembering that one eigenvalue is 1, and the rest less than 1 in absolute value. Let $U^{-1}PU = L = \text{diag}\{1, \lambda_2, \ldots, \lambda_t\}$, where

$$U = \begin{pmatrix} 1 & u_{12} & u_{13} & \cdots \\ 1 & u_{22} & u_{23} & \cdots \\ & \vdots & \vdots & \\ 1 & u_{t2} & u_{t3} & \cdots \end{pmatrix}; \; U^{-1} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_t \\ u^{21} & u^{22} & \cdots & u^{2t} \\ & \vdots & \vdots & \\ u^{t1} & u^{t2} & \cdots & u^{tt} \end{pmatrix}.$$

This implies $P = ULU^{-1} \Rightarrow P^n = UL^nU^{-1}$. Because each λ_j for $j > 1$ satisfies $|\lambda_j| < 1$, we have $|\lambda_j|^n \to 0$ as $n \to \infty$. This fact, together with the explicit forms of U, U^{-1} given immediately above, leads to the result that each row of UL^nU^{-1} converges to the fixed row vector π, which is the statement in part (a).

We assumed that our chain is regular for the fundamental theorem. An exercise asks us to show that regularity is not necessary for the existence of a stationary distribution. Regular chains are of course irreducible. But irreducibility alone is not enough for the existence of a stationary distribution. More will be said on the issue of existence of a stationary distribution a bit later. For finite chains, irreducibility plus aperiodicity is enough for the validity of the fundamental theorem for the simple reason that such chains are regular in the finite case. It is worth mentioning this as a formal result.

Theorem 14.6. *Let $\{X_n\}$ be a stationary Markov chain with a finite state space S. If $\{X_n\}$ is irreducible and aperiodic, then the fundamental theorem holds.*

Example 14.18 (Weather Pattern). Consider the two-state Markov chain with the transition probability matrix

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}.$$

Assume $0 < \alpha, \beta < 1$, so that the chain is regular. The stationary probabilities π_1, π_2 are to be found from the equation

$$\begin{aligned} (\pi_1, \pi_2)P &= (\pi_1, \pi_2) \\ &\Rightarrow \alpha\pi_1 + (1-\beta)\pi_2 = \pi_1; \\ &\Rightarrow (1-\alpha)\pi_1 = (1-\beta)\pi_2 \Rightarrow \pi_2 = \frac{1-\alpha}{1-\beta}\pi_1. \end{aligned}$$

Substituting this into $\pi_1 + \pi_2 = 1$ gives $\pi_1 + \frac{1-\alpha}{1-\beta}\pi_1 = 1$, so $\pi_1 = \frac{1-\beta}{2-\alpha-\beta}$, which then gives $\pi_2 = 1 - \pi_1 = \frac{1-\alpha}{2-\alpha-\beta}$. For example, if $\alpha = \beta = .8$, then we get $\pi_1 = \pi_2 = \frac{1-.8}{2-.8-.8} = .5$, which is the numerical limit we saw in our example by computing P^n explicitly for large n. For general $0 < \alpha, \beta < 1$, each of the states is positive recurrent. For instance, if $\alpha = \beta = .8$, then $E_i(T_i) = \frac{1}{.5} = 2$ for each of $i = 1, 2$.

Example 14.19. With the row vector $\pi = (\pi_1, \pi_2, \ldots, \pi_t)$ denoting the vector of stationary probabilities of a chain, π satisfies the vector equation $\pi P = \pi$, and taking a transpose on both sides, $P'\pi' = \pi'$. That is, the column vector π' is a right eigenvector of P', the transpose of the transition matrix. For example, consider the voting preferences example with

$$P = \begin{pmatrix} .8 & .05 & .15 \\ .03 & .9 & .07 \\ .1 & .1 & .8 \end{pmatrix}.$$

The transpose of P is

$$P' = \begin{pmatrix} .8 & .03 & .1 \\ .05 & .9 & .1 \\ .15 & .07 & .8 \end{pmatrix}.$$

A set of its three eigenvectors is

$$\begin{pmatrix} .38566 \\ .74166 \\ .54883 \end{pmatrix}, \begin{pmatrix} .44769 \\ -.81518 \\ .36749 \end{pmatrix}, \begin{pmatrix} -.56867 \\ -.22308 \\ .79174 \end{pmatrix}.$$

Of these, the last two cannot be the eigenvector we are looking for because they contain negative elements. The first eigenvector contains only nonnegative (actually strictly positive) elements, and when normalized to give elements that add to 1 results in the stationary probability vector $\pi = (.2301, .4425, .3274)$. We could have also obtained it using the method of elimination as in our preceding example, but the eigenvector method is a general clean method and is particularly convenient when the number of states t is not small.

Example 14.20 (Ehrenfest Urn). Consider the symmetric version of the Ehrenfest urn model in which a certain number among m balls are initially in urn I, the rest in urn II, and at each successive time one of the m balls is selected completely at random and transferred to the other urn with probability $\frac{1}{2}$ (and left in the same urn with probability $\frac{1}{2}$). The one-step transition probabilities are $p_{i,i-1} = \frac{i}{2m}$, $p_{i,i+1} = \frac{m-i}{2m}$, $p_{ii} = \frac{1}{2}$.

A stationary distribution π would satisfy the equations

$$\pi_j = \frac{m-j+1}{2m}\pi_{j-1} + \frac{j+1}{2m}\pi_{j+1} + \frac{\pi_j}{2}, 1 \le j \le m-1;\ \pi_0 = \frac{\pi_0}{2} + \frac{\pi_1}{2m};$$
$$\pi_m = \frac{\pi_m}{2} + \frac{\pi_{m-1}}{2m}.$$

These are equivalent to the equations

$$\pi_0 = \frac{\pi_1}{m};\ \pi_m = \frac{\pi_{m-1}}{m};\ \pi_j = \frac{m-j+1}{m}\pi_{j-1} + \frac{j+1}{m}\pi_{j+1}, 1 \le j \le m-1.$$

Starting with π_1, one can solve these equations just by successive substitution, leaving π_0 as an undetermined constant to get $\pi_j = \binom{m}{j}\pi_0$. Now use the fact that $\sum_{j=0}^{m} \pi_j$ must equal 1. This forces $\pi_0 = \frac{1}{2^m}$ and hence $\pi_j = \frac{\binom{m}{j}}{2^m}$. We now realize that these are exactly the probabilities in a binomial distribution with parameters m and $\frac{1}{2}$. That is, in the symmetric Ehrenfest urn problem, there is a stationary distribution and it is the $Bin(m, \frac{1}{2})$ distribution. In particular, after the process has evolved for a long time, we would expect close to half the balls to be in each urn. Each state is positive recurrent, i.e., the chain is sure to return to its original configuration with a finite expected value for the time it takes to return to that configuration. As a specific example, suppose $m = 10$ and that initially there were five balls in each urn. Then, the stationary probability $\pi_5 = \frac{\binom{10}{5}}{2^10} = \frac{63}{256} = .246$, so we can expect that after about four transfers the urns will once again have five balls each.

Example 14.21 (Asymmetric Random Walk). Consider a random walk $\{S_n\}, n \ge 0$ starting at zero, and taking independent steps of length 1 at each time, either to the left or to the right, with the respective probabilities depending on the current position of the walk. Formally, S_n is a Markov chain with initial state zero and with the one-step transition probabilities $p_{i,i+1} = \alpha_i, p_{i,i-1} = \beta_i, \alpha_i + \beta_i = 1$ for any $i \ge 0$. In order to restrict the state space of the chain to just the nonnegative integers $S = \{0, 1, 2, \ldots\}$, we assume that $\alpha_0 = 1$. Thus, if you ever reach zero, then you start over.

If a stationary distribution π exists, by virtue of the matrix equation $\pi = \pi P$, it satisfies the recursion

$$\pi_j = \pi_{j-1}\alpha_{j-1} + \pi_{j+1}\beta_{j+1}$$

with the initial equation

$$\pi_0 = \pi_1\beta_1.$$

This implies, by successive substitution,

$$\pi_1 = \frac{1}{\beta_1}\pi_0 = \frac{\alpha_0}{\beta_1}\pi_0, \pi_2 = \frac{\alpha_0\alpha_1}{\beta_1\beta_2}\pi_0; \cdots,$$

and for a general $j > 1$,

$$\pi_j = \frac{\alpha_0 \alpha_1 \cdots \alpha_{j-1}}{\beta_1 \beta_2 \cdots \beta_j} \pi_0.$$

Since each π_j, $j \geq 0$ is clearly nonnegative, the only issue is whether they constitute a probability distribution; i.e., whether $\pi_0 + \sum_{j=1}^{\infty} \pi_j = 1$. This is equivalent to asking whether $\left(1 + \sum_{j=1}^{\infty} c_j\right) \pi_0 = 1$, where $c_j = \frac{\alpha_0 \alpha_1 \ldots \alpha_{j-1}}{\beta_1 \beta_2 \ldots \beta_j}$. In other words, the chain has a stationary distribution if and only if the infinite series $\sum_{j=1}^{\infty} c_j$ converges to some positive finite number δ, in which case $\pi_0 = \frac{1}{1+\delta}$ and, for $j \geq 1, \pi_j = \frac{c_j}{1+\delta}$.

Consider now the special case where $\alpha_i = \beta_i = \frac{1}{2}$ for all $i \geq 1$. Then, for any $j \geq 1, c_j = \frac{1}{2}$, and hence $\sum_{j=1}^{\infty} c_j$ diverges. Therefore, the case of the symmetric random walk does not possess a stationary distribution, in the sense that no stationary distribution exists that is a valid probability distribution.

The stationary distribution of a Markov chain is not just the limit of the n-step transition probabilities; it also has important interpretations in terms of the marginal distribution of the state of the chain. Suppose the chain has run for a long time and we want to know what the chances are that it is now in some state j. It turns out that the stationary probability π_j approximates that probability, too. The approximations are valid in a fairly strong sense, to be made precise below. Even more, π_j is approximately equal to the fraction of the time *so far* that the chain has spent visiting state j. To describe these results precisely, we need a little notation.

Given a stationary chain $\{X_n\}$, we denote $f_n(j) = P(X_n = j)$. Also let $I_k(j) = I_{\{X_k = j\}}$ and $V_n(j) = \sum_{k=1}^{n} I_k(j)$. Thus, $V_n(j)$ counts the number of times up to time n that the chain has been in state j, and $\delta_n(j) = \frac{V_n(j)}{n}$ measures the fraction of the time up to time n that it has been in state j. Then, the following results hold.

Theorem 14.7 (Weak Ergodic Theorem). *Let $\{X_n\}$ be a regular Markov chain with a finite state space and the stationary distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_t)$. Then:*

(a) *Whatever the initial distribution of the chain, for any $j \in S$, $P(X_n = j) \to \pi_j$ as $n \to \infty$.*
(b) *For any $\epsilon > 0$ and for any $j \in S$, $P(|\delta_n(j) - \pi_j| > \epsilon) \to 0$ as $n \to \infty$.*
(c) *More generally, given any bounded function g, and any $\epsilon > 0$, $P(|\frac{1}{n} \sum_{k=1}^{n} g(X_k) - \sum_{j=1}^{t} g(j)\pi_j| > \epsilon) \to 0$ as $n \to \infty$.*

Remark. See Norris (1997) for a proof of this theorem. The theorem provides a basis for estimating the stationary probabilities of a chain by following its trajectory for a long time. Part (c) of the theorem says that time averages of a general bounded function will ultimately converge to the state-space average of the function with respect to the stationary distribution. In fact, a stronger convergence result than the one we state here holds and is commonly called the *ergodic theorem for stationary Markov chains*; see Brémaud (1999) or Norris (1997).

14.7 Synopsis

(a) The one-step transition probabilities of a stationary Markov chain with state space S are $p_{ij} = P(X_{n+1} = j \,|X_n = i), i, j \in S$. The n-step transition probabilities are $p_{ij}^{(n)} = P(X_{m+n} = j \,|X_m = i)$.

(b) The Chapman-Kolmogorov equation says that

$$p_{ij}(m+n) = \sum_{k \in S} p_{ik}(m) p_{kj}(n)$$

for all $m, n \geq 1$. In matrix notation, $P^{(n)} = P^n$, where $P^{(n)}$ is the n-step transition probability matrix and P is the one-step transition probability matrix.

(c) The (row) vector λ_n of the probabilities $P(X_n = i), i \in S$, satisfies

$$\lambda_n = \lambda P^n,$$

where λ is the vector of the initial probabilities $P(X_0 = i), i \in S$.

(d) A specific state i is recurrent if and only if $\sum_{n=1}^{\infty} p_{ii}(n) = \infty$.

(e) Recurrence is a class property. Either every state in a communicating class is recurrent or it is transient.

(f) For Markov chains with a finite state space, at least one state must be recurrent. If the chain is also regular, then every state is recurrent and even positive recurrent.

(g) For Markov chains with a finite state space, every state is recurrent if the chain is just irreducible. However, irreducibility alone does not imply that every state is positive recurrent.

(h) Finite regular chains admit a stationary distribution π, which can be found by solving the system of equations

$$\pi_j = \sum_{i \in S} \pi_i p_{ij}, \sum_{j \in S} \pi_j = 1.$$

(i) For finite regular chains, both $P(X_n = j)$ and $P(X_n = j \,|X_0 = i)$ converge to the stationary probability π_j, and this is true whatever the initial state i. Moreover, the mean first-passage time $E_i(T_i) = \frac{1}{\pi_i}$ for every i.

14.8 Exercises

Exercise 14.1. A particular machine is either in working order or broken on any particular day. If it is in working order on some day, it remains so the next day with probability .7, while if it is broken on some day, it stays broken the next day with probability .2.

(a) If it is in working order on Monday, what is the probability that it is in working order on Saturday?
(b) If it is in working order on Monday, what is the probability that it remains in working order all the way through Saturday?

Exercise 14.2. Consider the voting preferences example in the text with the transition probability matrix

$$P = \begin{pmatrix} .8 & .05 & .15 \\ .03 & .9 & .07 \\ .1 & .1 & .8 \end{pmatrix}.$$

Suppose a family consists of the two parents and a son. The three follow the same Markov chain described above in deciding their votes. Assume that the family members act independently and that in this election the father voted Conservative, the mother voted Labor, and the son voted Independent.

(a) Find the probability that they will all vote the same parties in the next election as they did in this election.
(b) * Find the probability that, as a whole, the family will split their votes among the three parties, one member for each party, in the next election.

Exercise 14.3. Suppose $\{X_n\}$ is a stationary Markov chain. Prove that for all n and all $x_i, i = 0, 1, \ldots, n+2$, $P(X_{n+2} = x_{n+2}, X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+2} = x_{n+2}, X_{n+1} = x_{n+1} | X_n = x_n)$.

Exercise 14.4. *(What the Markov Property Does Not Mean). Give an example of a stationary Markov chain with a small number of states such that $P(X_{n+1} = x_{n+1} | X_n \leq x_n, X_{n-1} \leq x_{n-1}, \ldots, X_0 \leq x_0) = P(X_{n+1} = x_{n+1} | X_n \leq x_n)$ is not true for arbitrary $x_0, x_1, \ldots, x_{n+1}$.

Exercise 14.5 (Ehrenfest Urn). Consider the Ehrenfest urn model when there are only two balls to distribute.

(a) Write the transition probability matrix P.
(b) Calculate P^2, P^3.
(c) Find general formulas for P^{2k}, P^{2k+1}.

Exercise 14.6. * (The Cat and Mouse Chain). In one of two adjacent rooms, say room 1, there is a cat, and in the other one, room 2, there is a mouse. There is a small hole in the wall through which the mouse can travel between the rooms, and there is a larger hole through which the cat can travel between the rooms. Each minute, the cat and the mouse decide the room they want to be in by following a stationary Markov chain with the transition probability matrices

$$P_1 = \begin{pmatrix} .5 & .5 \\ .5 & .5 \end{pmatrix}; \; P_2 = \begin{pmatrix} .1 & .9 \\ .6 & .4 \end{pmatrix}.$$

At time n, let X_n be the room in which the cat is and Y_n the room in which the mouse is. Assume that the chains $\{X_n\}$ and $\{Y_n\}$ are independent.

(a) Write the transition matrix for the chain $Z_n = (X_n, Y_n)$.
(b) Let $p_n = P(X_n = Y_n)$. Compute p_n for $n = 1, 2, 3, 4, 5$, taking the initial time to be $n = 0$.
(c) The very first time that they end up in the same room, the cat will eat the mouse. Let q_n be the probability that the cat eats the mouse at time n. Compute q_n for $n = 1, 2, 3$.

Exercise 14.7 (Diagonalization in the Two-State Case). Consider a two-state stationary chain with the transition probability matrix

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}.$$

(a) Find the eigenvalues of P. When are they distinct?
(b) Diagonalize P when the eigenvalues are distinct.
(c) Find a general formula for $p_{11}(n)$.

Exercise 14.8. A flea is initially located on the top face of a cube that has six faces, top and bottom, left and right, and front and back. Every minute it moves from its current location to one of the other five faces, chosen at random.

(a) Find the probability that after four moves it is back to the top face.
(b) Find the probability that after n moves it is on the top face; repeat for the bottom face.
(c) * Find the probability that the next five moves are distinct. This is the same as the probability that the first six locations of the flea are the six faces of the cube, each location being chosen exactly once.

Exercise 14.9 (Subsequences of Markov Chains). Suppose $\{X_n\}$ is a stationary Markov chain. Let $Y_n = X_{2n}$. Prove or disprove that $\{Y_n\}$ is a stationary Markov chain. How about $\{X_{3n}\}$? $\{X_{kn}\}$ for a general k?

Exercise 14.10. Let $\{X_n\}$ be a three-state stationary Markov chain with the transition probability matrix

$$P = \begin{pmatrix} 0 & x & 1-x \\ y & 1-y & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Define a function g as $g(1) = 1, g(2) = g(3) = 2$, and let $Y_n = g(X_n)$. Is $\{Y_n\}$ a stationary Markov chain?

Give an example of a function g such that $g(X_n)$ is not a Markov chain.

Exercise 14.11 (An IID Sequence). Let $X_i, i \geq 1$ be iid Poisson random variables with some common mean λ. Prove or disprove that $\{X_n\}$ is a stationary Markov chain. If it is, describe the transition probability matrix.

How important is the Poisson assumption? What happens if $X_i, i \geq 1$ are independent but not iid?

Exercise 14.12. Let $\{X_n\}$ be a stationary Markov chain with transition matrix P and g a one-to-one function. Define $Y_n = g(X_n)$. Prove that $\{Y_n\}$ is a Markov chain, and characterize as well as you can the transition probability matrix of $\{Y_n\}$.

Exercise 14.13. * **(Loop Chains).** Suppose $\{X_n\}$ is a stationary Markov chain with state space S and transition probability matrix P.

(a) Let $Y_n = (X_n, X_{n+1})$. Show that Y_n is also a stationary Markov chain.
(b) Find the transition probability matrix of Y_n.
(c) How about $Y_n = (X_n, X_{n+1}, X_{n+2})$? Is this also a stationary Markov chain?
(d) How about $Y_n = (X_n, X_{n+1}, \ldots, X_{n+d})$ for a general $d \geq 1$?

Exercise 14.14 (Dice Experiments). Consider the experiment of rolling a fair die repeatedly. Define

(a) X_n = the number of sixes obtained up to the nth roll;
(b) X_n = the number of rolls, at time n, that a six has not been obtained since the last six.

Prove or disprove that each $\{X_n\}$ is a Markov chain, and if they are, obtain the transition probability matrices.

Exercise 14.15. Suppose $\{X_n\}$ is a regular stationary Markov chain with transition probability matrix P. Prove that there exists $m \geq 1$ such that every element in P^n is strictly positive for all $n \geq m$.

Exercise 14.16 (Communicating Classes). Consider a finite-state stationary Markov chain with the transition matrix

$$P = \begin{pmatrix} 0 & .5 & 0 & .5 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ .5 & 0 & 0 & 0 & .5 \\ 0 & .25 & .25 & .25 & .25 \\ .5 & 0 & 0 & 0 & .5 \end{pmatrix}.$$

(a) Identify the communicating classes of this chain.
(b) Identify those classes that are closed.

Exercise 14.17. * **(Periodicity and Simple Random Walk).** Consider the Markov chain corresponding to the simple random walk with general step probabilities $p, q, p + q = 1$.

(a) Identify the periodic states of the chain and the periods.
(b) Find the communicating classes.
(c) Are there any communicating classes that are not closed? If there are, identify them. If not, prove that there are no communicating classes that are not closed.

Exercise 14.18. *(**Gambler's Ruin).** Consider the Markov chain corresponding to the problem of the gambler's ruin with initial fortune a and absorbing states at 0 and b.

(a) Identify the periodic states of the chain and the periods.
(b) Find the communicating classes.
(c) Are there any communicating classes that are not closed? If there are, identify them.

Exercise 14.19. Prove that a stationary Markov chain with a finite state space has at least one closed communicating class.

Exercise 14.20. * **(Chain with No Closed Classes).** Give an explicit example of a stationary Markov chain with no closed communicating classes.

Exercise 14.21 (Skills Exercise). Consider the stationary Markov chains corresponding to the following transition probability matrices:

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & 0 & \frac{1}{3} \end{pmatrix}; \; P = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{3}{4} & \frac{1}{8} & \frac{1}{8} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

(a) Are the chains irreducible?
(b) Are the chains regular?
(c) For each chain, find the communicating classes.
(d) Are there any periodic states? If there are, identify them.
(e) Do both chains have stationary distributions? Is there anything special about the stationary distribution of either chain? If so, what is special?

Exercise 14.22. * **(Recurrent States).** Let $Z_i, i \geq 1$ be iid Poisson random variables with mean 1. For each of the sequences

$$X_n = \sum_{i=1}^{n} Z_i, X_n = \max\{Z_1, \ldots, Z_n\}, X_n = \min\{Z_1, \ldots, Z_n\} :$$

(a) Prove or disprove that $\{X_n\}$ is a stationary Markov chain.
(b) If it is, write the transition probability matrix.
(c) Find the recurrent and the transient states of the chain.

Exercise 14.23 (Irreducibility and Aperiodicity). For stationary Markov chains with the following transition probability matrices, decide whether the chains are irreducible and aperiodic.

$$P = \begin{pmatrix} 0 & 1 \\ p & 1-p \end{pmatrix};\ P = \begin{pmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix};\ P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ p & 1-p & 0 \end{pmatrix}.$$

Exercise 14.24 (Irreducibility of the Machine Maintenance Chain). Consider the machine maintenance example given in the text. Prove that the chain is irreducible if and only if $p_0 > 0$ and $p_0 + p_1 < 1$. Do some numerical computing that reinforces this theoretical result.

Exercise 14.25. * (Irreducibility of Loop Chains). Let $\{X_n\}$ be a stationary Markov chain, and consider the *loop chain* defined by $Y_n = (X_n, X_{n+1})$. Prove that if $\{X_n\}$ is irreducible, then so is $\{Y_n\}$.

Do you think this generalizes to $Y_n = (X_n, X_{n+1}, \ldots, X_{n+d})$ for general $d \geq 1$?

Exercise 14.26. * (Functions of a Markov Chain). Consider the Markov chain $\{X_n\}$ corresponding to the simple random walk with general step probabilities $p, q, p + q = 1$.

(a) If $f(.)$ is any strictly monotone function defined on the set of integers, show that $\{f(X_n)\}$ is a stationary Markov chain.
(b) Is this true for a general chain $\{Y_n\}$? Prove it or give a counterexample.
(c) Show that $\{|X_n|\}$ is a stationary Markov chain, although $x \to |x|$ is not a strictly monotone function.
(d) Give an example of a function f such that $\{f(X_n)\}$ is not a Markov chain.

Exercise 14.27 (A Nonregular Chain with a Stationary Distribution). Consider a two-state stationary Markov chain with the transition probability matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

(a) Show that the chain is not regular.
(b) Prove that, nevertheless, the chain has a unique stationary distribution, and identify it.

Exercise 14.28. * (Immigration-Death Model). At time $n, n \geq 1, U_n$ particles enter into a box. $U_1, U_2, \ldots$ are assumed to be iid with some common distribution F. The lifetimes of all the particles are assumed to be iid with common distribution G. Initially, there are no particles in the box. Let X_n be the number of particles in the box just after time n.

(a) Take F to be a Poisson distribution with mean 2, and G to be geometric with parameter $\frac{1}{2}$. That is, G has the mass function $\frac{1}{2^x}, x = 1, 2, \ldots$. Write the transition probability matrix for $\{X_n\}$.
(b) Does $\{X_n\}$ have a stationary distribution? If it does, find it.

Exercise 14.29. * **(Betting on the Basis of a Stationary Distribution).** A particular stock either retains the value that it had at the close of the previous day, gains a point, or loses a point, the respective states being denoted as $1, 2, 3$. Suppose X_n is the state of the stock on the nth day; thus, X_n takes the values $1, 2$, or 3. Assume that $\{X_n\}$ forms a stationary Markov chain with the transition probability matrix

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}.$$

A friend offers you the following bet: if the stock goes up tomorrow, he pays you 15 dollars, while if it goes down, you pay him 10 dollars. If it remains the same as where it closed today, a fair coin will be tossed and he will pay you 10 dollars if a head shows up and you will pay him 15 dollars if a tail shows up. Will you accept this bet? Justify your answer with appropriate calculations.

Exercise 14.30. * **(Absent-Minded Professor).** A mathematics professor has two umbrellas, both of which were originally at home. The professor walks back and forth between his home and office, and if it is raining when he starts a journey, he carries an umbrella with him unless both his umbrellas are at the other location. If it is clear when he starts a journey, he does not take an umbrella with him. We assume that at the time he starts a journey, it rains with probability p and the states of weather are mutually independent.

(a) Find the limiting proportion of journeys in which the professor gets wet.
(b) What if the professor had three umbrellas to begin with, all of which were originally at home?
(c) Is the limiting proportion affected by how many umbrellas were originally at home?

Exercise 14.31. * **(Wheel of Fortune).** A pointed arrow is set on a circular wheel marked with m positions labeled as $0, 1, \ldots, m-1$. The hostess turns the wheel during each game so that the arrow either remains where it was before the wheel was turned or moves to a different position. Let X_n denote the position of the arrow after n turns.

(a) Suppose that at any turn the arrow has an equal probability $\frac{1}{m}$ of ending up at any of the m positions. Does $\{X_n\}$ have a stationary distribution? If it does, identify it.
(b) Suppose that at each turn the hostess keeps the arrow where it was or moves it one position clockwise or one position counterclockwise, each with an equal probability $\frac{1}{3}$. Does $\{X_n\}$ have a stationary distribution? If it does, identify it.
(c) Suppose again that at each turn the hostess keeps the arrow where it was or moves it one position clockwise or one position counterclockwise, but now with

respective probabilities $\alpha, \beta, \gamma, \alpha + \beta + \gamma = 1$. Does $\{X_n\}$ have a stationary distribution? If it does, identify it.

Exercise 14.32 (Wheel of Fortune Continued). Consider again the Markov chains corresponding to the wheel of fortune. Prove or disprove that they are irreducible and aperiodic.

Exercise 14.33. * (Stationary Distribution in Ehrenfest Model). Consider the general Ehrenfest chain defined in the text, with m balls, and transfer probabilities $\alpha, \beta, 0 < \alpha, \beta < 1$. Identify a stationary distribution if it exists.

Exercise 14.34. * (Time Until Break away). Consider a general stationary Markov chain $\{X_n\}$, and let $T = \min\{n \geq 1 : X_n \neq X_0\}$.

(a) Can T be equal to ∞ with a positive probability?
(b) Give a simple necessary and sufficient condition for $P(T < \infty) = 1$.
(c) For the *weather pattern, Ehrenfest urn, and the cat and mouse* chain, compute $E(T \mid X_0 = i)$ for a general i in the corresponding state space S.

Exercise 14.35. ** (Constructing Examples). Construct an example of each of the following phenomena:

(a) a Markov chain with only absorbing states;
(b) a Markov chain that is irreducible but not regular;
(c) a Markov chain that is irreducible but not aperiodic;
(d) a Markov chain on an infinite state space that is irreducible and aperiodic, but not regular;
(e) a Markov chain in which there is at least one null recurrent state;
(f) a Markov chain on an infinite state space such that every state is transient;
(g) a Markov chain such that each first-passage time T_{ij} has all moments finite;
(h) a Markov chain without a proper stationary distribution;
(i) independent irreducible chains $\{X_n\}, \{Y_n\}$, such that $Z_n = (X_n, Y_n)$ is not irreducible;
(j) Markov chains $\{X_n\}, \{Y_n\}$ such that $Z_n = (X_n, Y_n)$ is not a Markov chain.

Exercise 14.36. * (Reversibility of a Chain). A stationary chain $\{X_n\}$ with transition probabilities p_{ij} is called reversible if there is a function $m(x)$ such that $p_{ij}m(i) = p_{ji}m(j)$ for all $i, j \in S$. Give a simple sufficient condition in terms of the function m that ensures that a reversible chain has a proper stationary distribution. Then, identify the stationary distribution.

Exercise 14.37. Give a physical interpretation for the property of reversibility of a Markov chain.

Exercise 14.38 (Reversibility). Give examples of a Markov chain that is reversible and one that is not.

Exercise 14.39 (Use Your Computer: Cat and Mouse). Take the *cat and mouse chain* and simulate it to find how long it takes for the cat and mouse to end up in the same room. Repeat the simulation and estimate the expected time until the cat and mouse end up in the same room. Vary the transition matrix and examine how the expected value changes.

Exercise 14.40 (Use Your Computer: Ehrenfest Urn). Take the symmetric *Ehrenfest chain*; that is, take $\alpha = \beta = .5$. Put all the m balls in the second urn to begin with. Simulate the chain and find how long it takes for the urns to have an equal number of balls for the first time. Repeat the simulation and estimate the expected time until both urns have an equal number of balls. Take $m = 10, 20$.

Exercise 14.41 (Use Your Computer: Gambler's Ruin). Take the *gambler's ruin problem* with $p = .4, .49$. Simulate the chain using $a = 10, b = 25$ and find the proportion of times that the gambler goes broke by repeating the simulation. Compare your empirical proportion with the exact theoretical value of the probability that the gambler will go broke.

References

Bhattacharya, R.N. and Waymire, E. (2009). *Stochastic Processes with Applications*, SIAM, Philadelphia.

Brémaud, P. (1999). *Markov Chains, Gibbs Fields, Monte Carlo, and Queues*, Springer, New York.

Diaconis, P. (1988). *Group Representations in Probability and Statistics*, IMS Lecture Notes and Monographs Series, Hayward, CA.

Feller, W. (1968). *An Introduction to Probability Theory, with Applications*, Wiley, New York.

Freedman, D. (1975). *Markov Chains*, Holden-Day, San Francisco.

Isaacson, D. and Madsen, R. (1976). *Markov Chains, Theory and Applications*, Wiley, New York.

Kemperman, J. (1950). *The General One-Dimensional Random Walk with Absorbing Barriers*, Geboren Te, Amsterdam.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*, Springer, New York.

Norris, J. (1997). *Markov Chains*, Cambridge University Press, Cambridge.

Seneta, E. (1981). *Nonnegative Matrices and Markov Chains*, Springer-Verlag, New York.

Stirzaker, D. (1994). *Elementary Probability*, Cambridge University Press, Cambridge.

Chapter 15
Urn Models in Physics and Genetics

Urn models conceptualize general allocation problems in which we distribute, withdraw, and redistribute certain objects or units into a specified number of categories. We think of the categories as urns and the objects as balls. Depending on the specific urn model, the balls may be of different colors and distinguishable or indistinguishable. Urn models are special because they can be successfully used to model real phenomena in diverse areas such as physics, ecology, genetics, economics, clinical trials, modeling of networks, and many others. The aim is to understand the evolution of the content of the urns as distribution and redistribution according to some prespecified scheme progresses. There are many urn models in probability, and the allocation scheme depends on exactly which model one wishes to study. We introduce and provide basic information on some key urn models in this chapter. Classic references are Feller (1968) and Johnson and Kotz (1977). Bernoulli (1713) and Whitworth (1901) are two historically important monographs on urn models. More recent references include Gani (2004), Lange (2003), and Ivchenko and Medvedev (1997). Other specific references are given in the various sections of this chapter.

It turns out that the study of most of the common urn models in physics and genetics involves a special sequence of numbers known as the *Stirling numbers*. We start with a brief introduction to the Stirling numbers and some of their basic properties.

15.1 Stirling Numbers and Their Basic Properties

Stirling numbers have a variety of combinatorial definitions. For our purpose, however, an algebraic definition seems proper. We will mention the combinatorial connections also.

Definition 15.1. The *Stirling numbers of the first kind* are the unique numbers $s(n,k), n \geq 1, 1 \leq k \leq n$, such that $x_{(n)} = x(x-1)\cdots(x-n+1) = \sum_{k=1}^{n} s(n,k)x^k$.

A. DasGupta, *Fundamentals of Probability: A First Course*, Springer Texts in Statistics,
DOI 10.1007/978-1-4419-5780-1_15,

Definition 15.2. The *Stirling numbers of the second kind* are the unique numbers $S(n,k), n \geq 1, 0 \leq k \leq n$, such that $x^n = \sum_{k=0}^{n} S(n,k)x_{(k)}$, where $x_{(0)} = 1$.

Here is an elementary example.

Example 15.1. By simple expansion, $x_{(3)} = x(x-1)(x-2) = x^3 - 3x^2 + 2x$. Therefore, by its definition, $s(3,1) = 2, s(3,2) = -3, s(3,3) = 1$.

On the other hand, $x^3 = x + 3x(x-1) + x(x-1)(x-2)$ by direct verification. Therefore, by its definition, $S(3,1) = 1, S(3,2) = 3, S(3,3) = 1$.

Of course, it is impractical to find the coefficients for large n and k by such direct verification. Fortunately, that is not necessary. One can use recursion relations to generate the coefficients sequentially, or write formulas for them, although the formulas are not very simple. The result below describes the recursions and the formulas; standard texts on enumerative combinatorics can be consulted for these results. One reference is Tomescu (1985).

Theorem 15.1.

(a) $s(n+1,k) = s(n,k-1) - ns(n,k), n \geq k \geq 1$;

(b) $s(n,0) = 0\ \forall n \geq 1; \sum_{k=1}^{n} s(n,k) = 0$;

(c) $s(n,1) = (-1)^{n-1}(n-1)!; s(n,n-1) = -\binom{n}{2}; s(n,n) = 1$;

(d) $S(n+1,k) = kS(n,k) + S(n,k-1), n \geq k \geq 1$;

(e) $S(n,k) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^n$;

(f) $S(n,0) = 0\ \forall n \geq 1; S(n,1) = S(n,n) = 1; S(n,n-1) = \binom{n}{2}$;

(g) $\sum_{k=m}^{n} S(n,k)s(k,m) = I_{\{m=n\}}$.

We will omit the proof of this theorem, as it is stated principally for ease of reference and evaluation of the coefficients in examples and exercises.

Numerical values of the Stirling numbers are of course useful whenever they arise in a specific problem. A table of Stirling numbers is provided below for quick reference.

Stirling Numbers of the First Kind

n	k									
	1	2	3	4	5	6	7	8	9	10
1	1									
2	−1	1								
3	2	−3	1							
4	−6	11	−6	1						
5	24	−50	35	−10	1					
6	−120	274	−225	85	−15	1				
7	720	−1764	1624	−735	175	−21	1			
8	−5040	13068	−13132	6769	−1960	322	−28	1		
9	40320	−109584	118124	−67284	22449	−4536	546	−36	1	
10	−362880	1026576	−1172700	723680	−269325	63273	−9450	870	−45	1

Stirling Numbers of the Second Kind

n	k									
	1	2	3	4	5	6	7	8	9	10
1	1									
2	1	1								
3	1	3	1							
4	1	7	6	1						
5	1	15	25	10	1					
6	11	31	90	65	15	1				
7	1	63	301	350	140	21	1			
8	1	127	966	1701	1050	266	28	1		
9	1	255	3025	7770	6951	2646	462	36	1	
10	1	511	9330	34105	42525	22827	5880	750	45	1

The Stirling numbers have interesting combinatorial interpretations. In turn, these combinatorial interpretations are sometimes useful in expressing otherwise complicated probabilities in terms of Stirling numbers. Here is a very important result that will be directly useful to us.

Theorem 15.2. *The Stirling number $S(n,k)$ of the second kind equals the total number of ways in which n distinct objects can be partitioned into k disjoint and nonempty subsets.*

Example 15.2 (Missing Faces in Die Rolls). Suppose a fair die is rolled n times. Let X be the number of faces of the die that are still missing after the n rolls. Note that $X = k$ if and only if Y, the number of faces that *have* shown up, is $6-k$. But, for $6-k$ faces to show up, the n rolls would each be assigned to some $6-k$ specific faces; i.e., the set of all the n rolls would be partitioned into $6-k$ subsets of rolls, one subset corresponding to all the rolls where a particular face occurred. Therefore,

$$P(X=k) = P(Y=6-k) = \binom{6}{6-k} \frac{(6-k)^n}{6^n}(6-k)!\frac{S(n,6-k)}{(6-k)^n}$$
$$= \frac{\binom{6}{k}(6-k)!S(n,6-k)}{6^n}.$$

15.2 Urn Models in Quantum Mechanics

Classical mechanics does not succeed in explaining the physical workings of systems at the subatomic level. For example, classical mechanics would predict that electrons would leave their orbits and collide with the nucleus. But, in reality, we see quite the contrary. We see that electrons maintain an energy state to keep them in

a stable orbit around the nucleus. If energy states are quantized, which is a name for discretization, then the behavior of particles can be understood in terms of suitable urn models. We will consider the different states of energy to be our urns and the particles to be the balls. Physics dictates exactly which urn model applies to a particular kind of particle. Particles can be of many different types, for example photons, electrons, Fermions, Bosons, and so forth. Three celebrated urn models with origins in quantum mechanics are the *Maxwell-Boltzmann* (M-B), *Bose-Einstein* (B-E), and *Fermi-Dirac* (F-D) models, also called the *M-B, B-E*, and *F-D statistics*. We now introduce these three models and describe some of their elementary properties.

Throughout this section, N will denote the total number of urns (i.e., energy states for the physicist) and n will denote the number of balls (i.e., the total number of particles of a particular type under consideration). Each particle resides in some energy state at a given time. We want to understand the conglomeration of particles by using an appropriate urn model. We will let X_i denote the number of balls in the ith urn and M_k denote the number of urns with k balls. In particular, M_0 is the number of empty urns. The fraction of urns among all the urns that have k balls is the ratio $r_k = \frac{M_k}{N}$.

In the M-B model, each of the n particles can be in any of the N energy states independently of each other and with an equal probability of being in any state. Conceptually, this is the simplest of the three models. If we now focus our attention on one specific state, say state i, then the number of particles out of n that are in this specific energy state has the $Bin(n, \frac{1}{N})$ distribution. Let X_i denote the number of particles in state i. Then, by the familiar binomial distribution formula,

$$P(X_i = k) = \binom{n}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{(n-k)}.$$

Writing I_i as the indicator of the event that the ith urn has k balls, we have $M_k = I_1 + \cdots + I_N$. Therefore,

$$E(r_k) = E\left(\frac{M_k}{N}\right) = \frac{1}{N}\sum_{i=1}^{N} E(I_i) = \frac{1}{N} N E(I_1) = P(X_1 = k)$$

$$= \binom{n}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{(n-k)}.$$

Of particular interest is M_0, the number of empty urns. Primary interest lies in the distribution and the expected value of M_0. It turns out that the Stirling numbers introduced in the previous section are now going to become directly useful in finding the distribution of M_0 in the Maxwell-Boltzmann model. Suppose we want to find $P(M_0 = m)$ for some specified $m, m \leq N-1$. The event $\{M_0 = m\}$ happens if and only if a subset of m urns are empty and the n balls are divided among the remaining $N-m$ urns, with the restriction that none of these $N-m$ urns can be empty. This can happen in $\binom{N}{m} S(n, N-m)(N-m)!$ ways, where $S(n,k)$ is the notation for Stirling

numbers of the second kind. The $(N-m)!$ factor is needed because the Stirling number $S(n, N-m)$ does not account for the different configurations of the $N-m$ distinguishable urns that contain the n balls. Hence, we now have

$$P(M_0 = m) = \frac{\binom{N}{m} S(n, N-m)(N-m)!}{N^n}.$$

No further simplification of this is possible for general n, N, and m. However, the formula is useful for numerical computation of the distribution of M_0 and its expected value, etc. We will see such a numerical example below.

In the *Bose-Einstein* model, it does not matter exactly which particles are in which energy states, but all that matters is how many particles are in each of the different energy states. In the terminology of urns and balls, the urns are distinguishable but the balls are not. This changes the total number of possible distributions of the balls in the urns.

A clever geometric argument gives us the total number of ways to distribute the n balls into the N urns. Let us take a specific example to understand this geometric argument. Suppose $N = 3$ and $n = 5$. Line up the five balls as points in a straight line. Now add to these five points two vertical lines. That gives us a total of seven objects. Arrange these seven objects in a line. For example, one possible arrangement is three points starting from the left, then two consecutive vertical lines, and then two more points. This will correspond to there being three balls in the first urn, *none* in the second urn, two in the third urn, etc. We can arrange the seven objects in 7! ways. But, of course, the vertical lines are just vertical lines and not to be distinguished, and the balls are not to be distinguished by the definition of the Bose-Einstein model. Therefore, the total number of ways to distribute the balls into the three urns is $\frac{7!}{2!5!} = \binom{7}{5}$. Exactly the same argument gives us the formula that the total number of ways to distribute n balls into N urns in the Bose-Einstein model is $\binom{n+N-1}{n}$.

It is still assumed that each of these possible configurations has an equal probability. In other words, we assume that the sample points are equally likely, and each sample point ω has the probability

$$P(\omega) = \frac{1}{\binom{n+N-1}{n}}.$$

On replacing N by $N-1$ and n by $n-k$, we find that

$$P(X_i = k) = \frac{\binom{n-k+N-2}{n-k}}{\binom{n+N-1}{n}},$$

which is therefore also equal to $E(r_k)$.

Coming to M_0, the number of empty urns, by using the same geometric argument as the one above, one has the formula that the number of ways to distribute n indistinguishable balls into m distinguishable urns so that each of the m urns is nonempty is $\binom{n-1}{m-1}$. Therefore,

$$P(M_0 = m) = \frac{\binom{N}{m}\binom{n-1}{N-m-1}}{\binom{N+n-1}{n}}$$
$$= \frac{\binom{N}{m}\binom{n-1}{N-m-1}}{\binom{N+n-1}{N-1}}.$$

In particular,

$$P(M_0 = 0) = \frac{\binom{n-1}{N-1}}{\binom{N+n-1}{N-1}}.$$

The *Fermi-Dirac* model imposes the additional restriction that there can be at most one particle in any particular energy state. That is, an urn can either remain empty or contain only one ball. Note that this automatically forces the number of balls to be no larger than the number of urns; i.e., we must have $n \leq N$. The number of possible sample points in the F-D model is simply the number of ways that we can pick n urns from the N urns that will not be empty. This can be done in $\binom{N}{n}$ ways. If these are assumed to be equally likely, then each sample point ω has the probability

$$P(\omega) = \frac{1}{\binom{N}{n}}.$$

Since urns can only contain at most one ball in the F-D model, only the value $P(X_i = 1)$ is of interest, and this equals, under the equally likely assumption,

$$P(X_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

The distribution of the number of empty urns in the F-D model is saved for the chapter exercises.

This finishes the most elementary description of the M-B, B-E, and F-D models. More advanced properties of these three urn models will be presented in a later section.

Example 15.3 (Empty Urns in the Bose-Einstein Scheme). One useful index of clumping in urn models is the number of empty urns. If the balls tend to clump, they will mostly drop into a few common urns, leaving the others sparsely occupied or unoccupied. Suppose, as a specific example, that $n = 100$ particles are quantized into $N = 20$ energy states and that the particles follow the Bose-Einstein model. The distribution of the number of empty urns, M_0, was worked out above, and $P(M_0 = m) = \frac{\binom{N}{m}\binom{n-1}{N-m-1}}{\binom{N+n-1}{n}} = \frac{\binom{20}{m}\binom{99}{19-m}}{\binom{119}{100}}, m = 0, 1, \ldots, 19$. For example, $P(M_0 = 0) = .02$, while $P(M_0 \geq 3) = .66$. The expected value of M_0 is 3.2, and the standard deviation is 1.5. A histogram of the distribution of M_0 is provided in Figure 15.1, and one can see a roughly symmetric *normal-like* distribution.

Example 15.4 (Empty Urns in the Maxwell-Boltzmann Scheme). We witnessed a roughly symmetric bell-shaped histogram for the number of empty urns under the Bose-Einstein scheme in our preceding example. This example will show that the shape of the histogram critically depends on the choice of the urn model. Consider now the case of a Maxwell-Boltzmann scheme, and suppose $n = 100$ balls are distributed into $N = 30$ urns according to a Maxwell-Boltzmann scheme. The formula for the exact distribution of M_0 was derived above in this section. If we use this exact formula, then we get, for instance, that $P(M_0 = 0) = .335$ and $P(M_0 \geq 3) \approx .01$. It appears that a fundamentally different kind of distribution for M_0 now emerges. Once again, a histogram will help us appreciate how the shape of the distribution

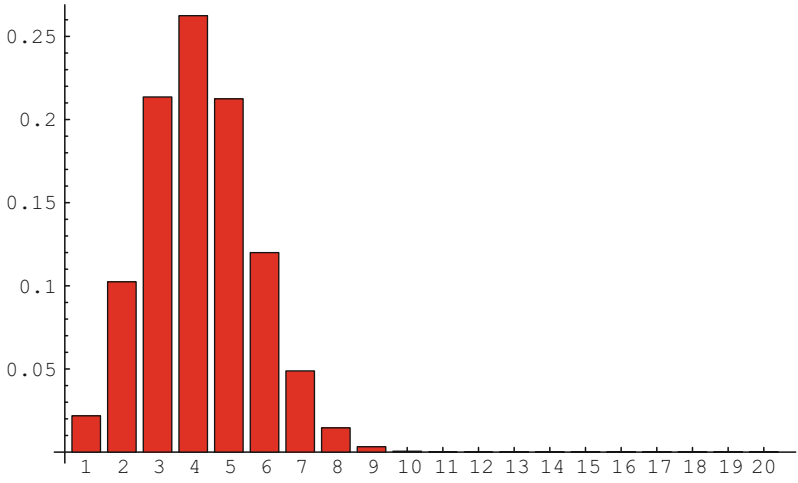


Fig. 15.1 Histogram of the number of empty urns in a Bose-Einstein scheme

changes under the Maxwell-Boltzmann scheme. We see an asymmetric skewed histogram. In fact, in contrast to the Bose-Einstein case, a Poisson distribution will approximate this histogram well under the Maxwell-Boltzmann scheme. The choice of the urn model matters; this is an important point.

15.3 ∗ Poisson Approximations

The exact distribution of the number of empty urns can be cumbersome to calculate when N and n are large because the formulas involve large factorials; see the exact formulas in the previous section. These cumbersome exact formulas can be approximated by more convenient expressions. However, exactly which approximation applies in a given case depends crucially on the relative magnitudes of n and N and *also on the exact urn model*. In the Bose-Einstein scheme, typically one does not get Poisson-type approximations; we had already noted this in our example and the histogram plot in Figure 15.1. But, fortunately, under the Maxwell-Boltzmann scheme, accurate Poisson approximations are available when N and n are large and satisfy suitable conditions on their relative magnitudes. Only the Poisson approximation is stated here, for purposes of simplicity. See Figure 15.2 for an illustration. One reference for this section and the theorem below is Johnson and Kotz (1977). Two other references are Kolchin et al. (1978) and Barbour et al. (1992).

Theorem 15.3.

(a) *Suppose* $N, n \to \infty$ *and that* $Ne^{-\frac{n}{N}} \to \lambda$ *for some positive and finite number* λ. *Then, for all* $k \geq 0$,

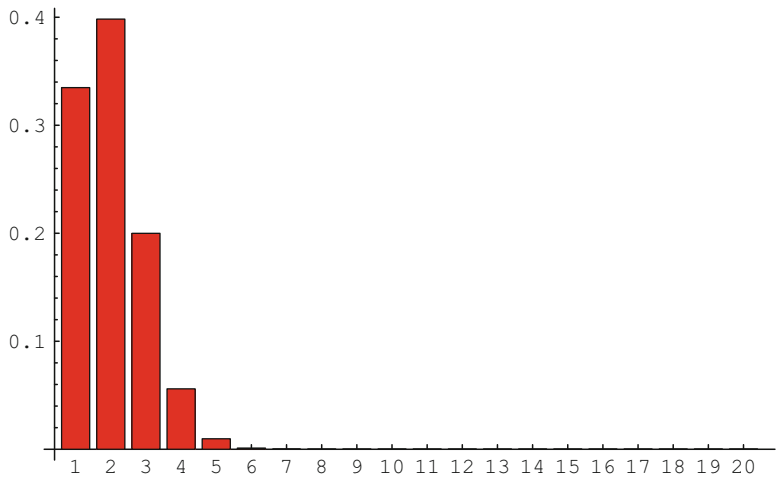


Fig. 15.2 Histogram of the number of empty urns in a Maxwell-Boltzmann scheme

$$P(M_0 = k) \to \frac{e^{-\lambda}\lambda^k}{k!}$$

as $N, n \to \infty$.

(b) *Suppose* $N, n \to \infty$ *and that* $\frac{n^2}{N} \to 2\lambda$ *for some positive and finite number* λ. *Then, for all* $k \geq 0$,

$$P(M_0 - (N - n) = k) \to \frac{e^{-\lambda}\lambda^k}{k!}$$

as $N, n \to \infty$.

Discussion. In the first case of this theorem, $n \gg N$, and there being many more balls than urns, not too many urns can be empty. Thus, a Poisson distribution with mean λ applies to the number of empty urns itself. In contrast, in the second case of the theorem, n is of the order of $\sqrt{N}$. So now there are far more urns than there are balls. Hence, we would expect to see a lot of empty urns. And, indeed, the second part of the theorem says that the number of empty urns would be about $N - n + \lambda$ on average, a large number! *An important case not covered by the theorem is when* N *and* n *are of comparable magnitude, i.e.,* $\frac{N}{n} \to \lambda$ *for some positive and finite number* λ. *In this case, a Poisson approximation does not apply.*

Example 15.5 (Testing the Poisson Approximation). To apply the Poisson approximation result above, we need to choose a value of λ. It is common to simply calculate $Ne^{-\frac{n}{N}}$ and apply part (a) of the theorem with this number as λ unless the number turns out to be too small. Some subjective judgment has to be used to decide if it is too small. If $n = 100$ and $N = 30$, then $Ne^{-\frac{n}{N}} = 1.07$. This is certainly not too small. If we use a Poisson distribution with mean $\lambda = 1.07$ as the approximation and use the exact distribution, which was derived theoretically for general n, N in the previous section, then here is how the two compare.

m	$P(M_0 = m)$(Exact)	Poisson Approximation
0	.3349	.3430
1	.3983	.3670
2	.1999	.1964
3	.0560	.0700
4	.0097	.0187
5	.0011	.0040

The maximum discrepancy is .031, which is reasonably small, although not extremely so. For most practical purposes, the Poisson approximation will probably suffice.

15.4 Pólya's Urn

Pólya's urn model is perhaps the most well-known urn model in which some form of replacement of the balls takes place as the drawing process evolves. The replacement is not as simple as in ordinary sampling with replacement. Pólya's urns were originally applied to model contagion processes, such as the spread of a contagious disease. The model has also been widely used for internal applications; i.e., mathematical results on the Pólya urn scheme have been useful in establishing properties of various methods in other areas of statistics. One example of such an internal application is the application of Pólya urns to *Bayesian statistics*, an area of statistics based on Bayes' theorem.

The Pólya urn scheme is defined as follows. Initially, an urn contains a white and b black balls, a total of $a + b$ balls. One ball is drawn at random from among all the balls in the urn. It, together with c more balls of its color, is returned to the urn, so that after the first draw, the urn has $a + b + c$ balls. This process is repeated.

The following notation will be used throughout this section: A_i is the event that the ith ball drawn in the Pólya urn scheme is white, X_i is the indicator of the event A_i, and, for given $n \geq 1$, $S_n = X_1 + \cdots + X_n$, which is the total number of times that a white ball has been drawn in the first n trials. First we will see a really interesting property of the sequence of indicator random variables $X_1, X_2, \ldots$.

To start with, evidently,

$$P(X_1 = 1) = \frac{a}{a+b}.$$

Next,

$$\begin{aligned} P(X_2 = 1) &= P(X_2 = 1 \,|X_1 = 1)P(X_1 = 1) + P(X_2 = 1 \,|X_1 = 0)P(X_1 = 0) \\ &= \frac{a}{a+b}\frac{a+c}{a+b+c} + \frac{b}{a+b}\frac{a}{a+b+c} = \frac{a^2 + ac + ab}{(a+b)(a+b+c)} \\ &= \frac{a(a+b+c)}{(a+b)(a+b+c)} = \frac{a}{a+b}. \end{aligned}$$

We notice that $P(X_2 = 1)$ and $P(X_1 = 1)$ are equal. Let us look at $P(X_3 = 1)$. This has to be found by conditioning on the colors of the balls chosen in the first two draws. Precisely,

$$\begin{aligned} P(X_3 = 1) = {} & P(X_3 = 1 \,|X_1 = 1, X_2 = 1)P(X_1 = 1, X_2 = 1) \\ & + P(X_3 = 1 \,|X_1 = 1, X_2 = 0)P(X_1 = 1, X_2 = 0) \\ & + P(X_3 = 1 \,|X_1 = 0, X_2 = 1)P(X_1 = 0, X_2 = 1) \\ & + P(X_3 = 1 \,|X_1 = 0, X_2 = 0)P(X_1 = 0, X_2 = 0) \\ = {} & \frac{a}{a+b}\frac{a+c}{a+b+c}\frac{a+2c}{a+b+2c} + \frac{a}{a+b}\frac{b}{a+b+c}\frac{a+c}{a+b+2c} \end{aligned}$$

$$+\frac{b}{a+b}\frac{a}{a+b+c}\frac{a+c}{a+b+2c}+\frac{b}{a+b}\frac{b+c}{a+b+c}\frac{a}{a+b+2c}$$
$$=\frac{a(a+c)(a+2c)+2ab(a+c)+ab(b+c)}{(a+b)(a+b+c)(a+b+2c)}=\frac{a}{a+b}$$

on factorizing the numerator in the last line as $a(a+b+c)(a+b+2c)$. So, now we see that $P(X_3=1)$, $P(X_2=1)$, and $P(X_1=1)$ are all equal. Indeed, the following general formulas hold. For notational simplicity, we have assumed that $c=1$ in the next theorem.

Theorem 15.4. *Consider the Pólya urn scheme with $c=1$. Then, for any $n\geq 1$,*

(a) $P(X_{n+1}=1\,|X_1=x_1,\ldots,X_n=x_n)=\dfrac{a+x_1+\cdots+x_n}{a+b+n}$;

(b) $P(X_1=x_1,\ldots,X_n=x_n)=\dfrac{a(a+1)\cdots(a+s_n-1)b(b+1)\cdots(b+n-s_n-1)}{(a+b)(a+b+1)\cdots(a+b+n-1)}$,
where $s_n=x_1+\cdots+x_n$; and

(c) $P(X_{n+1}=1)=P(X_n=1)=\cdots=P(X_1=1)=\dfrac{a}{a+b}$.

This last statement can be rewritten as $P(A_i)=P(A_1)$ for any i.

Hint to the proof. Part (a) is an easy exercise. Part (b) is proved by induction on n and by using part (a). Part (c) follows by combining part (a) and part (b) and summing over all $x_1, x_2, \ldots, x_n$.

One can similarly show (without too much algebraic effort) that probabilities of all pairwise intersections are the same; i.e., whatever indices i, j we take, $P(A_i \cap A_j)=P(A_1\cap A_2)$. In fact, a much stronger result is true. Here is the result.

Theorem 15.5. *Let $k\geq 1$ be any fixed integer. Let $j_1<j_2<\cdots<j_k$ be any k given indices. Then*

$$P(A_{j_1}\cap A_{j_2}\cap\cdots\cap A_{j_k})=P(A_1\cap A_2\cap\cdots\cap A_k).$$

A proof of this can be seen in Feller (1968).

An infinite sequence of events $A_1, A_2, \ldots$ that has this property, namely that for any k and any indices $j_1<j_2<\cdots<j_k$ the intersection probabilities $P(A_{j_1}\cap A_{j_2}\cap\cdots\cap A_{j_k})$ are all equal (i.e., *the choice of the indices $j_1,\ldots,j_k$ does not matter*), is called an *exchangeable sequence* of events. So, we can restate this last theorem as follows.

Theorem 15.6 (Exchangeability in the Pólya Urn Scheme). *For $i\geq 1$, let A_i be the event that the ith ball drawn according to the general Pólya urn scheme is white when trials are repeated indefinitely. Then the infinite sequence $A_1, A_2, \ldots$ is exchangeable.*

This is regarded as a classic fact in combinatorial probability.

15.5 Pólya-Eggenberger Distribution

A consequence of this exchangeability fact is that we can now write down an explicit formula for the distribution of S_n, the number of white balls drawn in the first n trials of the Pólya urn scheme. Once again, for notational simplicity, we take $c = 1$ in the next theorem.

Theorem 15.7. *Consider the Pólya urn scheme with $c = 1$. Take any fixed $n \geq 1$, and let $0 \leq k \leq n$. Then,*

$$P(S_n = k) = \binom{n}{k} \frac{a(a+1)\cdots(a+k-1)b(b+1)\cdots(b+n-k-1)}{(a+b)(a+b+1)\cdots(a+b+n-1)}.$$

Proof. We have already established that

$$P(X_1 = x_1, \ldots, X_n = x_n) = \frac{a(a+1)\cdots(a+s_n-1)b(b+1)\cdots(b+n-s_n-1)}{(a+b)(a+b+1)\cdots(a+b+n-1)},$$

where $s_n = x_1 + \cdots + x_n$. Consider any n-tuple $(x_1, \ldots, x_n)$ such that $s_n = x_1 + \cdots + x_n = k$. Then, by exchangeability,

$$\begin{aligned} P(S_n = k) &= \sum_{(x_1,\ldots,x_n);s_n=k} P(X_1 = x_1, \cdots, X_n = x_n) \\ &= \binom{n}{k} \frac{a(a+1)\cdots(a+k-1)b(b+1)\cdots(b+n-k-1)}{(a+b)(a+b+1)\cdots(a+b+n-1)}. \end{aligned}$$

Remark. For any given n, this is a distribution on the integers $0, 1, \ldots, n$, with parameters a, b. For a general c, the formula becomes

$$P(S_n = k) = \binom{n}{k} \frac{a(a+c)\cdots(a+(k-1)c)b(b+c)\cdots(b+(n-k-1)c)}{(a+b)(a+b+c)\cdots(a+b+(n-1)c)}.$$

This is the famous *Pólya-Eggenberger distribution* with parameters a, b, c. The case $c = 0$ specializes to the binomial distribution with parameters n and $p = \frac{a}{a+b}$, and the case $c = -1$ specializes to the hypergeometric distribution with parameters $n, D = a, N = a + b$.

A plot of the Pólya-Eggenberger distribution when $a = 10, b = 5, c = 1$, and $n = 20$ is provided in Figure 15.3 and shows the affinity of the distribution toward larger values caused by a being larger than b.

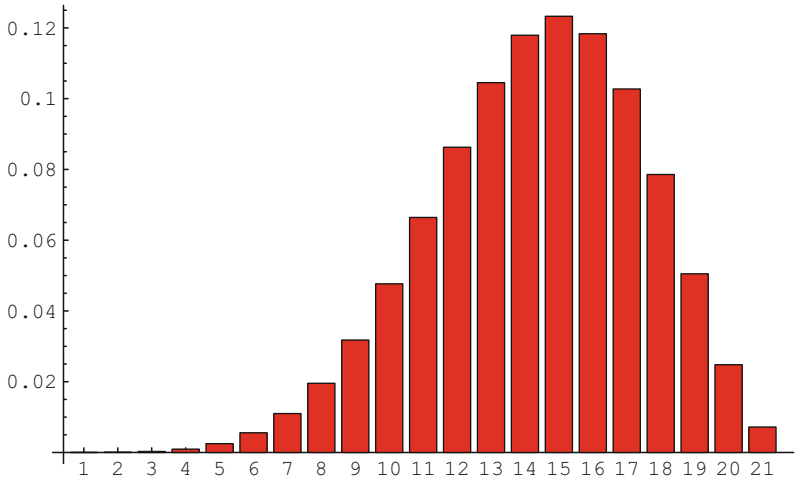


Fig. 15.3 Polya-Eggenberger distribution for $a = 10, b = 5, c = 1, n = 20$

15.6 ∗ de Finetti's Theorem and Pólya Urns

Exchangeability is a very fundamental concept in probability. It allows us to wriggle out of the assumption of mutual independence in a very neat way while still preserving abundant symmetry in the structure of the problem. It is not surprising that exchangeability has been studied carefully by probabilists. Three specific references are Regazzini (1987), Diaconis (1988), and Rao and Shanbhag (2001). An exposition at the textbook level is available in the classic book of Feller (1968) and in DasGupta (2008). The most profound result in the study of exchangeability is a theorem due to de Finetti, an Italian mathematician and probabilist. If we put together de Finetti's theorem on exchangeability and our result above on exchangeability of the sequence of events $A_1, A_2, \ldots$ in the Pólya urn scheme, then a remarkable result for Pólya urns emerges. The purpose of this section is to describe this result. First, we state de Finetti's (1931) theorem. A small word of caution is needed here. A really rigorous statement of de Finetti's theorem cannot be given without the use of some measure theory terminology. The statement given below is not fully rigorous; nevertheless, it makes the point that we need in this context.

Theorem 15.8. *Let $\{A_1, A_2, \ldots\}$ be an infinite sequence of exchangeable events. Take any fixed $n \geq 1$, and let S_n be the number of events among $A_1, A_2, \ldots, A_n$ that occur. Then there is a unique nonnegative function f on $[0,1]$ such that $\int_0^1 f(p)dp = 1$, and for any $k, 0 \leq k \leq n$,*

$$P(S_n{=}k){=}\int_0^1 \left[\binom{n}{k} p^k (1-p)^{n-k}\right] f(p)dp = \binom{n}{k}\int_0^1 p^k (1-p)^{n-k} f(p)dp.$$

Remark. Suppose, hypothetically, that the events $A_1, A_2, \ldots$ were mutually independent with a common probability of $p = .5$. Then, we know that our random variable S_n in de Finetti's theorem will be distributed as $Bin(n, p)$ with $p = .5$. De Finetti's theorem is saying that the most general exchangeable sequence of events must be a mixture of such binomial distributions, the mixing property being obtained via the integration with the function $f(p)$, a very profound result. Furthermore, we cannot have two different such functions f; for any particular exchangeable sequence of events, we can have just one such function f.

We will now apply this result to learn something about the Pólya urn scheme. To this end, recall that we have in fact already derived an explicit formula for $P(S_n = k)$ for the Pólya urn scheme by direct arguments. What we will now do is to take that direct formula and relate it to de Finetti's theorem in order to reach some interesting conclusions. We will work out an illustrative example to help us understand the general case.

Example 15.6. Consider the Pólya urn scheme with $a = b = c = 1$. Then, the formula for the distribution of S_n reduces to

$$P(S_n = k) = \binom{n}{k} \frac{k!(n-k)!}{2 \times 3 \times \cdots \times (n+1)} = \frac{1}{n+1}, 0 \le k \le n.$$

That is, if $a = b = c = 1$, then, for any n, S_n has a discrete uniform distribution on $\{0, 1, \cdots, n\}$. However, there is more. We know from our previous discussion of exchangeability that there is an underlying nonnegative function f on $[0, 1]$ with $\int_0^1 f(p)dp = 1$ and $P(S_n = k) = \binom{n}{k} \int_0^1 p^k (1-p)^{n-k} f(p)dp$. In particular, using $k = n$, for any $n \ge 1$,

$$P(S_n = n) = \frac{1}{n+1} = \int_0^1 p^n f(p)dp.$$

By inspection, the special function $f(p) \equiv 1$ satisfies this, and now we find on simple integration that this choice of f also satisfies $P(S_n = k) = \frac{1}{n+1}$ for any k between 0 and n. Thus, in the special case $a = b = c = 1$, we have the de Finetti representation

$$P(S_n = k) = \binom{n}{k} \int_0^1 p^k (1-p)^{n-k} dp.$$

The point is that we can explicitly identify the required function f for general a, b, c. Indeed, writing $\alpha = \frac{a}{c}, \beta = \frac{b}{c}$, the required function f is

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, 0 < p < 1,$$

where $\Gamma(z)$ is the *Gamma function* defined by the integral $\Gamma(z) = \int_0^\infty e^{-x} x^{z-1} dx$, $z > 0$. When $a = b = c = 1$, this reduces to $f(p) \equiv 1$. It is a Beta density. For now, we simply note the important finding that *the Pólya-Eggenberger distribution*

is a Beta mixture of binomial distributions. Techniques of more advanced probability theory can be usefully exploited to derive from this that, for large n, the distribution of $\frac{a+S_n}{n}$*, which is the proportion of white balls in the urn after n trials, is well approximated by this Beta density function* f*. To put it in simple terms, if we compute the exact distribution of* $\frac{S_n}{n}$ *and plot a histogram, the histogram will look like a plot of the function* $f(p)$ *on* $[0, 1]$*. This is a useful conclusion in analyzing the Pólya urn scheme.*

15.7 Urn Models in Genetics

Some basic and simple models for evolutionary processes in population genetics correspond to urn models with the balls having different colors. Perhaps the most basic such model of historical importance is the *Wright-Fisher model*. The Wright-Fisher model gives a mathematical model for how a specific allele frequency in a finite population changes over generations under certain assumptions on the population and the organism's mating behavior. The idea is that if in the long run a particular allele becomes extinct, then it contributes to a decrease in the genetic diversity in that population. Population geneticists call this *genetic drift*. Genetic drift accounts for the change in the genetic composition of a population over time due to purely random fluctuations. Other forces that act on the evolutionary mechanism include natural selection and mutation. Geneticists want to understand the relative weight of each factor in the evolutionary process. Two references for this section are Lange (2003) and Balding et al. (2007). See also Johnson and Kotz (1977).

15.7.1 Wright-Fisher Model

The assumptions we make under the Wright-Fisher model are that:

(a) The population size is a finite constant N and remains fixed from generation to generation.
(b) We consider one gene, and assume that it has two different forms or alleles, say A and B. A particular individual may have two copies of the same form or one of each. In other words, we have a diploid population with a total of $2N$ copies of the gene in each generation.
(c) The generations are nonoverlapping.
(d) The first generation has a certain initial supply, say i, of the allele form A, and the rest, namely $2N - i$, of the allele form B. The $2N$ genes of the next generation are produced using a *binomial model* in which each copy is a random pick from the gene pool of the previous generation and the $2N$ copies in the second generation are picked mutually independently.
(e) This process is continued indefinitely over generations.

It is clear from the definition of the model that if by chance in some generation each of the $2N$ alleles turns out to be of one kind, say all A alleles or all B alleles,

then this configuration will be preserved for all future generations. Geneticists call this *allele uniformity*.

We can phrase all of this in the form of the following urn model. Urn 1 has i red balls and $2N - i$ green balls. $2N$ balls are chosen from Urn 1 at random and with replacement. These balls are used to fill up another urn, say Urn 2. Thereafter, $2N$ balls are chosen from Urn 2 at random and with replacement, and these form the contents of Urn 3, and so on. The genetic composition of the nth generation corresponds to the number of red balls in Urn n.

We now need some notation. Let

$$\begin{aligned} p &= \frac{i}{2N} = \text{fraction of } A \text{ alleles in the first generation;} \\ X_n &= \text{total number of } A \text{ alleles in the } n\text{th generation;} \\ p_{jk} &= P(X_{n+1} = k \,|\, X_n = j); \\ p_\infty &= P(X_n = 2N \text{ for some finite } n); \\ E_x &= \text{expected number of generations needed to achieve allele uniformity.} \end{aligned}$$

From the assumption of binomial sampling, it follows that

$$p_{jk} = \binom{2N}{k} \left(\frac{j}{2N}\right)^k \left(1 - \frac{j}{2N}\right)^{2N-k}.$$

This is the *Wright-Fisher equation*. Let us work out an example.

Example 15.7. Genetic drift is known to be a more important factor in small isolated populations. Consider a situation where the size of the population is small, say $N = 50$. Also suppose that of the $2N = 100$ copies of the gene in the first generation, 40 are of the allele form A. Thus, $i = 40$ and $p = .4$. We therefore have $X_2 \sim Bin(100, .4)$. From the binomial distribution mean formula, $E(X_2) = 100 \times .4 = 40 = i$. This is a characteristic of the Wright-Fisher model; for any n, $E(X_n) = i$. We show a simulated pattern of the genetic composition in the first ten generations. For any n, X_n is simulated as a value from the $Bin(2N, \frac{x_{n-1}}{2N})$ distribution, where x_{n-1} is the realized value of X_{n-1}.

n	X_n
1	40
2	51
3	50
4	53
5	53
6	51
7	42
8	41
9	45
10	49

We can see in this table how purely random errors will lead to fluctuations in gene frequency over generations. We have not gotten anywhere close to allele uniformity in ten generations.

15.7.2 Time until Allele Uniformity

It turns out, however, that advanced probability-theoretic methods show that in the Wright-Fisher model allele uniformity will eventually take place. In the language of urn models, sooner or later all balls in an urn will be of the same color. The higher the initial proportion of red balls, the higher the probability that all balls in the urn will be red from some point onward. Indeed, the following neat result holds.

Theorem 15.9. $p_\infty = p$.

Thus, in our numerical example above, there is a 40% *chance that the allele form B will eventually vanish from the population. The expected number of generations that have to pass to obtain allele uniformity satisfies the following system of equations.*

Theorem 15.10.

$$E_i = 1 + \sum_{j=0}^{2N-1} \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} E_j, 1 \le i \le 2N-1.$$

Proof. A heuristic proof is as follows. Starting with i genes of the allele form A in the first generation, we obtain some j genes of the allele form A in the second generation, and then with j as our new initial frequency for the allele form A, we wait until we achieve allele uniformity, which is expected to take E_j more generations. The probability that a specific number j is the frequency of the allele form A in the second generation is $\binom{2N}{j}(\frac{i}{2N})^j(1-\frac{i}{2N})^{2N-j}$ by the Wright-Fisher equation. Now, simply sum over all possible values of j.

These equations are all linear in the required quantities $E_i, 1 \le i \le 2N-1$. Therefore, matrix methods can be used to successively generate the values of $E_1, E_2, E_3, \ldots$. Use of a computer is essential because solving for the E_i values involves inverting a matrix of order $(2N-1) \times (2N-1)$. By simple manipulation of the linear equations given in the theorem above, one can show that the vector of the E_i values is given by

$$\mathbf{E} = (I - P)^{-1}\mathbf{1},$$

where

$$\mathbf{E} = (E_1, \ldots, E_{2N-1})',$$

I is the $(2N-1) \times (2N-1)$ identity matrix,

P is the $(2N-1)\times(2N-1)$ matrix with elements $p_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^j \left(1-\frac{i}{2N}\right)^{2N-j}$,

$\mathbf{1}$ is the $(2N-1) \times$ 1-dimensional vector with all entries equal to 1.

It is the inversion of the matrix $I-P$ that requires a computer, and this inversion can become numerically unreliable or impossible for large N. Let us see an example.

Example 15.8. Consider a small population with $N = 25$ individuals. The number of genes in the allele form A is some number i between 1 and 49. We want to know the mean number of generations that has to pass for one of the two allele forms to become extinct.

Evaluating the inverse of $I-P$ and using the formula $\mathbf{E} = (I-P)^{-1}\mathbf{1}$, we find, for instance, that $E_1 = 9.13$ and $E_5 = 31.18$. That is, if there was just one gene of allele form A to start with, even then it would take more than nine generations on average to obtain allele uniformity; if there were five genes of allele form A to start with, it would take more than 31 generations to obtain allele uniformity. This is a general phenomenon. *Genetic drift progresses at a slow rate and gives ample opportunity for the other forces in evolution, such as natural selection, to take place.*

15.8 Mutation and Hoppe's Urn

The Wright-Fisher model assumes a single gene at a particular locus with two allele forms in a population of a fixed size and that binomial sampling from the gene pool of one generation forms the gene pool of the next generation. In the terminology of urn models, if each allele form is thought of as a color, then the number of colors is always two and the total number of balls is always the same. A biological generalization of this simple model is to envision an underlying set of infinitely many alleles that arise gradually over generations via a process of mutation of a previously existing special gene. The urn model formulation is the following. We start with an urn with some θ balls of a distinguished color (say black), To construct the nth urn for a general $n, n \geq 2$, we sample a ball at random from the $(n-1)$th urn. If this chosen ball happens to be black, then we put the black ball together with an additional ball of a previously unseen color back into the urn. The colors are given labels $1, 2, 3, \ldots$, in the order of their emergence. The labels do not have any other significance. If the chosen ball happens to be of some other color (that is, not black), then it is returned to the urn together with an additional ball of the same color. The appearance of a new color corresponds to the emergence of a new species. On the other hand, when a ball that is not black is chosen and is returned to the urn with another representative of the same color, that is supposed to correspond to a preexisting species simply multiplying in the population. It is important to note that the number of black balls (that is, balls of that distinguished color) always remains equal to θ. It is also important to note that if at some stage we choose a ball of the special color, then that adds a ball of a new nonspecial color into our urn. At any stage, each nonspecial color corresponds to one distinct species. The special color is not considered to be a species; the special color balls only generate new species. This is the well-known Hoppe urn scheme (Hoppe (1984)).

Biologically, the important questions are how many distinct species exist in the population after a prescribed number, say n, of generations, and what the respective sizes of these various species are. To investigate these questions, we need some notation:

$$n_i = \theta + i = \text{number of balls in the urn after } i \text{ iterations}, i \geq 0;$$
$$p_i = \frac{\theta}{n_{i-1}}, i \geq 1, W_i = I_{\{\text{the ball drawn at the } i\text{th iteration is a black ball}\}};$$
$$S_n = \text{number of balls in the urn of nonspecial colors after } n \text{ iterations}, n \geq 1.$$

Thus, $W_i \sim Ber(p_i)$, $S_n = \sum_{i=1}^{n} W_i$; also, from the drawing mechanism, we have that $W_1, W_2, \ldots$ is an independent (but not iid) sequence of random variables.

It follows immediately that

$$E(S_n) = E\left(\sum_{i=1}^{n} W_i\right) = \theta \sum_{i=1}^{n} \frac{1}{\theta + i - 1}$$
$$= \theta\left(\frac{1}{\theta} + \frac{1}{\theta+1} + \cdots + \frac{1}{\theta+n-1}\right).$$

For fixed θ and large n,

$$\frac{1}{\theta} + \frac{1}{\theta+1} + \cdots + \frac{1}{\theta+n-1} \approx \log n,$$

and hence $E(S_n) \approx \theta \log n$. Similarly,

$$\text{Var}(S_n) = \text{Var}\left(\sum_{i=1}^{n} W_i\right) = \sum_{i=1}^{n} \frac{\theta}{\theta+i-1}\left(1 - \frac{\theta}{\theta+i-1}\right)$$
$$= \theta \sum_{i=1}^{n} \frac{1}{\theta+i-1} - \theta^2 \sum_{i=1}^{n} \frac{1}{(\theta+i-1)^2}.$$

Now, for fixed θ and large n, $\sum_{i=1}^{n} \frac{1}{(\theta+i-1)^2}$ is small compared with the first term $\theta \sum_{i=1}^{n} \frac{1}{\theta+i-1}$ because, as we saw above, the first term $\theta \sum_{i=1}^{n} \frac{1}{\theta+i-1} \approx \theta \log n$, whereas $\sum_{i=1}^{n} \frac{1}{(\theta+i-1)^2}$ stays bounded as $n \to \infty$. Therefore, an approximation to the variance of S_n is $\text{Var}(S_n) \approx \theta \log n$. That is, for fixed θ and large n, the mean and the variance of S_n are both approximately equal to $\theta \log n$. *It is interesting that, even in the long run, the effect of the initial value θ does not go away.* As a matter of practical approximation, it is better to approximate the mean of S_n by $\theta(\log n - \log \theta)$; the next example will show some evidence of it.

Example 15.9 (Evolution of New Species). It is clear that, in the Hoppe urn scheme, new species arise according to a jump process. That is, S_n does not change if a nonblack ball is drawn at some iteration and increases by one when a black ball is

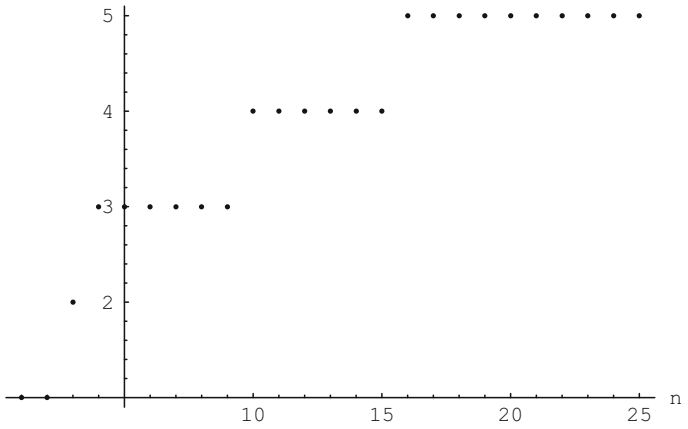


Fig. 15.4 Emergence of new species; plot of $S(n)$ vs. n; theta $=1$

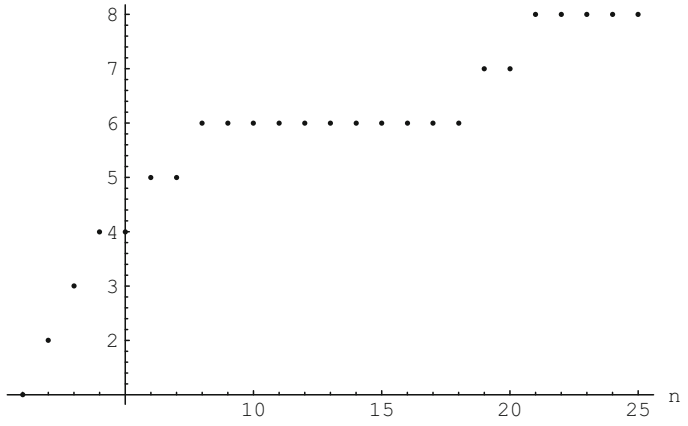


Fig. 15.5 Emergence of new species; plot of $S(n)$ vs. n; theta $= 10$

drawn at some iteration. Plots of two simulations are shown in Figures 15.4 and 15.5. The population is followed for up to 25 generations in these plots. In Figure 15.4, $\theta = 1$, and in Figure 15.5, $\theta = 10$. In the first case, the population ends up with five different species after 25 generations, and in the second case it ends up with eight different species after 25 generations. The approximation $E(S_n) \approx \theta \log n = 10 \log 25 = 32.2$ is very far from the realized value $S_{25} = 8$ in the second case. In comparison, the approximation $E(S_n) \approx \theta(\log n - \log \theta) = 10(\log 25 - \log 10) = 9.2$ is much closer to the realized value $S_{25} = 8$. *An interesting feature of the plots is that species seem to arise in spurts. We have new species arising in quick successions, and then we have long periods of inactivity. Real-life evolution seems to show similar spurt activity*.

The distribution of S_n itself is also of interest. Note that although S_n is a sum of n independent Bernoulli variables $W_1, \ldots, W_n$, the W_i are not iid. Thus, S_n is certainly not binomially distributed. However, the generating function of S_n can still be found in closed form, from which the mass function can be derived. Indeed, the generating function of S_n equals

$$G_{S_n}(s) = E(s^{S_n}) = \prod_{i=1}^{n} E(s^{W_i}) = \prod_{i=1}^{n} \left(\frac{i-1}{\theta + i - 1} + \frac{\theta}{\theta + i - 1} s \right)$$
$$= \frac{\prod_{j=0}^{n-1} (\theta s + j)}{\prod_{j=0}^{n-1} (\theta + j)}$$

on writing $j = i - 1$ in the products.

Recall now that from the definition of *Stirling numbers of the first kind*, for a given real number x, $x(x+1)\cdots(x+n-1) = \sum_{k=1}^{n} (-1)^{n-k} s(n,k) x^k$. Substituting into our expression for $G_{S_n}(s)$ above, we get the formula

$$G_{S_n}(s) = \frac{\sum_{k=1}^{n} (-1)^{n-k} s(n,k) \theta^k s^k}{\prod_{j=0}^{n-1} (\theta + j)}.$$

This expression is easier to manipulate for deriving the mass function of S_n. Indeed,

$$P(S_n = k) = \frac{G_{S_n}^{(k)}(0)}{k!} = \frac{(-1)^{n-k} s(n,k) \theta^k}{\prod_{j=0}^{n-1} (\theta + j)}$$

for $k \leq n$. Of course, for $k > n$, $P(S_n = k) = 0$. We have thus proved the following important result.

Theorem 15.11. (Distribution of Number of Distinct Species). *In the Hoppe urn scheme, the mass function of S_n is given by $P(S_n = k) = \frac{(-1)^{n-k} s(n,k) \theta^k}{\prod_{j=0}^{n-1} (\theta + j)}$, $1 \leq k \leq n$.*

For large n, computing these exact probabilities cannot be done without a computer because the formula involves the Stirling numbers. It can be shown that, for large n, S_n is approximately *normally distributed* with mean and variance $\theta(\log n - \log \theta)$.

15.9 ∗ The Ewens Sampling Formula

The Ewens sampling formula studies a question on species diversity. Suppose that in the population we have a total of n animals of different species. An interesting question is whether there are a few species of large abundance or many species of

more or less comparable abundances. For example, suppose there were 50 animals of different species in the population. One possible configuration would be that there are two species of size 20 each and another five species of size 2 each. A completely different type of configuration would be that there are 50 species, each of size 1! Which one is more likely? The Ewens sampling formula gives an analytic expression for the long-run probability that, in a population with n total animals, there are s_1 species each with one animal, s_2 species each with two animals, etc. Specifically, let $(s_1, s_2, \ldots, s_n)$ be any particular configuration that is physically possible; i.e., the s_i must be nonnegative integers such that $\sum_{i=1}^{n} i s_i = n$. Then the Ewens sampling formula says what is an analytic expression for $P(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n)$, where the uppercase $S_1, S_2, \ldots, S_n$ denote the number of species of sizes $1, 2, \ldots, n$, respectively, and $(s_1, s_2, \ldots, s_n)$ denotes a particular configuration. If the particular configuration is not physically possible, then of course $P(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n)$ would be zero.

Here is the Ewens sampling formula.

Theorem 15.12. *In a Hoppe urn scheme with θ balls of black (the special) color and n balls of other colors, let S_i denote the total number of different colors that have i balls each and $s_1, s_2, \ldots, s_n$ denote any arbitrary nonnegative integers satisfying $\sum_{i=1}^{n} i s_i = n$.*

Then

$$P(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) = \frac{n! \theta^{\sum_{i=1}^{n} s_i}}{(\prod_{i=1}^{n} i^{s_i})(\prod_{i=1}^{n} s_i!)(\prod_{i=1}^{n} (\theta + i - 1))}.$$

We will not prove this theorem here. A proof can be found in the original paper of Ewens (1972).

A second, different question is that of species abundance; in particular, whether the oldest species are more abundant in the population. Suppose we label the species according to their order of appearance. That is, the oldest species is called species number 1, the next oldest species is called species number 2, etc. Then, one has the following analytic expression for the long-run probability that, in a population of a total of n animals, there are m different species with the respective species sizes $N_1 = n_1, N_2 = n_2, \ldots, N_m = n_m$, where $\sum_{i=1}^{m} n_i = n$ and of course each $n_i \geq 1$. This was proved in Donnelly and Tavaré (1986); we will not prove it here.

Theorem 15.13. *In a Hoppe urn scheme with θ balls of black (the special) color and n balls of m other colors, let N_i denote the number of balls of color number $i, i = 1, 2, \ldots, m$. Suppose colors are labeled according to their order of appearance. Then,*

$$P(m; N_1 = n_1, \ldots, N_m = n_m) = \frac{\theta^m n!}{\prod_{i=1}^{n} (\theta + i - 1) n_m (n_m + n_{m-1}) \cdots (n_m + n_{m-1} + \cdots + n_1)},$$

where $n_i \geq 1$ and $\sum_{i=1}^{m} n_i = n$.

This has the interesting property that, for a given set of $n_1, n_2, \ldots, n_m$, the expression would be maximized if n_1 is the largest among the n_i, n_2 the second largest among the n_i, etc. As a simple example of what we mean, suppose n is 50 and there are just two species, so that $m = 2$. Then, it is more likely that the older species has 30 animals and the younger has 20 than the other way around. *So, we have the interesting result that older species are likely to be more abundant in the population*. We end with the converse question: Is the most abundant species the oldest one? Here is a neat formula for this probability.

Theorem 15.14. *Consider the Hoppe urn scheme. Suppose there are n balls of different colors (other than black) in the urn. Then the probability that a color with c balls of its kind is the first color to have arisen is* $\frac{c}{n}$.

15.10 Synopsis

(a) If the total number of balls (particles) is n, the total number of urns (states, or energy states) is N, and X_i is the number of balls that get distributed to the ith urn, then the pmf of X_i is as follows:

M-B scheme

$$P(X_i = k) = \binom{n}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{n-k};$$

B-E scheme

$$P(X_i = k) = \frac{\binom{n-k+N-2}{n-k}}{\binom{n+N-1}{n}};$$

F-D scheme

$$P(X_i = 1) = \frac{n}{N}, P(X_i = 0) = 1 - \frac{n}{N}.$$

(b) For the M-B scheme, the pmf of the number of empty urns has the exact pmf

$$P(M_0 = k) = \frac{\binom{N}{k} S(n, N-k)(N-k)!}{N^n}$$

and the Poisson approximation result

$$P(M_0 = k) \to \frac{e^{-\lambda}\lambda^k}{k!}$$

if $n, N \to \infty$ in such a way that $Ne^{-\frac{n}{N}} \to \lambda$, a finite nonzero number.

(c) For the general Pólya urn scheme,

$$P(S_n = k) = \binom{n}{k} \frac{a(a+c)\cdots(a+(k-1)c)b(b+c)\cdots(b+(n-k-1)c)}{(a+b)(a+b+c)\cdots(a+b+(n-1)c)},$$

where S_n denotes the number of times up to the nth draw that a white ball has been picked. This distribution is called the *Pólya-Eggenberger distribution*.

(d) This formula can be written in the alternative form

$$P(S_n = k) = \binom{n}{k} \int_0^1 p^k (1-p)^{n-k} f(p)dp,$$

where $f(p)$ is the density of a Beta distribution with parameters $\alpha = \frac{a}{c}, \beta = \frac{b}{c}$. This is an important result, and many other properties of the distribution of S_n can be derived from this connection with a Beta density.

(e) Under the *Wright-Fisher model*, if

$$\begin{aligned} X_n &= \text{total number of } A \text{ alleles in the } n\text{th generation,} \\ p &= \text{fraction of } A \text{ alleles in the first generation,} \\ p_{jk} &= P(X_{n+1} = k \,|X_n = j), \\ p_\infty &= P(X_n = 2N \text{ for some finite } n), \end{aligned}$$

then

$$p_{jk} = \binom{2N}{k} \left(\frac{j}{2N}\right)^k \left(1 - \frac{j}{2N}\right)^{2N-k}$$

and $p_\infty = p$.

(f) In the *Hoppe urn scheme*, the distribution of S_n, which is the number of distinct species in the population after n generations, has the exact pmf

$$P(S_n = k) = \frac{(-1)^{n-k} s(n,k)\theta^k}{\prod_{j=0}^{n-1}(\theta + j)},$$

where $s(n,k)$ denotes a Stirling number of the first kind. The mean and variance of S_n have the formulas

$$\begin{aligned} E(S_n) &= \theta\left(\frac{1}{\theta} + \frac{1}{\theta+1} + \cdots + \frac{1}{\theta+n-1}\right) \\ &\approx \theta(\log n - \log\theta), \\ \text{Var}(S_n) &= \theta \sum_{i=1}^{n} \frac{1}{\theta+i-1} - \theta^2 \sum_{i=1}^{n} \frac{1}{(\theta+i-1)^2} \\ &\approx \theta \log n. \end{aligned}$$

(g) In the *Hoppe urn scheme*, there is an exact formula for the joint distribution of the sizes of the distinct species in the population. The formula is known as the Donnelly-Tavaré formula and is related to the Ewens sampling formula. See the text for these two formulas.

15.11 Exercises

Exercise 15.1. Compute the first four moments of a discrete uniform distribution on $\{1, 2, 3, 4\}$ by using the factorial moments and the Stirling numbers of the second kind.

Exercise 15.2. * **(Geometric Moments).** Find a general factorial moment of a geometric distribution and convert them into formulas for the moments of the distribution.

Exercise 15.3. * **(Negative Binomial Factorial Moments).** Prove first that

(a) $(x + y)_{(n)} = \sum_{k=0}^{n} \binom{n}{k} x_{(k)} y_{(n-k)}$

for all reals x, y and for all $n \geq 1$.

(b) Hence find the factorial moments of the $NB(2, p)$ distribution, and generalize to the $NB(r, p)$ distribution.

Exercise 15.4. For each of the following cases, find the probability of one sample point under the M-B, B-E, and F-D schemes:

(a) $n = 5, N = 3$;
(b) $n = 20, N = 5$;
(c) $n = 20, N = 20$.

Exercise 15.5. Suppose five balls are distributed into three urns according to the Bose-Einstein scheme. Find the probability that at least one urn contains three or more balls.

Exercise 15.6. Suppose five balls are distributed into three urns according to the Bose-Einstein scheme. Find the expected value of the number of empty urns.

Exercise 15.7. Suppose five balls are distributed into three urns according to the Maxwell-Boltzmann scheme, but the probabilities that a particular ball drops into the three urns are .6, .3, .1, respectively. Find the expected value of the number of empty urns.

Hint: Try indicator variables.

Exercise 15.8. * Suppose n balls are distributed into N urns according to the Maxwell-Boltzmann scheme and that the probabilities that a particular ball drops into the N urns are $p_1, p_2, \ldots, p_N$, respectively. Prove that the expected value of the number of empty urns is minimized when each $p_i = \frac{1}{N}$.

Exercise 15.9. * Suppose n balls are distributed into N urns according to the Bose-Einstein scheme. Find a formula for the mean and the variance of M_0, the number of empty urns.

Hint: For the variance, try $E[M_0(M_0-1)]$ first.

Exercise 15.10. Fifty balls are distributed into ten urns according to the M-B scheme. Find the expected number of urns with k balls for $k = 2, 5, 10$. Repeat if the balls are distributed according to the B-E scheme.

Exercise 15.11. * Derive a formula for the distribution of the number of empty urns when n balls are distributed into N urns according to the Fermi-Dirac scheme.

Exercise 15.12. Suppose $n = 250$ balls are distributed into $N = 50$ cells according to the Maxwell-Boltzmann scheme. Find the Poisson approximation to $P(M_0 = k)$ for $k = 0, 1, 2, 3$.

Exercise 15.13. * In the Maxwell-Boltzmann scheme, if $\frac{n}{\sqrt{N}} \approx 3$, then what approximately is the expected number of empty urns?

Hint: Use the Poisson approximation result.

Exercise 15.14. For the Pólya urn scheme, write down the details of the proof of $P(X_n = 1) = \frac{a}{a+b}$ for all $n \geq 1$.

Exercise 15.15. * For the Pólya urn scheme, find the expected value of the number of white balls drawn in the first n trials by three methods:

(a) using $S_n = X_1 + \cdots + X_n$;
(b) using the Pólya-Eggenberger distribution;
(c) using the de Finetti representation of the Pólya-Eggenberger distribution.

Exercise 15.16. * Prove or disprove: The Pólya-Eggenberger distribution is unimodal for any values of a, b, c.

Exercise 15.17. For the Pólya urn scheme with $c = 1$, prove that the variance of S_n equals $\frac{na}{a+b}[\frac{(n-1)(a+1)}{a+b+1} + 1 - \frac{na}{a+b}]$.

Hint: First derive a formula for $E[S_n(S_n-1)]$.

Exercise 15.18. * Consider the Pólya urn scheme with $c = 1$. Let T be the first trial at which a white ball is drawn.

(a) First prove that $P(T = n) = \frac{P(S_n=1)}{n}$.
(b) From this, or by direct methods, find a formula for $P(T = n)$.
(c) Generalize to the case of a general value of c.

Hint: Look at the de Finetti representation.

Exercise 15.19. * **(Negative Binomial as Limit of Pólya Distribution).** In the Pólya urn scheme, denote $\frac{a}{a+b} = p, q = 1-p, r = \frac{1}{a+b}$. Suppose $n \to \infty$, $np \to \lambda, 0 < \lambda < \infty, nr \to \delta, 0 < \delta < \infty$. Prove that, for each fixed k, $P(S_n = k) \to \binom{k+\frac{\lambda}{\delta}-1}{k} \left(\frac{1}{1+\delta}\right)^{\frac{\lambda}{\delta}} \left(\frac{\delta}{1+\delta}\right)^k$.

Exercise 15.20 (Variance in Wright-Fisher Model). Consider a population of $N = 50$ individuals following the Wright-Fisher mating scheme, and suppose that i alleles of form A are present in the first generation. Find formulas for the variance of the number of A alleles in the population in the second generation.

Exercise 15.21. * (Allele Parity in Wright-Fisher Model). Let θ_n denote the probability that two alleles chosen independently at random from the gene pool of the nth generation are of the same form. Assume that the Wright-Fisher model holds. Show that

(a) $\theta_n = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\theta_{n-1}$;

(b) $\theta_1 = \dfrac{\binom{i}{2} + \binom{2N-i}{2}}{\binom{2N}{2}}$,

where i denotes the number of A alleles in the first generation;

(c) $\theta_n = 1 - \left(1 - \frac{1}{2N}\right)^{n-1}(1 - \theta_1)$;

(d) Evaluate θ_n for $n = 1, 2, \ldots, 10$ when $N = 25, i = 20$.

Exercise 15.22 (Hoppe's Urn). In Hoppe's urn scheme, which probability is larger, $P(S_n = 1)$ or $P(S_n = n - 1)$?

Exercise 15.23 (Weak Law in Hoppe's Urn). In Hoppe's urn scheme, show that for some suitable sequence c_n and a constant c, for any $\epsilon > 0$, $P(|\frac{S_n}{c_n} - c| > \epsilon) \to 0$ as $n \to \infty$, and identify such a sequence c_n and the constant c.

Exercise 15.24 (Conditional Distributions in Hoppe's Urn). In Hoppe's urn scheme, find expressions for

(a) $P(S_{n+1} = j \mid S_n = k)$;
(b) $P(S_{n+2} = j \mid S_n = k)$;
(c) * $P(S_{n+m} = j \mid S_n = k)$.

In the above, for suitable j, depending on k, the conditional probabilities will be zero.

Exercise 15.25 (Poisson Approximation in Hoppe's Urn). Consider a Hoppe urn with $\theta = 1$. Find the exact distribution of S_{10} and also a suitable Poisson approximation. Compare them. Why do you think that a Poisson approximation is worth considering?

Exercise 15.26 (Size of the Oldest Species). In Hoppe's urn scheme, let Z_n denote the number of animals of the oldest species after n iterations. Find a closed-form formula for the mean and the variance of Z_n for a general θ.

Exercise 15.27 (Simple Application of Ewens' Formula). In a population with $n = 10$ animals, compute the probability that there are two species, each with five animals, and the probability that there are five species, each with two animals.

Exercise 15.28. * (Ewens' Formula). In a population with n animals, derive an expression for the probability that there are no species with strictly larger than two animals.

Exercise 15.29 (Use Your Computer). Plot the Pólya-Eggenberger distribution for each of the following cases and then superimpose the Beta density function $f(p)$ as defined in the text for each corresponding case. Comment on the accuracy.

(a) $a = b = 5, c = 1, n = 10$;
(b) $a = b = 5, c = 1, n = 25$;
(c) $a = 5, b = 20, c = 5, n = 100$.

Exercise 15.30 (Use Your Computer). Compute the expected number of generations until allele uniformity in a population with $N = 50$ individuals satisfying the Wright-Fisher model. Take i to be between 1 and 50. Plot the expected values as a function of i.

Exercise 15.31 (Use Your Computer). Plot the exact distribution of the number of species after n generations in a Hoppe urn scheme with $\theta = 5$ and $n = 10, 20, 30, 50$. Plot by using histograms. How does the histogram evolve as n increases?

Exercise 15.32 (Use Your Computer). Generate the Stirling numbers of the second kind $S(n, k)$ for n between 2 and 20. Verify that, for each n, $S(n, k)$ has exactly one turning point. Tabulate the turning point for each n considered. What relation do you see for the turning points corresponding to consecutive values of n?

References

Balding, D., Bishop, M., and Cannings, C. (2007). *Handbook of Statistical Genetics*, third ed., Wiley, New York.

Barbour, A., Holst, L., and Janson, S. (1992). *Poisson Approximation*, Clarendon Press, Oxford.

Bernoulli, J. (1713). *Ars Conjectandi*, translated by Edith Sylla, 2005, Johns Hopkins University Press, Baltimore.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer, New York.

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio, Atti R. Accad. Naz. Lincei, Ser. 6, Mem. Cl. Sci. Fis. Mat. Nat., 4, 251–299.

Diaconis, P. (1988). Recent progress on de Finetti's notions of exchangeability, in *Bayesian Statistics*, Vol. 3, J. Bernardo ed. 111–125, Oxford University Press, New York.

Donnelly, P. and Tavaré, S. (1986). The ages of alleles and a coalescent, Adv. Appl. Prob., 18, 1–19.

Ewens, W. (1972). The sampling theory of selectively neutral alleles, Theort. Pop. Biol., 3, 87–112.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Wiley, New York.

Gani, J. (2004). Random allocation and urn models, J. Appl. Prob., 41A, 313–320.
Hoppe, F. (1984). Pólya-like urns and the Ewens sampling formula, J. Math. Biol., 20, 91–94.
Ivchenko, G. and Medvedev, Y. (1997). *The Contribution of the Russian Mathematicians to the Study of Urn Models*, VSP, Utrecht.
Johnson, N. and Kotz, S. (1977). *Urn Models and Their Application*, Wiley, New York.
Kolchin, V., Sevast'yanov, B., and Chistyakov, V. (1978). *Random Allocations*, V. H. Winston & Sons, Washington, DC.
Lange, K. (2003). *Applied Probability*, Springer, New York.
Rao, C.R. and Shanbhag, D. (2001). Exchangeability, functional equations, and characterizations, in *Stochastic Processes: Theory and Methods*, Shanbhag, D. and Rao, C.R. eds. 733–763, North-Holland, Amsterdam.
Regazzini, E. (1987). On the origins of the concept of exchangeability in probability and statistics, a reminiscence of Bruno de Finetti, Rend. Semin. Mat. Fis. Milano, 57, 261–273.
Tomescu, I. (1985). *Problems in Combinatorics and Graph Theory*, Wiley Interscience, New York.
Whitworth, W. (1901). *Choice and Chance*, Hafner, New York.

Appendix I: Supplementary Homework and Practice Problems

I.1 Word Problems

Chapters 1–4

Exercise I.1. How many possible initials can be formed if each person has two given names and a last name? What if each person has at most two given names and a last name?

Exercise I.2. Three numbers are chosen at random with replacement from $0, 1, \ldots, 9$. Find the probabilities that the three are alike, that the three are distinct, and that exactly two are alike.

Exercise I.3. A, B, C, D are four independent events, each with probability .5. Find the probability that at least two of the four events occur.

Exercise I.4. The birthdays of five random people are known to fall in exactly three calendar months. Find the probability that exactly two of the five were born in January. State your assumptions.

Exercise I.5. There are two good bulbs and two bad bulbs in a package. These will be tested one by one in a random order. Find the probabilities that the second bad bulb is the second bulb tested, the third bulb tested, and the fourth bulb tested.

Exercise I.6. Suppose $B_1, B_2, \ldots$ are infinitely many events, and let B be their union. An event A is independent of each individual B_i. Prove, or give a counterexample, that A and B are independent.

Exercise I.7. A man seeks advice from three oracles on whether or not to accept a particular job offer. He acts according to the advice of the majority. The three oracles have probabilities .95, .9, .95 of giving the correct advice. Find the probability that the man will take the correct action. State your assumptions.

Exercise I.8. Three people, say A, B, C, take turns rolling a fair die. A rolls first, then B, and then C. The first to roll a five wins. Find the probabilities of each player winning.

Exercise I.9.
(a) A six-sided die is manipulated in such a way that the face with the number i has probability proportional to i. Find the probability that the die will produce an even number if it is rolled once.
(b) An n-sided die is manipulated in such a way that the face with the number i has probability proportional to i. Find the probability that the die will produce an even number if it is rolled once.

Exercise I.10. There are ten black, ten white, and ten blue balls in an urn. Ten of these are chosen at random without replacement. Find the probability that there is at least one ball of each color among the ten drawn.

Exercise I.11. There are ten black, ten white, and ten blue balls in an urn. Ten of these are chosen at random without replacement. Find the probability that the first blue ball is drawn on the sixth draw.

Exercise I.12. There are ten black, ten white, and ten blue balls in an urn. Ten of these are chosen at random without replacement. Find the probability that the second ball drawn is blue if the third ball drawn is known to be blue.

Exercise I.13. Box 1 has two good bulbs and two bad ones; box 2 has three good bulbs and two bad ones. One bulb is chosen at random from box 1 and transferred to box 2. Then, one bulb is chosen at random from box 2. It is found to be a good bulb. What is the probability that the bulb from box 1 that was transferred to box 2 was a good bulb?

Exercise I.14. Find the probability that a hand in bridge has four cards of one suit and three cards each of three other suits.

Exercise I.15. Which is more likely: that a bridge hand will contain one card of each denomination or that it will contain cards of only two suits?

Exercise I.16. An urn contains four black, four white, and four blue balls. Three balls are drawn at random from the urn. Is it more likely that the balls will all be of the same color if sampling is with replacement or without replacement?

Exercise I.17. Find the probability that a randomly selected bridge hand will be void in at least one suit.

Exercise I.18. Find the probability that a randomly selected bridge hand will contain exactly five cards of at least one suit.

Exercise I.19. Cards are taken out one at a time from a well-shuffled deck. What is the probability that it will take at least five and at most ten draws to take out the first club?

Exercise I.20. A fair coin is tossed six times. Given that there are at least three heads, what is the probability that there are exactly four heads?

Exercise I.21. A fair die is rolled 12 times. Given that there are exactly two ones, what is the probability that there are exactly two sixes?

Exercise I.22. A fair die is rolled twice. Compute the probability that the sum of the two rolls is 3, 5, 7, 9, 11, respectively, given that the sum is odd.

Exercise I.23. A fair die was rolled four times. The faces 1 and 2 never appeared. What is the probability that the other four faces each appeared exactly once?

Exercise I.24. Jeff, Jen, and Cathy shoot at a bull's eye. They can hit the bull's eye 70%, 80%, and 75% of the time, respectively. One of the three is known to have hit the bull's eye. Find the probability that it was Jen.

Exercise I.25. A fair coin is tossed ten times. Given that at least seven heads were obtained, what is the probability that the first toss was a head? That at least one of the first two tosses was a head? That the first two tosses were both heads?

Exercise I.26. From a town of 25 Republicans and 25 Democrats, pollsters A and B each sampled ten residents at random without replacement. Find the probability that the two polls contained exactly the same number of Republicans.

Exercise I.27. A, B, C are three events. If A is independent of B given C and if C is independent of B, are A and B independent events? Prove or give a counterexample.

Exercise I.28. A library patron has decided to try five libraries for a particular book. Each library has a 50% chance of having the book, and if a library has the book, there is a 20% chance that it will be checked out. Find the probability that the patron can find the book. State your assumptions.

Exercise I.29. An urn has two white and three green balls. A number is selected at random from $1, 2, 3, 4, 5$, and then that many balls are taken out from the urn. Find the probability that they are all green.

Exercise I.30. An urn has five white and five green balls. Five balls are drawn at random without replacement. Find the probability that in each odd-numbered draw a green ball is drawn.

Exercise I.31. On a table, there are two dice. One is a fair die, and the faces of the other die are 1, 1, 2, 2, 6, 6. One die is selected at random and rolled, and it gives a six. What is the probability that it was the fair die?

Exercise I.32. A number is chosen at random from $1, 2, \ldots, 200$. Find the probability that it is even if it is not divisible by 7.

Exercise I.33. Team X plays against team Y in a best of seven series. In each game, team X has a 70% chance of winning, and assume that the games are independent. Find the probabilities that X wins, that X wins within five games, and that the series ends within five games.

Exercise I.34. A, B, C are three pairwise independent events. Also, A and $B \cap C^c$ are independent. Show that A, B, C are mutually independent.

Exercise I.35. Suppose a discrete random variable X has the distribution $P(X = n) = 2^{-n}, n \geq 1$.

(a) Find the mean of X.
(b) Find all medians of X.
(c) Find the variance of X.
(d) Find $P(|X - \mu| \geq 2\sigma)$, and compare it with the bound of Chebyshev's inequality.

Exercise I.36. Suppose X has a finite variance. Does $|X|$ have the same, a smaller, or a larger variance than X?

Exercise I.37. Suppose a discrete random variable X has a distribution such that $P(X > n + 1 | X > n) = \frac{n+1}{n+2}$ for all $n \geq 1$. Find the probability mass function of X.

Exercise I.38. Cards are drawn one at a time, without replacement, from a deck of 52 cards until the first club card is obtained. Let X be the number of draws required.

(a) Find the mass function of X.
(b) Find the mean of X.

Exercise I.39. X is uniformly distributed on $\{1, 2, \ldots, n\}$, and Y is uniformly distributed on $\{2, 4, \ldots, 2n\}$; X and Y are independent variables.

(a) Find the variance of $X + Y$.
(b) Find the variance of XY.
(c) Find $P(Y > X)$.

Exercise I.40. Given positive numbers M, ϵ, show a random variable X such that $\sigma^2 \geq M$ but $P(|X - \mu| > .01) < \epsilon$.

Exercise I.41. Suppose X has a positive mean μ and that $E(X^2)$ is also equal to μ. Prove that $\text{Var}(X) \leq \frac{1}{4}$.

Exercise I.42. X takes the values $1, 2, 3, 4$, and we know that $P(X = 1) = P(X = 2) = 2P(X = 3) = 3P(X = 4)$. Find the distribution of X.

Exercise I.43. Three fair dice are continually rolled until a sum of 15 on the three dice is obtained. Find the expected number of times the three dice would have to be rolled.

Exercise I.44. Four distinguishable balls are distributed independently at random into three distinguishable cells. Let X be the number of balls that land in the first

cell, Y the number of balls that land in the second cell, and Z the number of cells that remain empty.

(a) Find $E(X)$ and $E(Y)$.
(b) Find $\text{Var}(X)$ and $\text{Var}(Y)$.
(c) Find $P(Z = 0)$ and $P(Z = 2)$.
(d) Find $E(Z)$.

Exercise I.45. A fair die is rolled six times. Let X be the sum of the first four rolls and Y the sum of the last four rolls. Find the variance of $X - Y$.

Exercise I.46. A fair die is rolled six times, and let $X_1, X_2, \ldots, X_6$ be the six rolls obtained, respectively. Find the mean and the variance of $\sum_{i=1}^{6} (-1)^{X_i} X_i$.

Exercise I.47. Twenty-five people will each toss a fair coin 20 times. Let X be the number of people among the 25 people who get exactly ten heads and ten tails. Find the mean and variance of X.

Exercise I.48. Give an example of a random variable such that $E(X) = 1$ and $\text{Var}(X) > 100$.

Exercise I.49. Give an example of a random variable such that $E(X) = 100$ and $\text{Var}(X) = 1$.

Exercise I.50. In bridge, find the expected number of players who receive no aces or no hearts.

Exercise I.51. Coupons are drawn, independently with replacement, from a set of ten coupons. Find the expected number of draws:

(a) until the first coupon drawn is drawn again;
(b) until a duplicate occurs.

Exercise I.52. A fair die is rolled one hundred times. Find the expected number of rolls such that it and the next roll show the same face.

Exercise I.53. A random variable X takes the values $1, 0$ with probabilities p, $1 - p$. It has a variance equal to .16, and we know that $E(X - p)^3 > 0$. Find p.

Exercise I.54. Consider couples that have children until they have a girl. What is the expected proportion of boys in such families?

Chapters 5 and 6

Exercise I.55. A random variable X takes the values $0, \pm 1, \pm 2$ with the equal probability $\frac{1}{5}$. Find the mgf of X and $E(X)$. Verify that $E(X) = \psi'(0)$, ψ being the mgf.

Exercise I.56. Suppose $X \sim Bin(n, p)$, where $p = \frac{1}{2}$. Define Y as $Y = X$ if X is even and zero if X is odd. Find the mgf of Y and hence $E(Y)$.

Exercise I.57. Suppose $X \sim Ber(p)$ and $Y \sim Poi(\lambda)$, and assume that X and Y are independent. Find the mgf of XY.

Exercise I.58. Suppose X has a finite mean μ and a finite variance σ^2, and that its mgf $\psi(t)$ exists in some interval around zero. Show that $\sigma^2 = \xi''(0)$, where $\xi(t) = e^{-t\mu}\psi(t)$.

Exercise I.59. Suppose $X_i \overset{indep.}{\sim} Ber(p_i), i = 1, 2, \ldots, n$. Find the mgf of $X_1 + \cdots + X_n$, and hence the variance of $X_1 + \cdots + X_n$.

Exercise I.60. Suppose X has the mgf $\psi(t) = \cosh t, -\infty < t < \infty$. Find the distribution of X.

Exercise I.61. Suppose X has the mgf $\psi(t) = \frac{\sinh t}{t}$ for $t \ni 0$, and $\psi(0) = 1$. Find the distribution of X.

Exercise I.62. Suppose $X \sim Poi(1)$, and define $X_n = XI_{\{X \leq n\}}, n \geq 1$.

(a) Find $\psi_n(t)$, the mgf of X_n.
(b) Find $\lim_{n\to\infty} \psi_n(t)$.

Exercise I.63. Find the factorial moments of the $Bin(n, p)$ distribution.

Exercise I.64. Find the factorial moments of the $Poi(\lambda)$ distribution.

Exercise I.65. A random variable X has the generating function (pgf) $c + \frac{1}{2}s + \frac{1}{4}s^2 + \frac{1}{8}s^3$ for some c.

(a) Find c.
(b) Find the distribution of X.
(c) Find the mean of X.

Exercise I.66. Find the generating function of a general Poisson distribution.

Exercise I.67. Find the generating function of the $NB(r, p)$ distribution.

Exercise I.68. Is it possible that neither of two random variables X and Y has a finite mgf in any interval around zero but $X + Y$ does in all intervals around zero?

Exercise I.69. Suppose X has a finite mgf in some interval around zero. Does $|X|$ also have a finite mgf in some interval around zero?

Exercise I.70. A binomial random variable has mean 6 and variance 2.4. Evaluate $P(X < 5)$.

Exercise I.71. Harry's experience is that 7% of the parcels he mails do not reach their destination. He has bought two books for 25 dollars apiece and wants to mail them to his brother. If he sends them in one parcel, the postage is 6 dollars, while for separate parcels the postage is 4 dollars for each parcel. To minimize his expected total cost (possible loss of books + postage), should he send one or two parcels?

Exercise I.72. You are promised a reward if you obtain exactly ten heads by tossing a coin. How many times you toss the coin is up to you, but you have to announce this before starting. Assuming the coin is a fair coin, what is the best number of times to toss the coin?

Exercise I.73. Suppose I roll three dice. Those that show a six are rolled again. Let X be the number of resulting sixes. Find the distribution, mean, and variance of X.

Exercise I.74. Printing errors occur on any specific page in a book with probability .01. A certain book has 400 pages.

(a) Find the probability that the book has ten or more printing errors.
(b) Find the probability that the first hundred pages are error-free.
(c) Find the probability that page 90 is error-free.
(d) Find the probability that the first error occurs on page 91.
(e) Find the probability that there are exactly three errors in pages 1 to 200 and exactly three errors in pages 201 to 400.

Exercise I.75. A telephone operator receives 25 calls on average per hour. What is the probability that in two consecutive five minute intervals she receives no calls at all?

Exercise I.76. A Poisson random variable has the property that $\psi(0) = \psi(1)$, where ψ denotes its mgf. Find $P(X > 1)$.

Exercise I.77. Peter has a coin that gives heads with probability p in individual tosses. Paul has a coin that gives heads with probability θ in individual tosses. Both toss their coins repeatedly. Let Y be the first toss at which Paul obtains a head and X be the number of heads Peter obtains up to and including the Yth toss. Find the mean of X.

Exercise I.78. Suppose $X \sim Bin(20, .1)$. Compute $P(X \leq k)$ for $k = 1, 2, 3$ exactly and then by using the normal approximation with and without a continuity correction. Compare the approximations with the exact values.

Exercise I.79. Let X be the number of people who will want to buy the daily newspaper from a vendor on a given day.

(a) Suppose $X \sim Poi(\lambda)$ with $\lambda = 10$. If the vendor stocks 14 papers, what is the probability that the demand will exceed the supply?
(b) Suppose $X \sim Bin(n, p)$ with $n = 20$, $p = .5$. If the vendor stocks 14 papers, what is the probability that the demand will exceed the supply?

(c) In each case, find the minimum number of papers the vendor should stock so that the chance that the demand will exceed the supply is at most 5%.

Exercise I.80.
(a) Compute the exact probability that a bridge hand is void in spades.
(b) Compute the exact probability that in one hundred independent plays at least twice a player finds his hand to be void in spades.
(c) Compute the Poisson approximation to the probability above.

Exercise I.81. For each of $p = .05, .1, .25, .4$, find the smallest value of n such that the $Bin(n, p)$ distribution has a skewness $\leq .2$ and kurtosis $\leq .1$.

Exercise I.82. Peter has a coin that gives heads with probability .6 in individual tosses, and Paul has a fair coin. Both toss their coins repeatedly. Let X and Y be the first tosses at which they obtain the first heads, respectively. Find the distribution and the mean of $\max\{X, Y\}$.

Exercise I.83. Suppose X and Y are independent Poisson random variables with means λ, μ. Find the mgf of $X - Y$.

Exercise I.84. Suppose $X_1, X_2, \ldots, X_{10}$ are independent Bernoulli variables with the common parameter p. Find the mgf of $X_1 - X_2 + X_3 - X_4 + \cdots - X_{10}$.

Exercise I.85. Find the first four moments of a Poisson distribution with mean 2.

Exercise I.86. Find the first four moments of a $Bin(10, .5)$ distribution.

Exercise I.87. Suppose $X \sim Bin(10, .5)$. Compute $E(X - 5)^5$.

Exercise I.88. Cards are drawn one by one from a deck of 52 cards.

(a) Compute the expected number of draws necessary to draw the first ace.
(b) Compute the expected number of draws necessary to draw the second ace.

Exercise I.89. Suppose a couple will have children until they have at least one boy and at least one girl, but they will not have more than four children. Compute the expected number of children they will have.

Exercise I.90. A coin with probability p for heads is repeatedly tossed until r heads or r tails are obtained, whichever happens first. Find the mass function of the number of tosses necessary.

Exercise I.91. For $i = 1, 2, \ldots 10$, let X_i be a randomly selected number from $\{1, 2, \ldots, i\}$. Find the expected number of even numbers drawn; i.e., the expected value of the number of X_i that are even.

Exercise I.92. Suppose X and Y are independent Poisson random variables with means λ, μ. Can XY have a Poisson distribution for any λ, μ?

Exercise I.93. Suppose $X_1, X_2, \ldots, X_n$ are n independent random variables. Show that $\text{Var}(X_1 X_2 \ldots X_n) \geq \text{Var}(X_1)\text{Var}(X_2) \ldots \text{Var}(X_n)$.

Exercise I.94. Suppose a random variable X is such that $E(X) = 0, E(X^2) = 1, E(X^6) = 1$. Find and plot the CDF of X.

Exercise I.95. Find all medians of the number of aces in a bridge hand.

Exercise I.96. In a small town of 100 people, there are 90 right-handed and ten left-handed people. If ten tosses of a fair coin produce eight or more heads, a sample of 20 people with replacement will be taken. If the number of heads is less than eight, a sample of ten people without replacement will be taken. Find the expected number of left-handed people in the sample.

Chapters 7–10

Exercise I.97. A density function is verbally described as follows: it is zero for $x < 1$, rises linearly between 1 and 2 to $\frac{1}{3}$, remains constant between 2 and 4, decreases linearly to zero from 4 to 5, and remains zero thereafter.

(a) Plot the density function.
(b) Find the corresponding CDF and plot it.
(c) Find the mean of the distribution.
(d) Find $P(2.5 < X < 4.5)$.

Exercise I.98. A random variable X has the density cx for x between 0 and .5 and $c(1 - x)$ for x between .5 and 1.

(a) Find the normalizing constant c.
(b) Let A, B, C be the three events $X < .5, X > .5, .25 < X < .75$. Find $P(A|B)$; $P(C)$; $P(C|A)$; $P(C|A \cap B)$.

Exercise I.99. Suppose we know that the following functions are valid CDFs. Find, for each case, the smallest number M such that $F(M) = 1$.

(a) $F(x) = x^2/4, x \geq 0$.
(b) $F(x) = \log x, x \geq 1$.
(c) $F(x) = \dfrac{1 - \cos ax}{2}, x, a > 0$.

Exercise I.100. Annual rainfall in a desert town is zero with probability .9, and if it rains in some year, then the amount is exponential with mean 2 in. Plot the CDF of the amount of rainfall in this town.

Exercise I.101. The pth quantiles for $p = .1, .2, \ldots, .9$ are called the deciles of a distribution. Compute approximately the deciles of the exponential distribution with mean 1, a Beta distribution with parameters 2 and 1, and a standard normal distribution.

Exercise I.102. Suppose X has the standard double exponential density. Compute each of the following probabilities:

(a) X is a prime number;
(b) X is an irrational number;
(c) $X^3 - X^2 - X - 2 > 0$;
(d) $|X| + |X - 3| > 3$;
(e) $|X|e^{-|X|} > e^{-1}$.

Exercise I.103. Suppose $X \sim U[0, 1]$. Find the density of e^{-X^2}.

Exercise I.104. Suppose $X \sim Exp(1)$. Find the density of $2e^{-X}$.

Exercise I.105. Suppose $X \sim Exp(1)$. Define a function $g(X)$ as $g(X) = X$ if $X < 1$ and $g(X) = \frac{1}{X}$ if $X > 1$. Find the density of $Y = g(X)$.

Exercise I.106. Suppose X has the density $\frac{1}{x^2}$ for $x \geq 1$. Define a function $g(X)$ as $g(X) = 2X$ for $X \leq 2$ and $g(X) = X^2$ for $X > 2$. Find the density of $Y = g(X)$.

Exercise I.107. Suppose $Z \sim U[-1, 1]$ and X takes values ± 1 with probability $\frac{1}{2}$ each. We know that X and Z are independent.

(a) Find the CDF of $Y = ZX$.
(b) Find the density of $Y = ZX$.

Exercise I.108. Household incomes in a town have a Pareto distribution with $\theta = 10$; the value of the α parameter is not explicitly given. We know that the mean income is 40,000 dollars.

(a) Find the value of α.
(b) What percentage of the families earn more than 50,000 dollars?

Exercise I.109. It is known that the shortest interval containing 95% of the total area in a normal distribution is $[2, 8]$. Find:

(a) the mean and variance of this normal distribution;
(b) the 90th percentile of this normal distribution;
(c) the area between 5 and 10 in this normal distribution.

Exercise I.110. Find the shortest interval with probability $\geq .5$ under the $N(0, 1)$, $U[-1, 1]$, and $C(0, 1)$ distributions, simultaneously.

Exercise I.111. Let $Z \sim N(0, 1)$. Evaluate $P(\Phi(Z)\Phi(-Z) > .1)$, where Φ denotes the standard normal CDF.

Exercise I.112. Suppose X has a normal distribution and $g(X)$ is a strictly increasing nonlinear function of X. Show that $g(X)$ cannot be normally distributed.

Exercise I.113. Suppose X has the Gamma density with parameters $\lambda = 1$, $\alpha = 2$. Find the expectation of the integer part and the fractional part of X.

Exercise I.114. Suppose $X_1, X_2, \ldots, X_n$ are n iid standard exponential variables. Find the mean, median, and variance of the minimum of $X_1, X_2, \ldots, X_n$.

Exercise I.115. Suppose X is uniformly distributed on $[0, 2\pi]$. Find $P(-.5 < \sin X < .5)$.

Exercise I.116. The diameter of a circular disk cut by a machine has the CDF $F(x) = \frac{(x-1)^3}{64}, 1 \le x \le 5$. Find the average diameter of disks coming from this machine.

Exercise I.117. Suppose $X \sim C(0, 1)$. Explicitly find a function $g(X)$ such that $Y = g(X) \sim Exp(1)$.

Exercise I.118. Let $f(x) = cx \sin x, 0 < x < \pi$.

(a) Evaluate a c that makes f a density function.
(b) Find the mean of this density function.

Exercise I.119. The waiting time at a teller's window in a bank has the density $f(x) = \frac{1}{3}e^{-x/3}, x > 0$.

(a) Find the average waiting time.
(b) Find the standard deviation of the waiting time.
(c) Find the probability that you have to wait longer than three minutes.
(d) Find a time such that the probability that you have to wait even longer than that time is only 5%.
(e) Find the probability that you have to wait at least three more minutes if you have already waited for three minutes.
(f) Interpret your result in part (e).

Exercise I.120. A square is to be constructed by choosing the common side length to be exponentially distributed with mean one inch. Find the expected area of the square.

Exercise I.121. A circle is to be constructed by choosing the radius of the circle to have the distribution of the absolute value of a standard normal. Find the expected perimeter of the circle.

Exercise I.122. A sphere is to be constructed by choosing the radius of the sphere such that it has a Beta distribution with both parameters equal to 3. Find the expected volume of the sphere.

Exercise I.123. Weights of individuals in some population are normally distributed with a mean of 150 lbs. and a standard deviation of 25 lbs. At least how many people must be sampled from this population if with a 90% probability we want at least one person in our sample who weighs more than 250 lbs.?

Exercise I.124. Suppose $X \sim N(0,1)$. For what values of a, b, c is $E(e^{aX^2+bX+c}) < \infty$?

Exercise I.125. Suppose $X \sim N(0,1)$. Find an expression for $P(|X| < 2a \mid |X| > a)$. Plot it as a function of a, and find the minima and the maxima.

Exercise I.126. Suppose $X \sim C(0,1)$. Find an expression for $P(|X| < 2a \mid |X| > a)$. Plot it as a function of a, and find the minima and the maxima.

Exercise I.127. Explicitly exhibit a density function $f(x)$ whose hazard rate has the bathtub shape; i.e., at first decreasing, then constant, and eventually increasing.

Exercise I.128. Suppose a positive continuous random variable has a finite mean. Write an expression for the mean in terms of the hazard rate function of the random variable.

Exercise I.129. $X_1, X_2, \ldots, X_{10}$ are ten iid $U[0,1]$ variables. Let m denote their minimum and M their maximum. Find $P(.05 < m < M < .95)$.

Exercise I.130. $X_1, X_2, \ldots, X_{10}$ are ten iid $N(0,1)$ variables. Let m denote their minimum and M their maximum. Find $P(-2 < m < M < 2)$.

Exercise I.131. Suppose X has a lognormal distribution with parameters $\mu = 0$, $\sigma = 1$. Find the deciles of X.

Exercise I.132. Suppose $X_1, X_2, \ldots, X_n$ are n independent lognormal variables. What is the name of the distribution of their product $X_1 X_2 \ldots X_n$?

Exercise I.133. The 25th, 50th, and 75th percentiles of a distribution are -1, 0, and 1.5. Can this be a normal distribution?

Exercise I.134. The 10th, 90th, and 95th percentiles of a distribution are 2, 5, and 8. Can this be a normal distribution?

Exercise I.135. Suppose we want to construct a confidence interval for a normal mean assuming that the variance σ^2 is known. What is the minimum n required for the margin of error of the confidence interval to be at most .1 if we want a 90% confidence interval? A 95% confidence interval? A 99% confidence interval? A 99.99% confidence interval?

Exercise I.136. Weights of adult males in some population are normally distributed with mean 160 lbs. and standard deviation 30 lbs. Weights of adult females in the same population are normally distributed with mean 130 lbs. and standard deviation 25 lbs. Find the probability that the weights of one randomly selected male and one randomly selected female differ by more than 50 lbs.

Exercise I.137. Suppose $Z \sim N(0,1)$. Find the mean, median, and mode of Z^5, $|Z|$, and $|Z-1|$.

Exercise I.138. A fair die is tossed 100 times. Approximate the probability that the sum of the rolls is between 300 and 400 inclusive. Next, suppose a fair die is tossed 1000 times. Approximate the probability that the sum of the rolls is between 3400 and 3600.

Exercise I.139. $X_1, X_2, \ldots, X_n$ are iid from a density $f(x)$ that equals $\frac{1}{3}$ on $[-1, 0]$ and equals $\frac{2}{3}$ on $[0, 1]$. Sketch the approximate density of $X_1 + X_2 + \cdots + X_n$ for $n = 50$.

Exercise I.140. Shipments of some equipment to a factory come in boxes of 1000 items. From past experience, the factory knows that (about) 1% of the items are defective. It returns a shipment if a sample of 50 items from the box contains two or more defective items.

(a) Approximate the probability that a shipment will be rejected.
(b) Suppose that on one occasion a bad shipment arrived with 5% defective items. Approximate the probability that the shipment will be rejected.

Exercise I.141. In approximately how many tosses of a fair coin is the probability of getting more than 52% heads at most .01?

Exercise I.142. In approximately how many tosses of a fair coin is the probability of getting more than 52% or less than 48% heads at most .01?

Exercise I.143. A certain congenital birth defect is found in some geographic region at the average rate of one a year. Approximate the probability that 60 or more people with this birth defect will be found in the next 50 years. State your assumptions.

Exercise I.144. A random variable X has the density $\frac{1}{x^2}$ for $x \geq 1$ and zero otherwise. An iid sample of size 100 is available from this density. Can we use a normal approximation to approximate the distribution of their sum? If so, sketch such an approximate normal density. If not, explain why we cannot do a normal approximation here.

Exercise I.145. A gambler repeatedly plays a game in which his earnings are iid $U[0, 1]$ in dollars. After each play, he tips the manager an amount equal to the square of the amount he just won. Approximate the probability that if he plays and tips 600 times, then his total winnings minus his total tip will exceed 105 dollars.

Exercise I.146. Suppose $X_1, X_2, \ldots, X_n$ are iid standard exponentials. For $n = 8$, sketch the exact density, the CLT approximation, and the first-order Edgeworth approximation for the density of their sum.

Exercise I.147. Suppose $X \sim Bin(n, p)$. For $n = 50, 100, 250$, plot the Berry-Esseen bound, as given in the text, as a function of p. Identify the peak value in the plot for each n.

Chapters 11–13

Exercise I.148. A fair die is rolled twice, and X and Y are the two rolls.

(a) Write the joint mass function of $X + Y$ and $\frac{X}{X+Y}$.
(b) From this, find the marginal pmf of $\frac{X}{X+Y}$.
(c) From this, find $E(\frac{X}{X+Y})$. Was the answer obvious to begin with?
(d) Find the conditional expectation of $\frac{X}{X+Y}$ given $X + Y = t, t = 2, 3, \ldots, 12$.
(e) By inspecting the numerical values in part (d), write a formula for $E(\frac{X}{X+Y} \,|X + Y = t)$. Was the answer obvious to begin with?

Exercise I.149. A fair coin is tossed 20 times. Let X be the number of heads in the first 15 tosses and Y the number of heads in the last 15 tosses. Find a formula for $E(Y \,|X = x)$.

Exercise I.150. Suppose X and Y are two random variables such that $E(Y|X) = X$. Assuming that the variances exist, prove that $\text{Var}(Y) \geq \text{Var}(X)$.

Exercise I.151. X, Y, Z have the joint pmf $p(x, y, z) = \frac{1}{8}$ for $x = \pm 1, y = \pm 1, z = \pm 1$.

(a) Find the marginal pmfs of each of X, Y, Z.
(b) Find the joint pmfs of each of $(X, Y), (Y, Z), (X, Z)$.
(c) Find the pairwise correlations between X and Y, Y and Z, and X and Z.
(d) Find the correlation between $X + Y$ and $Y + Z$.

Exercise I.152. A fair coin is tossed n times, and suppose X heads are obtained. Given $X = x$, a Poisson random variable Y with mean x is generated. Here, a Poisson with zero mean is the constant zero.

(a) Find the variance of the marginal distribution of Y.
(b) Evaluate the limit

$$\lim_{n\to\infty} P\left(|Y - \frac{n}{2}| > n^{\frac{3}{4}}\right).$$

Exercise I.153. Midterm grades in a class of 40 students are normally distributed with mean 50 and variance 100. The cutoffs for A, B, C, D are 70, 60, 40, 30, and a grade less than 30 is an F.

By recognizing it as a suitable multinomial distribution problem, calculate the probability that the number of students receiving each of the five letter grades is eight.

Exercise I.154. Suppose X, Y, Z are three independent Poisson variables with means λ, μ, η. Prove that the conditional distribution of (X, Y, Z) given $X + Y + Z = t$ is a trivariate multinomial distribution. Identify all the parameters of this multinomial distribution.

Exercise I.155. Suppose X has a discrete uniform distribution on $\{-n, -n+1, \ldots, 0, 1, \ldots, n-1, n\}$. Find the conditional expectation of X given $X^2 = t$ for a general t.

Exercise I.156. A fair coin is tossed repeatedly until the first head is obtained. Let X be the first toss at which the first head is obtained, and let $Y = \min(X, k)$, for a general $k \geq 1$.

(a) Find $E(Y)$.
(b) Find $E(Y|X = x)$.
(c) Find the correlation between X and Y in as simple a form as you can.
(d) Where does this correlation converge as $k \to \infty$?

Exercise I.157. From an urn with N balls numbered $1, 2, \ldots, N$, two balls are taken out without replacement. Let X, Y denote the numbers on the first ball chosen and the second ball chosen, respectively.

(a) Find $E(X), E(Y)$.
(b) Find $E(Y|X = n), n = 1, 2, \ldots, N$.
(c) Find $\text{Cov}(X, Y)$.
(d) Find the correlation between X and Y as a function of N.
(e) Compute the correlation for $N = 2, 3, 5, 10$.
(f) Find the limit of the correlation as $N \to \infty$. Is the answer what you would intuitively expect?

Exercise I.158. Let X be the number of kings and Y the number of hearts in a hand in bridge. Find the correlation between X and Y.

Exercise I.159. A fair die is rolled three times. Let X, Y, Z be the three individual rolls. Define $U = X, V = \max(X, Y), W = \max(X, Y, Z)$.

(a) Find $P(U = V)$.
(b) Find $P(V = W)$.
(c) Find $P(U = W)$.

Exercise I.160. Suppose (X_1, X_2, X_3) is jointly multinomially distributed with parameter vector (n, p_1, p_2, p_3). By using the joint mgf, find $E(X_1 X_2 X_3)$.

Exercise I.161. Suppose (X_1, X_2, X_3) is jointly multinomially distributed with parameter vector (n, p_1, p_2, p_3). Find the correlation between $X_1 + X_2$ and $X_2 + X_3$.

Exercise I.162. In bridge, find the conditional expectation of the number of aces in the hands of South given that North has k aces in his hand, $k = 0, 1, \ldots, 4$. Does your answer make intuitive sense?

Exercise I.163. Consider the joint density function

$$f(x, y) = cx^2 y^2, 0 < x, y; x + y < 1.$$

(a) Find the normalizing constant c.
(b) Find the marginal densities of X, Y.
(c) Prove or disprove that X and Y are independent.

Exercise I.164. Consider again the joint density function

$$f(x, y) = cx^2y^2, 0 < x, y; x + y < 1,$$

as in the problem above.

(a) Find a formula for $E(X \mid Y = y)$.
(b) Find a formula for $E(Y \mid X = x)$.
(c) Find $E(XY)$.
(d) Find $E(X^2Y^2)$.

Exercise I.165. Suppose X, Y are iid standard normal variables. Find

(a) $P(|X + Y| < |X - Y|)$.
(b) $E(XI_{\{Y<c\}})$.
(c) $E(XI_{\{\max(X,Y)<c\}})$.
(d) $P(X < Y < 2X)$.

Exercise I.166. Suppose X, Y are iid $U[0, 1]$. Find the density of $X - Y$.

Exercise I.167. A foot-long stick is broken at a random point, and then the longer of the two pieces is again broken at a random point. Find the probability that a triangle can be made with these three pieces.

Exercise I.168. X, Y, Z are iid $U[0, 1]$. Find the probability that the largest of the three is larger than the sum of the other two.

Exercise I.169. X, Y, Z are iid standard exponential. Find the joint density of (X, XY, XYZ).

Exercise I.170. X, Y, Z are iid $U[0, 1]$. Find the joint density of (X, XY, XYZ).

Exercise I.171. Suppose X, Y, Z are iid $Exp(1)$. Define $U_1 = \sqrt{X_1X_2}, U_2 = \sqrt{X_2X_3}, U_3 = \sqrt{X_1X_3}$.

(a) Find the mean and variance of each U_i.
(b) Let $T = \frac{U_1+U_2+U_3}{3}$. Find the mean and variance of T.

Remark. U_1, U_2, U_3 are not independent.

Exercise I.172. Suppose X_1, X_2 are iid standard normal variables. Show that:

(a) $X_1 + X_2$ and $X_1 - X_2$ are independent.
(b) $X_1 + X_2$ and $|X_1 - X_2|$ are independent.
(c) $X_1^2 + X_2^2$ and $\frac{X_1}{X_2}$ are independent.

Exercise I.173. Suppose (X, Y) has a bivariate normal distribution with means equal to zero, standard deviations equal to 1, and a correlation .5.

(a) Find the mean and variance of XY.
(b) Find the mean of X^2Y.
(c) Find the correlation between X and XY.
(d) Find a constant c such that $X + Y$ and $X + cY$ are independent.

Exercise I.174. The heights of husbands and wives in some population are jointly distributed as bivariate normal, with means 71 in. and 66 in. and standard deviations 2 in. and 1 in., respectively. Furthermore, the correlation between the heights of the husband and wife is .7. Find the probability that, for a randomly selected couple, the wife is taller than the husband.

Exercise I.175. Suppose (X, Y) is jointly uniformly distributed inside the unit circle in two dimensions.

(a) Find $P(X^2 + Y^2 < .5)$.
(b) For general $0 < r < s < 1$, find $P(r \le X^2 + Y^2 \le s)$.
(c) Find $E(e^{-X^2-Y^2})$.

Exercise I.176. Suppose X, Y are iid $U[0, 1]$. Let $U = \max(X, Y), V = \min(X, Y)$. Find $P(U > 2V)$.

Exercise I.177. Suppose X, Y are iid standard exponentials. Let $U = \max(X, Y)$, $V = \min(X, Y)$. Find $P(U > 2V)$.

Exercise I.178. Suppose X, Y are iid standard normal variables. Let $R = \frac{X}{Y}$ and $U = \sqrt{|R|}$. Is $E(U) < \infty$? If it is, find its value.

Exercise I.179. Suppose X, Y are iid random variables with the common density function $f(x) = \frac{c}{1+x^4}, -\infty < x < \infty$, where c is a normalizing constant. Show that $R = \frac{X}{Y}$ has the standard Cauchy distribution.

Exercise I.180. Let X be standard normal and Y independent of X.

(a) Show that the density of $X + Y$ is uniformly bounded. Give such an explicit bound.
(b) Is the density of XY necessarily uniformly bounded? Prove it, or give a counterexample.

Exercise I.181. Let X be standard normal and Y independent of X. Find the density of $X + Y$ for each of the following cases:

(a) $Y \sim Bin(2, .5)$.
(b) $Y \sim U[a, b]$.
(c) $Y \sim Exp(\lambda)$.
(d) $Y \sim Gamma(\alpha, \lambda)$ with $\alpha = 2$.

Exercise I.182. Suppose $X_1, X_2, \ldots$ are iid $U[0,1]$. Let $U = \sum_{i=1}^{\infty} \frac{X_i}{10^i}$ and $V = \sum_{i=1}^{\infty} \frac{X_i}{2^i}$. Find the expectation of $|U - V|$.

Exercise I.183. Suppose $X \sim Geo(p), Y \sim Geo(\theta)$, and that X and Y are independent. Find $P(X > Y)$.

Exercise I.184. A number N is chosen according to a Poisson distribution with mean 10. One hundred balls are then distributed completely at random into $N + 1$ cells. What are the mean and the variance of the number of balls received by the first cell?

Exercise I.185. A number N is chosen according to a Poisson distribution with mean 10. A fair coin is then tossed until $N + 1$ heads are obtained. What is the expected number of tosses it will take to stop the experiment?

I.2 True-False Problems

For each of the following questions, answer whether the statement is true (T) or false (F).

Chapters 1–4

1. A and B are two events such that $P(A) = .5, P(B) = .25$. Then, $P(A \cup B) \leq .75$.
2. A, B, C are three events such that $P(A) = P(B) = .5$, $P(A \cap B) = .25$, and if either A or B occurs, then C also occurs. Then, $P(C) < .75$.
3. A, B, C are three events such that A and B are independent, and if both A and B occur, then C cannot occur. Furthermore, $P(A) = P(B) = P(C) = .5$. Then, P(Either A and C both occur or B and C both occur) = .75.
4. Ten numbers are drawn without replacement from $1, 2, \ldots, 100$. The probability that the second number drawn will be an even number is .5.
5. The six letters in the word CHEESE are rearranged in a random manner. The probability that it will still spell CHEESE is less than .5.
6. Two calculus and two history books are placed on a shelf in random order. The probability that the two calculus books will be placed next to each other is less than .5.
7. A fair die is rolled three times. It is more likely that the sum will be 16 or more than that two or more of the rolls will be a six.
8. It is possible for the total number of events in an experiment with probabilities strictly between 0 and 1 to be 62.
9. In bridge, it is more likely that North has no spades than that he has no aces.
10. If three distinguishable balls are distributed completely at random into three distinguishable cells, then it is more likely that no cell will remain empty than that only one cell remains nonempty.

11. Tim chose one number at random from $1, 2, \ldots, 10$, and Tom chose one number at random from $1, 2, \ldots, 10$. They chose independently. The probability that they happened to choose the number is less than 5%.
12. If A and B are independent and B and C are also independent, then A, B, C are mutually independent.
13. If $P(A|B) = .5$, then for $P(B|A)$ also to be .5, $P(A)$ and $P(B)$ must both be .5.
14. $P(A|A^c \cap B)$ is always zero.
15. $P(A|A^c \cup B)$ is always zero.
16. If $P(A) > P(B)$, then $P(A|B) > P(B|A)$.
17. Among five people in a room, two are twins and the other three are three random people. The probability that there are three or more people in the room with the same birthday is less than 5%.
18. A fair die is rolled three times. The probability that at least two of the rolls are even if we know that at least one of the rolls is even is $\frac{2}{3}$.
19. If $P(A) = P(B) = .8$, then $P(B|A)$ cannot be .6.
20. If $P(A|B)$ and $P(B|C)$ are both strictly positive, then $P(A|C)$ is also strictly positive.
21. Tim and Doug shoot simultaneously at the bull's eye. Tim misses 80% of the time, and Doug hits 80% of the time. We know that one of the two shots hit the eye and the other missed. The probability that it was Doug who hit is .8.
22. A random variable X has a CDF $F(x)$ such that $F(x) - F(x-) = .2$ at $x = 1, 2, 3, 4, 5$. Then $F(2.5) = .4$.
23. A random variable X takes values 0, .5, and 1 and has mean .5. Then $P(X = 0)$ and $P(X = 1)$ are equal.
24. A discrete random variable X assumes the values $0, 1, 2, 3, \ldots$. Then, $E(X^2) = \sum_{n=0}^{\infty} P(X > \sqrt{n})$.
25. A fair coin is tossed 20 times. Then the expected number of times that a head is followed by four or more heads is larger than .25.
26. A couple wants to have at least two boys or at least two girls, whichever happens first. Then the expected number of children they will have is 2.5.
27. If X, Y, Z are independent random variables, then $\text{Var}(XYZ) \geq \text{Var}(X)\text{Var}(Y)\text{Var}(Z)$.
28. If X, Y, Z are independent random variables, then $\text{Var}(XYZ)$ cannot be equal to $\text{Var}(X)\text{Var}(Y)\text{Var}(Z)$.
29. An urn contains three green and three red balls. Four of them are taken out at random, without replacement, one at a time. Let X be the first draw at which a green ball is taken out. Then $E(X) < 3$.
30. A fair coin is tossed repeatedly until both a head and a tail are obtained. Let X be the number of tosses it will take. Then $E(X) = 3$.
31. A nonnegative random variable X has variance 100. Then $P(X > 20)$ cannot be zero.
32. It is not possible that neither X nor Y has a finite variance but $X + Y$ does.
33. It is not possible for a random variable X to be such that both $E(X)$ and $E(\frac{1}{X})$ are strictly larger than 1.

34. If $X_1, X_2, \ldots, X_{100}$ are 100 independent variables, and if $\text{Var}(X_1 + X_2 + \cdots + X_{100}) = 100$, then it cannot be true that $\text{Var}(X_i) < 1$ for each i.
35. X and Y are two random variables such that $E(X + Y) = 2$. Then at least one of $E(|X|)$ and $E(|Y|)$ must be ≥ 1.
36. A fair coin is tossed repeatedly until the first head is obtained. If we know that two tosses did not suffice, then the expected value of the number of tosses it actually took to obtain the first head is larger than 3.5.
37. X_1 and X_2 are iid random variables with the common pmf $p(x) = \frac{1}{2}$, $x = \pm 1$, and $p(x) = 0$ otherwise. If we define $X_3 = X_1 X_2$, then X_1, X_2, X_3 are mutually independent.
38. If X and Y are independent random variables with a finite variance, then necessarily $E(X^2 Y^2) = E(X^2)E(Y^2)$.
39. If X and Y are iid random variables with mean 1 and a finite and nonzero variance, then necessarily $E(X - Y)^2 > E(X - 1)^2$.
40. For any random variable X with a finite variance, $\text{Var}(|X|) \leq \text{Var}(X)$.
41. A random variable X has finite variance and another random variable Y takes only the values ± 1 with probability $\frac{1}{2}$ each; X and Y are independent. Then X and XY have the same variance.

Chapters 5–9

42. A nonnegative integer-valued random variable X has a finite mgf at some $t > 0$. Another random variable Y equals X if $X > 1$ but is zero if $X = 0$ or 1. Then Y also has a finite mgf at that t.
43. If X and Y are independent random variables and each has a finite mgfs for $-1 < t < 1$, then $X + Y$ and $X - Y$ also have finite mgfs for $-1 < t < 1$.
44. X, Y, Z are three iid random variables. Then XY and YZ are necessarily equal; i.e., $P(XY = YZ) = 1$.
45. X, Y, Z are three iid random variables. Then XY and YZ necessarily have the same distribution.
46. A certain positive random variable X does not have a finite mgf at any $t > 0$. However, $Y = Xe^{-X}$ must still have a finite mgf at all $t > 0$.
47. X is a standard normal variable. Then $Y = 2\Phi(x) - 1$ is distributed uniformly on $[-1, 1]$.
48. X is a Bernoulli random variable with parameter $p = .5$. Let $F(x)$ be the CDF of X. Then $2F(X) - 1$ is also a Bernoulli random variable.
49. X is a standard normal variable. Then no integer power of X can be normally distributed.
50. X is a standard Cauchy variable. Then no strictly monotone function of X can also be a standard Cauchy variable.
51. If X_1 and X_2 are iid random variables and all their moments exist, then all odd moments of $X_1 - X_2$ must also exist and be zero.
52. A continuous random variable X has all odd moments equal to zero. Then the density of X is symmetric about zero.

53. X has a Poisson distribution. Then no function of X can be normally distributed.
54. X and Y are independent Poisson random variables. Then $\max\{X, Y\}$ is also Poisson distributed.
55. X and Y are independent Poisson random variables. Then $\min\{X, Y\}$ is also Poisson distributed.
56. X and Y are iid Poisson random variables with mean 1. Then $\frac{X+Y}{2}$ is also Poisson with mean 1.
57. If X and Y are independent continuous random variables and each has a density symmetric about zero, then $X + Y$ also has a density symmetric about zero.
58. If X and Y are independent continuous random variables and each has a density symmetric about zero, then XY also has a density symmetric about zero.
59. If a continuous random variable X has zero mean, then its density $f(x)$ has to be strictly positive at zero.
60. If a continuous random variable X has zero mean, then its density $f(x)$ has to be finite at zero.
61. A continuous random variable X has a density symmetric about zero, and X^2 has a chi-square distribution with one degree of freedom. Then X must be standard normal.
62. If X has a Pareto distribution with $\theta = 1$, then $\frac{1}{X}$ has a Beta distribution.
63. The variance of a Beta distribution cannot be 2.
64. If X has a lognormal distribution, then $E(\frac{1}{X})$ must exist.
65. If X, Y, Z are three independent lognormal variables, then XY^2Z^3 is another lognormal variable.
66. If X, Y, Z, W are four iid standard normal variables, then $\frac{X}{Y} + \frac{Z}{W}$ is a Cauchy variable.
67. If X has a standard double exponential density, then $|X|$ has an exponential density.
68. If X is a positive random variable and $E(X^2) = E(X^6) = 1$, then X is constantly equal to 1.
69. If $f(x)$ is a density function on $[0, 1]$, then $\int_0^1 f^2(x)dx < \infty$.
70. If $f(x)$ is a density function on $[0, 1]$, then $\int_0^1 \sqrt{f(x)}dx < \infty$.
71. If X is a positive random variable and $E[g(X)] < \infty$, then $E[g(-X)]$ must also be finite.

Chapter 10

72. The sum of 50 independent Poisson variables with mean 1 and the sum of 50 independent exponential variables with mean 1 have approximately the same distribution.
73. One hundred numbers are chosen at random independently with replacement from $1, 2, \ldots, 9$. Their sum should be 500 ± 50 with about a 95% probability.

74. One hundred numbers are chosen independently from the unit interval $[0, 1]$ according to a uniform distribution. Their sum should be 50 ± 5 with about a 92% probability.
75. If a fair coin is tossed 500 times, the probability that exactly 250 heads will be obtained is about 4%.
76. If a fair coin is tossed 5000 times, the probability that exactly 2500 heads will be obtained is about 1%.
77. The length of an approximate 95% confidence interval for a Poisson mean increases with the data value X.
78. The center of an approximate 95% confidence interval for a Poisson mean moves to the right with the data value X.
79. The sum of the squares of 90 iid $U[0, 1]$ variables should be approximately normal with mean 30 and variance 7.5.
80. The sum of the squares of 90 iid $U[0, 1]$ variables should be approximately normal with mean 30 and variance 8.
81. X is a Poisson variable with mean λ. If $P(X \leq 10) \approx .95$, then $\lambda \approx 6$.

Chapters 11–13

82. If X and Y are discrete random variables with the joint pmf $p(x, y) = \frac{1}{9}$, $1 \leq x \leq 3, 1 \leq y \leq 3$, then X and Y are independent random variables.
83. If X and Y are discrete random variables with the joint pmf $p(x, y) = \frac{1}{6}$, $1 \leq x \leq y \leq 3$, then $E(Y - X) > 0$.
84. A fair coin is tossed eight times. X is the number of heads in the first four tosses and Y the number of tails in the last four tosses. Then, $E[(Y - 2)^2 \mid X = 2] > 1$.
85. Given a positive random variable X, let $Y = e^{X \log X}$. Then $E(Y \mid X = 1) = 1$.
86. Given a positive random variable X, let $Y = e^{X \log X}$. Then $\text{Var}(Y \mid X = 1) > 1$.
87. X and Y are independent random variables and $E(Y) = 0$. Then $E(XY \mid X = 1) = 0$.
88. It is not possible that $\text{Var}(Y) > 0$ but $\text{Var}(Y \mid X = x) = 0$ for some particular x and some particular random variable X.
89. If the correlation between X and Y is strictly positive, then the correlation between X^2 and Y is also strictly positive.
90. Always, $\text{Var}(X) \geq E_Y[\text{Var}(X \mid Y = y)]$.
91. If $E(X \mid Y = y)$ exists for every y, then $E(X)$ also exists.
92. If X and Y are independent, then the correlation between $\sin X$ and $\cos Y$ is zero.
93. If 50 balls are distributed independently and with equal probability into ten cells, then the correlation between the number of balls that are allocated to the first cell and the number of balls that are allocated to the tenth cell is $< -.1$.
94. A fair die is rolled repeatedly. X is the first roll where a five is obtained, and Y is the first roll where a six is obtained. Then $E(Y \mid X = x) = x$.

95. If X and Y are two random variables with finite variances, then $\sigma_{X+Y} \leq \sigma_X + \sigma_Y$.
96. If X, Y, Z are three random variables with finite variances, then $\sigma_{X+Y+Z} \leq \sigma_X + \sigma_Y + \sigma_Z$.
97. A fair die is rolled repeatedly. X is the first roll where a six is obtained, and Y is the roll where the second six is obtained. Then $E(Y \mid X = 4) = 10$.
98. If X and Y are continuous random variables with joint density $f(x, y) = 2$, $x, y \geq 0, x + y \leq 1$, then marginally X and Y are both $U[0, 1]$.
99. If X and Y are both marginally $U[0, 1]$, then the joint density must be $f(x, y) = 1, x, y \in [0, 1]$.
100. If X and Y have a joint uniform density in the unit circle $C = \{(x, y) : x^2 + y^2 \leq 1\}$, then each of $E(X)$, $E(Y)$, $E(XY)$, and $E(XY^2)$ is zero.
101. One thousand observations are generated independently according to a uniform distribution in the ten dimensional unit cube. The number of observations among these 1000 observations that fall inside the inscribed sphere of the cube has an expected value of only about 2.
102. If X and Y have a joint uniform density in the unit circle $C = \{(x, y) : x^2 + y^2 \leq 1\}$, then $P(X^2 + Y^2 \leq .5) = .5$.
103. If X and Y are iid $U[0, 1]$ random variables, then $E[\min\{X, Y\}] = 1 - E[\max\{X, Y\}]$.
104. Whatever the joint distribution of two positive random variables X and Y, if $E(X) = 2$ and $E(Y) = 1$, then $E(|X - Y|) \geq 1$.
105. $X \sim N(\mu, \sigma^2)$. Let $Y = I_{\{X>0\}}$. Then $|X|, Y$ are independent if and only if $\mu = 0$.
106. If X and Y are random variables such that $\frac{X}{Y}$ has a standard Cauchy distribution, then X and Y must be independent standard normal.
107. If $X_1, \ldots, X_n$ are iid $U[0, 1]$ random variables and $X_{(1)}$ and $X_{(n)}$ are the smallest and the largest order statistics, then $\rho_{X_{(1)}, X_{(n)}} \to 0$ as $n \to \infty$.
108. If $X_1, \ldots, X_n$ are iid $U[0, 1]$ random variables and $X_{(1)}$ and $X_{(n)}$ are the smallest and the largest order statistics, then $P(X_{(n)} - X_{(1)} > .99) \to 1$ as $n \to \infty$.
109. If $X_1, \ldots, X_n$ are iid $U[0, 1]$ random variables and $X_{(1)}$ and $X_{(n)}$ are the smallest and the largest order statistics, then $X_{(n)} + X_{(1)}$ and $X_{(n)} - X_{(1)}$ are uncorrelated.
110. If $X_1, \ldots, X_n$ are iid $U[0, 1]$ random variables and $X_{(1)}$ and $X_{(n)}$ are the smallest and the largest order statistics, then $X_{(n)} + X_{(1)}$ and $X_{(n)} - X_{(1)}$ are independent.
111. If $X_1, \ldots, X_5$ are iid standard normal variables, then $E[X_{(5)} + X_{(4)} + X_{(3)} + X_{(2)} + X_{(1)}] = 0$.
112. If $X_1, \ldots, X_n$ are iid $U[0, 1]$ random variables, then the density of $X_{(i)}$ is unimodal for any $i, 1 \leq i \leq n$.
113. If $X_1, \ldots, X_n$ are iid standard exponential variables, then $\frac{E(X_{(n)})}{\log n} \to 1$ as $n \to \infty$.
114. If X and Y are jointly bivariate normal with marginal variance 1 and $\text{Var}(X \mid Y = 0) = .36$, then $\rho_{X,Y} = +.8$.

115. If X and Y are jointly bivariate normal, then $\frac{d^2}{dx^2} E(Y \mid X = x) = 0$ at any x.
116. If X has a t distribution, then X^2 has an F distribution.
117. If X and Y are iid standard exponentials, than $\text{Var}(\frac{X}{X+Y}) < .1$.
118. If X and Y are iid standard exponentials, than $E(X + Y \mid \frac{X}{X+Y} = .5) = 2$.
119. If X and Y are iid standard normal, then $P(\frac{X}{Y} < 1) = P(X < Y)$.
120. If X and Y have the joint density $f(x, y) = \frac{1}{4} e^{-|x|-|y|}, x, y \in \mathcal{R}$, then the polar coordinates r, θ are not independent.
121. If X and Y have the joint density $f(x, y) = \frac{c}{(1+x^2+y^2)^{5/2}}, x, y \in \mathcal{R}$, where c is a normalizing constant, then the polar coordinates r, θ are independent.

Appendix II: Symbols and Formulas

II.1 Glossary of Symbols

$n!$	$n(n-1)\cdots 1$
$\binom{n}{k}$	$\frac{n!}{k!(n-k)!}$
$a_n \sim b_n$	$0 < \liminf \frac{a_n}{b_n} \leq \limsup \frac{a_n}{b_n} < \infty$
$a_n = O(b_n)$	$\lvert a_n \rvert \leq K b_n$ for some finite positive constant K
$a_n = o(b_n)$	$\lim \frac{a_n}{b_n} = 0$
$a_n \approx b_n$	$\lim \frac{a_n}{b_n} = 1$
$\mathcal{R}$	real line
$\mathcal{R}^d$	d-dimensional Euclidean space
f'	first derivative of f
f''	second derivative of f
$\frac{\partial}{\partial x}$	partial derivative
$\Gamma(\alpha)$	Gamma function
$B(\alpha, \beta)$	Beta function
${}_1F_1$	hypergeometric function
I_z	Bessel function
H_j	Hermite polynomials
log	natural logarithm
$\log_b$	Logarithm to the base b
$\lfloor x \rfloor$	Integer part
$\{\}$	fractional part
ω	sample point
Ω	sample space
$P(A)$	probability of A
$P(A \mid B)$	conditional probability of A given B
A^c	complement of A
$\cup_{i=1}^n A_i$	union of $A_1, \cdots, A_n$
$\cap_{i=1}^n A_i$	intersection of $A_1, \cdots, A_n$
$G_X(s), G(s)$	generating function
$\psi_X(t), \psi(t)$	moment generating function

iid	independent and identically distributed
A, B, C, D	events
X, Y, Z, U, V, W	random variables
I_A	indicator function of A
sgn, sign	signum function
x_+, x^+	$\max\{x, 0\}$
max, min	maximum, minimum
sup, inf	supremum, infimum
$MN(n, p_1, \ldots, p_k)$	multinomial distribution with these parameters
$N(\mu, \sigma^2)$	normal distribution
$t_n, t_{(n)}$	Student's t distribution with n degrees of freedom
$Ber(p)$, $Bin(n, p)$	Bernoulli and binomial distributions
$Poi(\lambda)$	Poisson distribution
$Geo(p)$	geometric distribution
$NB(r, p)$	negative binomial distribution
$Hypergeom(n, D, N)$	hypergeometric distribution
$Exp(\lambda)$	exponential distribution with mean λ
$Gamma(\alpha, \lambda)$	Gamma distribution with shape parameter α and scale parameter λ
$\chi^2_n, \chi^2_{(n)}$	chi-square distribution
$C(\mu, \sigma)$	Cauchy distribution
$U[a, b]$	uniform distribution
$Be(\alpha, \beta)$	Beta distribution
$Pa(\theta, \alpha)$	Pareto distribution
$\phi(x)$	standard normal density
$\Phi(x)$	standard normal CDF
$R(x)$	Mills ratio
$f(x)$	general density
$F(x)$	general CDF
$\bar{F}(x)$	survival function
$F^{-1}(p)$, $Q(p)$	quantile function
$p(x)$	general pmf
$p(x, y)$	bivariate pmf
$f(x, y)$	bivariate density
$p(x_1, \cdots, x_n)$	multivariate pmf
$f(x_1, \cdots, x_n)$	multidimensional density
$F(x_1, \cdots, x_n)$	multidimensional CDF
$p(x\|y)$	conditional pmf
$f(x\|y)$	conditional density
f_X, f_Y	marginal densities
F_X, F_Y	marginal cdfs
$F_{X\|Y}$	conditional CDF
$E(X), \mu$	expected value
Var, σ^2	variance

μ_k	$E(X-\mu)^k$
κ_r	rth cumulant
β	skewness
γ	kurtosis
Cov	covariance
ρ	correlation
r, θ	polar coordinates in two dimensions
J	Jacobian matrix
$\|J\|$	determinant of J
$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$	order statistics
W_n	sample range
p_{ij}	transition probabilities in a Markov chain
P	one-step transition probability matrix
$P^{(n)}$	n-step transition probability matrix
S	state space of a Markov chain
T_i, T_{ij}, T_{iD}	first-passage times in a Markov chain
π	stationary distribution of a Markov chain
$x_{(n)}$	$x(x-1)\cdots(x-n+1)$
$s(n,k)$	Stirling numbers of the first kind
$S(n,k)$	Stirling numbers of the second kind

II.2 Formula Summaries

II.2.1 Moments and MGFs of Common Distributions

Discrete Distributions

Distribution	$p(x)$	Mean	Variance	Skewness	Kurtosis	MGF
Uniform	$\frac{1}{n}, x = 1, \cdots, n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	0	$-\frac{6(n^2+1)}{5(n^2-1)}$	$\frac{e^{(n+1)t}-e^t}{n(e^t-1)}$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}, x = 0, \cdots, n$	np	$np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{1-6p(1-p)}{np(1-p)}$	$(pe^t + 1 - p)^n$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, \cdots$	λ	λ	$\frac{1}{\sqrt{\lambda}}$	$\frac{1}{\lambda}$	$e^{\lambda(e^t-1)}$
Geometric	$p(1-p)^{x-1}, x = 1, 2, \cdots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{2-p}{\sqrt{1-p}}$	$6 + \frac{p^2}{1-p}$	$\frac{pe^t}{1-(1-p)e^t}$
Negative Binomial	$\binom{x-1}{r-1} p^r (1-p)^{x-r}, x \geq r,$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\frac{2-p}{\sqrt{r(1-p)}}$	$\frac{6}{r} + \frac{p^2}{r(1-p)}$	$\left(\frac{pe^t}{1-(1-p)e^t}\right)^r$
Hypergeometric	$\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$	$n\frac{D}{N}$	$n\frac{D}{N}(1-\frac{D}{N})\frac{N-n}{N-1}$	Complex	Complex	Complex
Benford	$\frac{\log(1+\frac{1}{x})}{\log 10}, x = 1, \cdots, 9$	3.44	6.057	.796	2.45	$\sum_{x=1}^{9} e^{tx} p(x)$

Continuous Distributions

Distribution	$f(x)$	Mean	Variance	Skewness	Kurtosis
Uniform	$\frac{1}{b-a}, a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	0	$-\frac{6}{5}$
Exponential	$\frac{e^{-x/\lambda}}{\lambda}, x \ge 0$	λ	λ^2	2	6
Gamma	$\frac{e^{-x/\lambda}x^{\alpha-1}}{\lambda^\alpha \Gamma(\alpha)}, x \ge 0$	$\alpha\lambda$	$\alpha\lambda^2$	$\frac{2}{\sqrt{\alpha}}$	$\frac{6}{\alpha}$
χ^2_m	$\frac{e^{-x/2}x^{m/2-1}}{2^{m/2}\Gamma\left(\frac{m}{2}\right)}, x \ge 0$	m	$2m$	$\sqrt{\frac{8}{m}}$	$\frac{12}{m}$
Weibull	$\frac{\beta}{\lambda}(\frac{x}{\lambda})^{\beta-1}e^{-(\frac{x}{\lambda})^\beta}, x > 0$	$\lambda\Gamma(1+\frac{1}{\beta})$	$\lambda^2\Gamma(1+\frac{2}{\beta})-\mu^2$	$\frac{\lambda^3\Gamma\left(1+\frac{3}{\beta}\right)-3\mu\sigma^2-\mu^3}{\sigma^3}$	Complex
Beta	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, 0 \le x \le 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{\sqrt{\alpha\beta}(\alpha+\beta+2)}$	Complex
Normal	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}, x \in \mathcal{R}$	μ	σ^2	0	0
lognormal	$\frac{1}{\sigma\sqrt{2\pi}x}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}, x > 0$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$	$e^{\sigma^2}+2\sqrt{e^{\sigma^2}-1}$	Complex
Cauchy	$\frac{1}{\sigma\pi(1+(x-\mu)^2/\sigma^2)}, x \in \mathcal{R}$	None	None	None	None
t_m	$\frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\Gamma\left(\frac{m}{2}\right)}\frac{1}{(1+x^2/m)^{(m+1)/2}}, x \in \mathcal{R}$	$0\ (m > 1)$	$\frac{m}{m-2}(m > 2)$	$0\ (m > 3)$	$\frac{6}{m-4}(m > 4)$
F	$\frac{\left(\frac{\beta}{\alpha}\right)^\beta x^{\alpha-1}}{B(\alpha,\beta)\left(x+\frac{\beta}{\alpha}\right)^{\alpha+\beta}}, x > 0$	$\frac{\beta}{\beta-1}(\beta > 1)$	$\frac{\beta^2(\alpha+\beta-1)}{\alpha(\beta-2)(\beta-1)^2}(\beta > 2)$	Complex	Complex
Double Exponential	$\frac{e^{-\lvert x-\mu\rvert/\sigma}}{2\sigma}, x \in \mathcal{R}$	μ	$2\sigma^2$	0	3
Pareto	$\frac{\alpha\theta^\alpha}{x^{\alpha+1}}, x \ge \theta > 0$	$\frac{\alpha\theta}{\alpha-1}(\alpha > 1)$	$\frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)}(\alpha > 2)$	$\frac{2(\alpha+1)}{\alpha-3}\sqrt{\frac{\alpha-2}{\alpha}}(\alpha > 3)$	Complex
Gumbel	$\frac{1}{\sigma}e^{\left(-e^{-\frac{x-\mu}{\sigma}}\right)}e^{-\frac{x-\mu}{\sigma}}, x \in \mathcal{R}$	$\mu+\gamma\sigma$	$\frac{\pi^2}{6}\sigma^2$	$\frac{12\sqrt{6}\zeta(3)}{\pi^3}$	$\frac{12}{5}$

Note: For the Gumbel distribution, $\gamma \approx .577216$ is the Euler constant and $\zeta(3)$ is Riemann's zeta function $\zeta(3) = \sum_{n=1}^{\infty}\frac{1}{n^3} \approx 1.20206$.

Table of MGFs of Continuous Distributions

Distribution	$f(x)$	MGF
Uniform	$\frac{1}{b-a}, a \le x \le b$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Exponential	$\frac{e^{-x/\lambda}}{\lambda}, x \ge 0$	$(1-\lambda t)^{-1} (t < 1/\lambda)$
Gamma	$\frac{e^{-x/\lambda} x^{\alpha-1}}{\lambda^\alpha \Gamma(\alpha)}, x \ge 0$	$(1-\lambda t)^{-\alpha} (t < 1/\lambda)$
χ^2_m	$\frac{e^{-x/2} x^{m/2-1}}{2^{m/2}\Gamma(\frac{m}{2})}, x \ge 0$	$(1-2t)^{-m/2} (t < \frac{1}{2})$
Weibull	$\frac{\beta}{\lambda}(\frac{x}{\lambda})^{\beta-1} e^{-(\frac{x}{\lambda})^\beta}, x > 0$	$\sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} \Gamma(1+\frac{n}{\beta})$
Beta	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, 0 \le x \le 1$	${}^1F_1(\alpha, \alpha+\beta, t)$
Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, x \in \mathcal{R}$	$e^{t\mu + t^2\sigma^2/2}$
lognormal	$\frac{1}{\sigma\sqrt{2\pi}x} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}, x > 0$	None
Cauchy	$\frac{1}{\sigma\pi(1+(x-\mu)^2/\sigma^2)}, x \in \mathcal{R}$	None
t_m	$\frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})} \frac{1}{(1+x^2/m)^{(m+1)/2}}, x \in \mathcal{R}$	None
F	$\frac{(\frac{\beta}{\alpha})^\beta x^{\alpha-1}}{B(\alpha,\beta)(x+\frac{\beta}{\alpha})^{\alpha+\beta}}, x > 0$	None
Double Exponential	$\frac{e^{-\lvert x-\mu\rvert/\sigma}}{2\sigma}, x \in \mathcal{R}$	$\frac{e^{t\mu}}{1-\sigma^2t^2} (\lvert t\rvert < 1/\sigma)$
Pareto	$\frac{\alpha\theta^\alpha}{x^{\alpha+1}}, x \ge \theta > 0$	None
Gumbel	$\frac{1}{\sigma} e^{(-e^{-\frac{x-\mu}{\sigma}})} e^{-\frac{x-\mu}{\sigma}}, x \in \mathcal{R}$	$e^{t\mu}\Gamma(1-t\sigma)\ (t < 1/\sigma)$

II.2.2 Useful Mathematical Formulas

$1+2+\cdots+n=\frac{n(n+1)}{2}$;
$1^2+2^2+\cdots+n^2=\frac{n(n+1)(2n+1)}{6}$;
$1^3+2^3+\cdots+n^3=\left(\frac{n(n+1)}{2}\right)^2$;
$1^4+2^4+\cdots+n^4=\frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$;
$\binom{n}{0}+\binom{n}{1}+\cdots+\binom{n}{n}=2^n$;
$\binom{n}{0}+\binom{n}{2}+\binom{n}{4}+\cdots=\binom{n}{1}+\binom{n}{3}+\binom{n}{5}+\cdots=2^{n-1}$;
$\binom{n}{0}-\binom{n}{1}+\binom{n}{2}-\binom{n}{3}+\cdots=0$;
$\binom{n}{0}^2+\binom{n}{1}^2+\cdots+\binom{n}{n}^2=\binom{2n}{n}$;
$(a+b)^n=a^n+\binom{n}{1}a^{n-1}b+\binom{n}{2}a^{n-2}b^2+\cdots+b^n$;
$1+x+x^2+\cdots+x^n=\frac{1-x^{n+1}}{1-x}$;
$1+x+x^2+x^3+\cdots=\frac{1}{1-x},-1<x<1$;
$x+2x^2+3x^3+4x^4+\cdots=\frac{x}{(1-x)^2},-1<x<1$;
$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots$;
$\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\cdots,-1<x\le 1$;
$\frac{1}{1-x}=1+x+x^2+x^3+\cdots,-1<x<1$;
$\sum_{n=1}^{\infty}\frac{1}{n}=\infty$;
$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}$;
$\sum_{n=1}^{\infty}\frac{1}{n(n+1)}=1$;
$\lim_{n\to\infty}[1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}-\log n]=\gamma$(Euler's constant);
$n!\approx e^{-n}n^{n+\frac{1}{2}}\sqrt{2\pi},n\to\infty$(Stirling's approximation);
$\arcsin x=\arccos\sqrt{1-x^2};\arccos x=\arcsin\sqrt{1-x^2}$;
$\arctan x=\arcsin\frac{x}{\sqrt{1+x^2}}$;
$\arctan x+\arctan y=\pi-\arctan\frac{x+y}{xy-1},x,y>0,xy>1$;
$\sin 2x=2\sin x\cos x;\sin 3x=3\sin x-4\sin^3 x$;
$\cos 2x=2\cos^2 x-1=\cos^2 x-\sin^2 x;\cos 3x=4\cos^3 x-3\cos x$;
$\tan 2x=\frac{2\tan x}{1-\tan^2 x};\tan\frac{x}{2}=\frac{\sin x}{1+\cos x}$;
$\Gamma(\alpha)=\int_0^{\infty}e^{-x}x^{\alpha-1}dx,\alpha>0;Be(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$;
$\Gamma(n)=(n-1)!,n=1,2,3,\cdots;\Gamma(x)=x\Gamma(x-1),x>1;\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$;
$\Gamma(2x)=\frac{2^{2x-1}\Gamma(x)\Gamma(x+\frac{1}{2})}{\sqrt{\pi}}$(Gamma duplication formula);
area of triangle $=\sqrt{s(s-a)(s-b)(s-c)},s=\frac{a+b+c}{2},a,b,c$ the side lengths;
area of circle $=\pi r^2,r$ the radius; volume of sphere in three dimensions $=\frac{4}{3}\pi r^3$;
volume of unit sphere in n dimensions $=V_n=\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$; surface area of unit sphere in n dimensions $=nV_n$;
volume of circular cylinder $=\pi r^2h$; volume of circular cone $=\frac{1}{3}\pi r^2h$.

II.2.3 Useful Calculus Facts

$(fg)' = f'g + fg'$; $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$; $\left(\frac{1}{f}\right)' = -\frac{f'}{f^2}$; $(\log f)' = \frac{f'}{f}$;
$(e^f)' = f'e^f$; $(f \circ g)' = f'(g)g'$(chain rule); $(fg)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(k)} g^{(n-k)}$;
$(\int_a^x f(t)dt)' = f(x)$; $(\int_x^a f(t)dt)' = -f(x)$.

Basic derivatives and indefinite integrals

$f(x)$	Derivative	Indefinite integral
$x^a, a \neq -1$	ax^{a-1}	$\frac{x^{a+1}}{a+1}$
$\frac{1}{x}$	$-\frac{1}{x^2}$	$\log \lvert x \rvert$
$\log x$	$\frac{1}{x}$	$x \log x - x$
e^{tx}	te^{tx}	$\frac{e^{tx}}{t}$
xe^{tx}	$(1 + tx)e^{tx}$	$\frac{e^{tx}}{t^2}(tx - 1)$
$\sin ax$	$a \cos ax$	$-\frac{1}{a} \cos ax$
$\cos ax$	$-a \sin ax$	$\frac{1}{a} \sin ax$
$x \sin ax$	$ax \cos ax + \sin ax$	$\frac{1}{a^2}[\sin ax - ax \cos ax]$
$\arcsin x$	$\frac{1}{\sqrt{1-x^2}}$	$x \arcsin x + \sqrt{1-x^2}$
$\arccos x$	$-\frac{1}{\sqrt{1-x^2}}$	$x \arccos x - \sqrt{1-x^2}$
$\frac{1}{a^2x^2+c^2}$	$-\frac{2a^2x}{(a^2x^2+c^2)^2}$	$\frac{1}{ac} \arctan \frac{ax}{c}$
$\frac{x}{a^2x^2+c^2}$	$\frac{c^2-a^2x^2}{(a^2x^2+c^2)^2}$	$\frac{1}{2a^2} \log \lvert a^2x^2 + c^2 \rvert$

II.3 Tables

II.3.1 Normal Table

Standard normal probabilities $P(Z \leq t)$ and standard normal percentiles

Quantity tabulated on the next page is $\Phi(t) = P(Z \leq t)$ for a given $t \geq 0$, where $Z \sim N(0, 1)$. For example, from the table, $P(Z \leq 1.52) = .9357$. For any positive t, $P(-t \leq Z \leq t) = 2\Phi(t) - 1$ and $P(Z < -t) = P(Z > t) = 1 - \Phi(t)$. Selected standard normal percentiles z_α are given below. Here, the meaning of z_α is $P(Z > z_\alpha) = \alpha$.

α	z_α
.25	.675
.2	.84
.1	1.28
.05	1.645
.025	1.96
.02	2.055
.01	2.33
.005	2.575
.001	3.08
.0001	3.72

	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

II.3.2 Poisson Table

The table below tabulates $P(X \leq x)$ for a given x when $X \sim Poi(\lambda)$. For example, if $\lambda = 2.5$, then $P(X \leq 4) = .8912$. The individual probabilities are found by subtraction. Thus, if $\lambda = 2.5$, then $P(X = 4) = P(X \leq 4) - P(X \leq 3) = .8912 - .7576 = .1336$.

	λ									
x	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
0	0.6065	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067
1	0.9098	0.7358	0.5578	0.4060	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404
2	0.9856	0.9197	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1736	0.1247
3	0.9982	0.9810	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.2650
4	0.9998	0.9963	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405
5	1.0000	0.9994	0.9955	0.9834	0.9580	0.9161	0.8576	0.7851	0.7029	0.6160
6	1.0000	0.9999	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622
7	1.0000	1.0000	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666
8	1.0000	1.0000	1.0000	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319
9	1.0000	1.0000	1.0000	1.0000	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682
10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9972	0.9933	0.9863
11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991	0.9976	0.9945
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Author Index

Subject Index