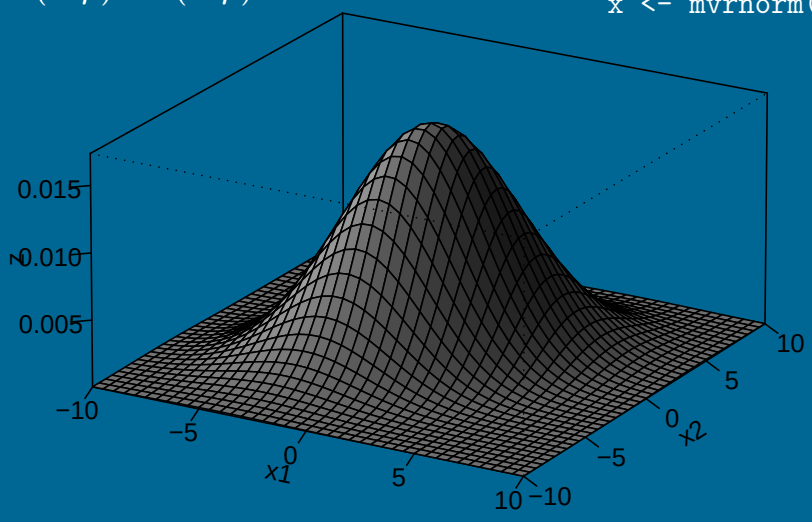
From Algorithms to Z-Scores:

Probabilistic and Statistical Modeling in Computer Science

Norm Matloff, University of California, Davis

$$f_X(t) = ce^{-0.5(t-\mu)'\Sigma^{-1}(t-\mu)}$$

```
library(MASS)
x <- mvrnorm(mu,sgm)
```



Contents

Preface

Why is this book different from all other books on mathematical probability and statistics? The key aspect is the book's consistently *applied* approach, especially important for engineering students.

The applied nature comes is manifested in a number of senses. First, there is a strong emphasis on intution, with less mathematical formalism. In my experience, defining probability via sample spaces, the standard approach, is a major impediment to doing good applied work. The same holds for defining expected value as a weighted average. Instead, I use the intuitive, informal approach of long-run frequency and long-run average. I believe this is especially helpful when explaining conditional probability and expectation, concepts that students tend to have trouble with. (They often think they understand until they actually have to work a problem using the concepts.)

On the other hand, in spite of the relative lack of formalism, all models and so on are described precisely in terms of random variables and distributions. And the material is actually somewhat more mathematical than most at this level in the sense that it makes extensive usage of linear algebra.

Second, the book stresses *real-world* applications. Many similar texts, notably the elegant and interesting book for computer science students by Mitzenmacher, focus on probability, in fact discrete probability. Their intended class of "applications" is the theoretical analysis of algorithms. I instead focus on the actual use of the material in the real world; which tends to be more continuous than discrete, and more in the realm of statistics than probability. This should prove especially valuable, as "big data" and machine learning now play a significant role in applications of computers.

Third, there is a strong emphasis on modeling. Considerable emphasis is placed on questions such as: What do probabilistic models really mean, in real-life terms? How does one choose a model? How do we assess the practical usefulness of models? This aspect is so important that there is a separate chapter for this, titled Introduction to Model Building. Throughout the text, there is considerable discussion of the real-world meaning of probabilistic concepts. For instance, when probability density functions are introduced, there is an extended discussion regarding the intuitive meaning of densities in light of the inherently-discrete nature of real data, due to the finite precision of measurement.

Finally, the R statistical/data analysis language is used throughout. Again, several excellent texts on probability and statistics have been written that feature R, but this book, by virtue of having a computer science audience, uses R in a more sophisticated manner. My open source tutorial on R programming, *R for Programmers* (`http://heather.cs.ucdavis.edu/~matloff/R/RProg.pdf`), can be used as a supplement. (More advanced R programming is covered in my book, *The Art of R Programming*, No Starch Press, 2011.)

There is a large amount of material here. For my one-quarter undergraduate course, I usually cover Chapters 2, 3, 6, 7, 8, 10, 16, 19, 20, 21, 22 and 27. My lecture style is conversational, referring to material in the book and making lots of supplementary remarks ("What if we changed the assumption here to such-and-such?" etc.). Students read the details on their own. For my one-quarter graduate course, I cover Chapters 10, 11, 24, 12, 14, 25, 26, 27, 28, 29 and 30.

As prerequisites, the student must know calculus, basic matrix algebra, and have some skill in programming. As with any text in probability and statistics, it is also necessary that the student has a good sense of math intuition, and does not treat mathematics as simply memorization of formulas.

The LaTeXsource **.tex** files for this book are in `http://heather.cs.ucdavis.edu/~matloff/132/PLN`, so readers can copy the R code and experiment with it. (It is not recommanded to copy-and-paste from the PDF file, as hidden characters may be copied.) The PDF file is searchable.

The following, among many, provided valuable feedback for which I am very grateful: Ibrahim Ahmed; Ahmed Ahmedin; Stuart Ambler; Earl Barr; Benjamin Beasley; Matthew Butner; Michael Clifford; Dipak Ghosal; Noah Gift; Laura Matloff; Nelson Max, Connie Nguyen, Jack Norman, Richard Oehrle, Michael Rea, Sana Vaziri, Yingkang Xie, and Ivana Zetko. The cover picture, by the way, is inspired by an example in Romaine Francois' old R Graphics Gallery, sadly now defunct.

Many of the data sets used in the book are from the UC Irvine Machine Learning Repository, `http://archive.ics.uci.edu/ml/`. Thanks to UCI for making available this very valuable resource.

The book contains a number of references for further reading. Since the audience includes a number of students at my institution, the University of California, Davis, I often refer to work by current or former UCD faculty, so that students can see what their professors do in research.

Chapter 1

Time Waste Versus Empowerment

I took a course in speed reading, and read War and Peace in 20 minutes. It's about Russia—comedian Woody Allen

I learned very early the difference between knowing the name of something and knowing something—Richard Feynman, Nobel laureate in physics

The main goal [of this course] is self-actualization through the empowerment of claiming your education—UCSC (and former UCD) professor Marc Mangel, in the syllabus for his calculus course

What does this really mean? Hmm, I've never thought about that—UCD PhD student in statistics, in answer to a student who asked the actual meaning of a very basic concept

You have a PhD in engineering. You may have forgotten technical details like $\frac{d}{dt}sin(t) = cos(t)$*, but you should at least understand the concepts of rates of change*—the author, gently chiding a friend who was having trouble following a simple quantitative discussion of trends in California's educational system

Give me six hours to chop down a tree and I will spend the first four sharpening the axe—Abraham Lincoln

The field of probability and statistics (which, for convenience, I will refer to simply as "statistics" below) impacts many aspects of our daily lives—business, medicine, the law, government and so on. Consider just a few examples:

- The statistical models used on Wall Street made the "quants" (quantitative analysts) rich—but also contributed to the worldwide financial crash of 2008.
- In a court trial, large sums of money or the freedom of an accused may hinge on whether the

judge and jury understand some statistical evidence presented by one side or the other.

- Wittingly or unconsciously, you are using probability every time you gamble in a casino—and every time you buy insurance.
- Statistics is used to determine whether a new medical treatment is safe/effective for you.
- Statistics is used to flag possible terrorists—but sometimes unfairly singling out innocent people while other times missing ones who really are dangerous.

Clearly, statistics *matters*. But it only has value when one really *understands* what it means and what it does. Indeed, blindly plugging into statistical formulas can be not only valueless but in fact highly dangerous, say if a bad drug goes onto the market.

Yet most people view statistics as exactly that—mindless plugging into boring formulas. If even the statistics graduate student quoted above thinks this, how can the students taking the course be blamed for taking that atititude?

I once had a student who had an unusually good understanding of probability. It turned out that this was due to his being highly successful at playing online poker, winning lots of cash. No blind formula-plugging for him! He really had to *understand* how probability works.

Statistics is *not* just a bunch of formulas. On the contrary, it can be mathematically deep, for those who like that kind of thing. (Much of statistics can be viewed as the Pythagorean Theorem in n-dimensional or even infinite-dimensional space.) But the key point is that *anyone* who has taken a calculus course can develop true understanding of statistics, of real practical value. As Professor Mangel says, that's empowering.

Statistics is based on probabilistic models. So, in order to become effective at data analysis, one must really master the principles of probability; this is where Lincoln's comment about "sharpening the axe" truly applies.

So as you make your way through this book, always stop to think, "What does this equation really mean? What is its goal? Why are its ingredients defined in the way they are? Might there be a better way? How does this relate to our daily lives?" Now THAT is empowering.

Chapter 2

Basic Probability Models

This chapter will introduce the general notions of probability. Most of it will seem intuitive to you, but pay careful attention to the general principles which are developed; in more complex settings intuition may not be enough, and the tools discussed here will be very useful.

2.1 ALOHA Network Example

Throughout this book, we will be discussing both "classical" probability examples involving coins, cards and dice, and also examples involving applications to computer science. The latter will involve diverse fields such as data mining, machine learning, computer networks, software engineering and bioinformatics.

In this section, an example from computer networks is presented which will be used at a number of points in this chapter. Probability analysis is used extensively in the development of new, faster types of networks.

Today's Ethernet evolved from an experimental network developed at the University of Hawaii, called ALOHA. A number of network nodes would occasionally try to use the same radio channel to communicate with a central computer. The nodes couldn't hear each other, due to the obstruction of mountains between them. If only one of them made an attempt to send, it would be successful, and it would receive an acknowledgement message in response from the central computer. But if more than one node were to transmit, a **collision** would occur, garbling all the messages. The sending nodes would timeout after waiting for an acknowledgement which never came, and try sending again later. To avoid having too many collisions, nodes would engage in random **backoff**, meaning that they would refrain from sending for a while even though they had something to send.

One variation is **slotted** ALOHA, which divides time into intervals which I will call "epochs." Each

epoch will have duration 1.0, so epoch 1 extends from time 0.0 to 1.0, epoch 2 extends from 1.0 to 2.0 and so on. In the version we will consider here, in each epoch, if a node is **active**, i.e. has a message to send, it will either send or refrain from sending, with probability p and 1-p. The value of p is set by the designer of the network. (Real Ethernet hardware does something like this, using a random number generator inside the chip.)

The other parameter q in our model is the probability that a node which had been inactive generates a message during an epoch, i.e. the probability that the user hits a key, and thus becomes "active." Think of what happens when you are at a computer. You are not typing constantly, and when you are not typing, the time until you hit a key again will be random. Our parameter q models that randomness.

Let n be the number of nodes, which we'll assume for simplicity is two. Assume also for simplicity that the timing is as follows. Arrival of a new message happens in the middle of an epoch, and the decision as to whether to send versus back off is made near the end of an epoch, say 90% into the epoch.

For example, say that at the beginning of the epoch which extends from time 15.0 to 16.0, node A has something to send but node B does not. At time 15.5, node B will either generate a message to send or not, with probability q and 1-q, respectively. Suppose B does generate a new message. At time 15.9, node A will either try to send or refrain, with probability p and 1-p, and node B will do the same. Suppose A refrains but B sends. Then B's transmission will be successful, and at the start of epoch 16 B will be inactive, while node A will still be active. On the other hand, suppose both A and B try to send at time 15.9; both will fail, and thus both will be active at time 16.0, and so on.

Be sure to keep in mind that in our simple model here, during the time a node is active, it won't generate any additional new messages.

(Note: The definition of this ALOHA model is summarized concisely on page 10.)

Let's observe the network for two epochs, epoch 1 and epoch 2. Assume that the network consists of just two nodes, called node 1 and node 2, both of which start out active. Let X_1 and X_2 denote the numbers of active nodes at the *very end* of epochs 1 and 2, *after possible transmissions*. We'll take p to be 0.4 and q to be 0.8 in this example.

Let's find $P(X_1 = 2)$, the probability that $X_1 = 2$, and then get to the main point, which is to ask what we really mean by this probability.

How could $X_1 = 2$ occur? There are two possibilities:

- both nodes try to send; this has probability p^2
- neither node tries to send; this has probability $(1 - p)^2$

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

Table 2.1: Sample Space for the Dice Example

Thus

$$P(X_1 = 2) = p^2 + (1-p)^2 = 0.52 \tag{2.1}$$

2.2 The Crucial Notion of a Repeatable Experiment

It's crucial to understand what that 0.52 figure really means in a practical sense. To this end, let's put the ALOHA example aside for a moment, and consider the "experiment" consisting of rolling two dice, say a blue one and a yellow one. Let X and Y denote the number of dots we get on the blue and yellow dice, respectively, and consider the meaning of $P(X + Y = 6) = \frac{5}{36}$.

In the mathematical theory of probability, we talk of a **sample space**, which (in simple cases) consists of the possible outcomes (X, Y), seen in Table 2.1. In a theoretical treatment, we place weights of 1/36 on each of the points in the space, reflecting the fact that each of the 36 points is equally likely, and then say, "What we mean by $P(X + Y = 6) = \frac{5}{36}$ is that the outcomes (1,5), (2,4), (3,3), (4,2), (5,1) have total weight 5/36."

Unfortunately, the notion of sample space becomes mathematically tricky when developed for more complex probability models. Indeed, it requires graduate-level math, called **measure theory**.

And much worse, under the sample space approach, one loses all the intuition. In particular, **there is no good way using set theory to convey the intuition underlying conditional probability** (to be introduced in Section 2.3). The same is true for expected value, a central topic to be introduced in Section 3.5.

In any case, most probability computations do not rely on explicitly writing down a sample space. In this particular example it is useful for us as a vehicle for explaining the concepts, but we will NOT use it much. Those who wish to get a more theoretical grounding can get a start in Section

notebook line	outcome	blue+yellow = 6?
1	blue 2, yellow 6	No
2	blue 3, yellow 1	No
3	blue 1, yellow 1	No
4	blue 4, yellow 2	Yes
5	blue 1, yellow 1	No
6	blue 3, yellow 4	No
7	blue 5, yellow 1	Yes
8	blue 3, yellow 6	No
9	blue 2, yellow 5	No

Table 2.2: Notebook for the Dice Problem

5.4.

But the intuitive notion—which is FAR more important—of what $P(X + Y = 6) = \frac{5}{36}$ means is the following. Imagine doing the experiment many, many times, recording the results in a large notebook:

- Roll the dice the first time, and write the outcome on the first line of the notebook.
- Roll the dice the second time, and write the outcome on the second line of the notebook.
- Roll the dice the third time, and write the outcome on the third line of the notebook.
- Roll the dice the fourth time, and write the outcome on the fourth line of the notebook.
- Imagine you keep doing this, thousands of times, filling thousands of lines in the notebook.

The first 9 lines of the notebook might look like Table 2.2. Here 2/9 of these lines say Yes. But after many, many repetitions, approximately 5/36 of the lines will say Yes. For example, after doing the experiment 720 times, approximately $\frac{5}{36} \times 720 = 100$ lines will say Yes.

This is what probability really is: In what fraction of the lines does the event of interest happen? **It sounds simple, but if you always think about this "lines in the notebook" idea, probability problems are a lot easier to solve.** And it is the fundamental basis of computer simulation.

2.3 Our Definitions

These definitions are intuitive, rather than rigorous math, but intuition is what we need. Keep in mind that we are making definitions below, not listing properties.

- We assume an "experiment" which is (at least in concept) repeatable. The experiment of rolling two dice is repeatable, and even the ALOHA experiment is so. (We simply watch the network for a long time, collecting data on pairs of consecutive epochs in which there are two active stations at the beginning.) On the other hand, the econometricians, in forecasting 2009, cannot "repeat" 2008. Yet all of the econometricians' tools assume that events in 2008 were affected by various sorts of randomness, and we think of repeating the experiment in a conceptual sense.

- We imagine performing the experiment a large number of times, recording the result of each repetition on a separate line in a notebook.

- We say A is an **event** for this experiment if it is a possible boolean (i.e. yes-or-no) outcome of the experiment. In the above example, here are some events:

  * X+Y = 6
  * X = 1
  * Y = 3
  * X-Y = 4

- A **random variable** is a numerical outcome of the experiment, such as X and Y here, as well as X+Y, 2XY and even sin(XY).

- For any event of interest A, imagine a column on A in the notebook. The k^{th} line in the notebook, k = 1,2,3,..., will say Yes or No, depending on whether A occurred or not during the k^{th} repetition of the experiment. For instance, we have such a column in our table above, for the event {A = blue+yellow = 6}.

- For any event of interest A, we define P(A) to be the long-run fraction of lines with Yes entries.

- For any events A, B, imagine a new column in our notebook, labeled "A and B." In each line, this column will say Yes if and only if there are Yes entries for both A and B. P(A and B) is then the long-run fraction of lines with Yes entries in the new column labeled "A and B."[1]

[1] In most textbooks, what we call "A and B" here is written A∩B, indicating the intersection of two sets in the sample space. But again, we do not take a sample space point of view here.

- For any events A, B, imagine a new column in our notebook, labeled "A or B." In each line, this column will say Yes if and only if at least one of the entries for A and B says Yes.[2]
- For any events A, B, imagine a new column in our notebook, labeled "A | B" and pronounced "A given B." In each line:
  * This new column will say "NA" ("not applicable") if the B entry is No.
  * If it is a line in which the B column says Yes, then this new column will say Yes or No, depending on whether the A column says Yes or No.

Think of probabilities in this "notebook" context:

- P(A) means the long-run fraction of lines in the notebook in which the A column says Yes.
- P(A or B) means the long-run fraction of lines in the notebook in which the A-or-B column says Yes.
- P(A and B) means the long-run fraction of lines in the notebook in which the A-and-B column says Yes.
- P(A | B) means the long-run fraction of lines in the notebook in which the A | B column says Yes—**among the lines which do NOT say NA.**

A hugely common mistake is to confuse P(A and B) and P(A | B). This is where the notebook view becomes so important. Compare the quantities $P(X = 1 \text{ and } S = 6) = \frac{1}{36}$ and $P(X = 1|S = 6) = \frac{1}{5}$, where S = X+Y:[3]

- After a large number of repetitions of the experiment, approximately 1/36 of the lines of the notebook will have the property that both X = 1 and S = 6 (since X = 1 and S = 6 is equivalent to X = 1 and Y = 5).
- After a large number of repetitions of the experiment, if **we look only at the lines in which S = 6**, then **among those lines**, approximately 1/5 of **those lines** will show X = 1.

The quantity P(A|B) is called the **conditional probability of A, given B**.

Note that *and* has higher logical precedence than *or*. For example, P(A and B or C) means P[(A and B) or C]. Also, *not* has higher precedence than *and*.

Here are some more very important definitions and properties:

[2] In the sample space approach, this is written A ∪ B.

[3] Think of adding an S column to the notebook too

- **Definition 1** *Suppose A and B are events such that it is impossible for them to occur in the same line of the notebook. They are said to be* **disjoint** *events.*

- If A and B are disjoint events, then

$$P(A \text{ or } B) = P(A) + P(B) \tag{2.2}$$

  Again, this terminology *disjoint* stems from the set-theoretic sample space approach, where it means that A $\cap$ B = ϕ. That mathematical terminology works fine for our dice example, but in my experience people have major difficulty applying it correctly in more complicated problems. This is another illustration of why I put so much emphasis on the "notebook" framework.

  By writing

$$\{A \text{ or } B \text{ or } C\} = \{A \text{ or } [B \text{ or } C]\} = \tag{2.3}$$

  (2.2) can be iterated, e.g.

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) \tag{2.4}$$

- If A and B are not disjoint, then

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{2.5}$$

  In the disjoint case, that subtracted term is 0, so (2.5) reduces to (2.2).

- **Definition 2** *Events A and B are said to be* **stochastically independent**, *usually just stated as* **independent**,[4] *if*

$$P(A \text{ and } B) = P(A) \cdot P(B) \tag{2.6}$$

- In calculating an "and" probability, how does one know whether the events are independent? The answer is that this will typically be clear from the problem. If we toss the blue and yellow dice, for instance, it is clear that one die has no impact on the other, so events involving the blue die are independent of events involving the yellow die. On the other hand, in the ALOHA example, it's clear that events involving X_1 are NOT independent of those involving X_2.

[4]The term *stochastic* is just a fancy synonym for *random*.

- If A and B are not independent, the equation (2.6) generalizes to

$$P(A \text{ and } B) = P(A)P(B|A) \tag{2.7}$$

This should make sense to you. Suppose 30% of all UC Davis students are in engineering, and 20% of all engineering majors are female. That would imply that 0.30 x 0.20 = 0.06, i.e. 6% of all UCD students are female engineers.

Note that if A and B actually are independent, then $P(B|A) = P(B)$, and (2.7) reduces to (2.6).

Note too that (2.7) implies

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \tag{2.8}$$

2.4 "Mailing Tubes"

If I ever need to buy some mailing tubes, I can come here—friend of the author's, while browsing through an office supplies store

Examples of the above properties, e.g. (2.6) and (2.7), will be given starting in Section 2.5. But first, a crucial strategic point in learning probability must be addressed.

Some years ago, a friend of mine was in an office supplies store, and he noticed a rack of mailing tubes. My friend made the remark shown above. Well, (2.6) and 2.7 are "mailing tubes"–make a mental note to yourself saying, "If I ever need to find a probability involving *and*, one thing I can try is (2.6) and (2.7)." **Be ready for this!**

This mailing tube metaphor will be mentioned often, such as in Section 3.5.5 .

2.5 Example: ALOHA Network

Please keep in mind that the notebook idea is simply a vehicle to help you understand what the concepts really mean. This is crucial for your intuition and your ability to apply this material in the real world. But the notebook idea is NOT for the purpose of calculating probabilities. Instead, we use the properties of probability, as seen in the following.

Let's look at all of this in the ALOHA context. Here's a summary:

- We have n network nodes, sharing a common communications channel.
- Time is divided in epochs. X_k denotes the number of active nodes at the end of epoch k, which we will sometimes refer to as the **state** of the system in epoch k.
- If two or more nodes try to send in an epoch, they collide, and the message doesn't get through.
- We say a node is active if it has a message to send.
- If a node is active node near the end of an epoch, it tries to send with probability p.
- If a node is inactive at the beginning of an epoch, then at the middle of the epoch it will generate a message to send with probability q.
- In our examples here, we have n = 2 and $X_0 = 2$, i.e. both nodes start out active.

Now, in Equation (2.1) we found that

$$P(X_1 = 2) = p^2 + (1-p)^2 = 0.52 \tag{2.9}$$

How did we get this? Let C_i denote the event that node i tries to send, i = 1,2. Then using the definitions above, our steps would be

$$\begin{aligned} P(X_1 = 2) &= P(\underbrace{C_1 \text{ and } C_2}\ \text{ or }\ \underbrace{\text{not } C_1 \text{ and not } C_2}) && (2.10)\\ &= P(C_1 \text{ and } C_2) + P(\text{ not } C_1 \text{ and not } C_2) \text{ (from (2.2))} && (2.11)\\ &= P(C_1)P(C_2) + P(\text{ not } C_1)P(\text{ not } C_2) \text{ (from (2.6))} && (2.12)\\ &= p^2 + (1-p)^2 && (2.13) \end{aligned}$$

(The underbraces in (2.10) do not represent some esoteric mathematical operation. There are there simply to make the grouping clearer, corresponding to events G and H defined below.)

Here are the reasons for these steps:

(2.10): We listed the ways in which the event $\{X_1 = 2\}$ could occur.

(2.11): Write $G = C_1$ and C_2, $H = D_1$ and D_2, where $D_i =$ not C_i, i = 1,2. Then the events G and H are clearly disjoint; if in a given line of our notebook there is a Yes for G, then definitely there will be a No for H, and vice versa.

(2.12): The two nodes act physically independently of each other. Thus the events C_1 and C_2 are stochastically independent, so we applied (2.6). Then we did the same for D_1 and D_2.

Now, what about $P(X_2 = 2)$? Again, we break big events down into small events, in this case according to the value of X_1:

$$\begin{aligned} P(X_2 = 2) &= P(X_1 = 0 \text{ and } X_2 = 2 \text{ or } X_1 = 1 \text{ and } X_2 = 2 \text{ or } X_1 = 2 \text{ and } X_2 = 2) \\ &= P(X_1 = 0 \text{ and } X_2 = 2) \\ &+ P(X_1 = 1 \text{ and } X_2 = 2) \\ &+ P(X_1 = 2 \text{ and } X_2 = 2) \end{aligned} \tag{2.14}$$

Since X_1 cannot be 0, that first term, $P(X_1 = 0 \text{ and } X_2 = 2)$ is 0. To deal with the second term, $P(X_1 = 1 \text{ and } X_2 = 2)$, we'll use (2.7). Due to the time-sequential nature of our experiment here, it is natural (but certainly not "mandated," as we'll often see situations to the contrary) to take A and B to be $\{X_1 = 1\}$ and $\{X_2 = 2\}$, respectively. So, we write

$$P(X_1 = 1 \text{ and } X_2 = 2) = P(X_1 = 1)P(X_2 = 2|X_1 = 1) \tag{2.15}$$

To calculate $P(X_1 = 1)$, we use the same kind of reasoning as in Equation (2.1). For the event in question to occur, either node A would send and B wouldn't, or A would refrain from sending and B would send. Thus

$$P(X_1 = 1) = 2p(1 - p) = 0.48 \tag{2.16}$$

Now we need to find $P(X_2 = 2|X_1 = 1)$. This again involves breaking big events down into small ones. If $X_1 = 1$, then $X_2 = 2$ can occur only if *both* of the following occur:

- Event A: Whichever node was the one to successfully transmit during epoch 1—and we are given that there indeed was one, since $X_1 = 1$—now generates a new message.
- Event B: During epoch 2, no successful transmission occurs, i.e. either they both try to send or neither tries to send.

Recalling the definitions of p and q in Section 2.1, we have that

$$P(X_2 = 2|X_1 = 1) = q[p^2 + (1 - p)^2] = 0.41 \tag{2.17}$$

Thus $P(X_1 = 1 \text{ and } X_2 = 2) = 0.48 \times 0.41 = 0.20$.

We go through a similar analysis for $P(X_1 = 2 \text{ and } X_2 = 2)$: We recall that $P(X_1 = 2) = 0.52$ from before, and find that $P(X_2 = 2|X_1 = 2) = 0.52$ as well. So we find $P(X_1 = 2 \text{ and } X_2 = 2)$ to be $0.52^2 = 0.27$. Putting all this together, we find that $P(X_2 = 2) = 0.47$.

Let's do another; let's find $P(X_1 = 1|X_2 = 2)$. [Pause a minute here to make sure you understand that this is quite different from $P(X_2 = 2|X_1 = 1)$.] From (2.8), we know that

$$P(X_1 = 1|X_2 = 2) = \frac{P(X_1 = 1 \text{ and } X_2 = 2)}{P(X_2 = 2)} \tag{2.18}$$

We computed both numerator and denominator here before, in Equations (2.15) and (2.14), so we see that $P(X_1 = 1|X_2 = 2) = 0.20/0.47 = 0.43$.

So, in our notebook view, if we were to look only at lines in the notebook for which $X_2 = 2$, a fraction 0.43 of *those lines* would have $X_1 = 1$.

You might be bothered that we are looking "backwards in time" in (2.18), kind of guessing the past from the present. There is nothing wrong or unnatural about that. Jurors in court trials do it all the time, though presumably not with formal probability calculation. And evolutionary biologists do use formal probability models to guess the past.

And one more calculation: $P(X_1 = 2 \text{ or } X_2 = 2)$. From (2.5),

$$P(X_1 = 2 \text{ or } X_2 = 2) = P(X_1 = 2) + P(X_2 = 2) - P(X_1 = 2 \text{ and } X_2 = 2) \tag{2.19}$$

Luckily, we've already calculated all three probabilities on the right-hand side to be 0.52, 0.47 and 0.27, respectively. Thus $P(X_1 = 2 \text{ or } X_2 = 2) = 0.72$.

Note by the way that events involving X_2 are NOT independent of those involving X_1. For instance, we found in (2.18) that

$$P(X_1 = 1|X_2 = 2) = 0.43 \tag{2.20}$$

yet from (2.16) we have

$$P(X_1 = 1) = 0.48. \tag{2.21}$$

2.6 Bayes' Rule

(This section should not be confused with Section 22.4. The latter is highly controversial, while the material in this section is not controversial at all.)

Following (2.18) above, we noted that the ingredients had already been computed, in (2.15) and (2.14). If we go back to the derivations in those two equations and substitute in (2.18), we have

$$P(X_1 = 1 | X_2 = 2) = \frac{P(X_1 = 1 \text{ and } X_2 = 2)}{P(X_2 = 2)} \tag{2.22}$$

$$= \frac{P(X_1 = 1 \text{ and } X_2 = 2)}{P(X_1 = 1 \text{ and } X_2 = 2) + P(X_1 = 2 \text{ and } X_2 = 2)} \tag{2.23}$$

$$= \frac{P(X_1 = 1) P(X_2 = 2 | X_1 = 1)}{P(X_1 = 1) P(X_2 = 2 | X_1 = 1) + P(X_1 = 2) P(X_2 = 2 | X_1 = 2)} \tag{2.24}$$

Looking at this in more generality, for events A and B we would find that

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)} \tag{2.25}$$

This is known as **Bayes' Theorem** or **Bayes' Rule**. It can be extended easily to cases with several terms in the denominator, arising from situations that need to be broken down into several subevents rather than just A and not-A.

2.7 ALOHA in the Notebook Context

Think of doing the ALOHA "experiment" many, many times.

- Run the network for two epochs, starting with both nodes active, the first time, and write the outcome on the first line of the notebook.
- Run the network for two epochs, starting with both nodes active, the second time, and write the outcome on the second line of the notebook.
- Run the network for two epochs, starting with both nodes active, the third time, and write the outcome on the third line of the notebook.
- Run the network for two epochs, starting with both nodes active, the fourth time, and write the outcome on the fourth line of the notebook.

notebook line	$X_1 = 2$	$X_2 = 2$	$X_1 = 2$ and $X_2 = 2$	$X_2 = 2 \mid X_1 = 2$
1	Yes	No	No	No
2	No	No	No	NA
3	Yes	Yes	Yes	Yes
4	Yes	No	No	No
5	Yes	Yes	Yes	Yes
6	No	No	No	NA
7	No	Yes	No	NA

Table 2.3: Top of Notebook for Two-Epoch ALOHA Experiment

- Imagine you keep doing this, thousands of times, filling thousands of lines in the notebook.

The first seven lines of the notebook might look like Table 2.3. We see that:

- Among those first seven lines in the notebook, 4/7 of them have $X_1 = 2$. After many, many lines, this fraction will be approximately 0.52.
- Among those first seven lines in the notebook, 3/7 of them have $X_2 = 2$. After many, many lines, this fraction will be approximately 0.47.[5]
- Among those first seven lines in the notebook, 3/7 of them have $X_1 = 2$ and $X_2 = 2$. After many, many lines, this fraction will be approximately 0.27.
- Among the first seven lines in the notebook, four of them do not say NA in the $X_2 = 2 \mid X_1 = 2$ column. **Among these four lines**, two say Yes, a fraction of 2/4. After many, many lines, this fraction will be approximately 0.52.

2.8 A Note on Modeling

Here is a question to test your understanding of the ALOHA model—not the calculation of probabilities, but rather the meaning of the model itself. What kinds of properties are we trying to capture in the model?

So, consider this question:

[5] Don't make anything of the fact that these probabilities nearly add up to 1.

Consider the ALOHA network model. Say we have two such networks, A and B. In A, the network typically is used for keyboard input, such as a user typing e-mail or editing a file. But in B, users tend to do a lot of file uploading, not just typing. Fill in the blanks: In B, the model parameter _____ is _____ than in A, and in order to accommodate this, we should set the parameter _____ to be relatively _____ in B.

In network B we have heavy traffic. Thus, when a node becomes idle, it is quite likely to have a new message to send right away.[6] Thus q is large.

That means we need to be especially worried about collisions, so we probably should set p to a low value.

2.9 Solution Strategies

The example in Section 2.5 shows typical stategies in exploring solutions to probability problems, such as:

- Name what seem to be the important variables and events, in this case X_1, X_2, C_1, C_2 and so on.
- Write the given probability in terms of those named variables, e.g.

$$P(X_1 = 2) = P(\underbrace{C_1 \text{ and } C_2}\ \text{ or }\ \underbrace{\text{not } C_1 \text{ and not } C_2}) \tag{2.26}$$

  above.
- Ask the famous question, "How can it happen?" Break big events down into small events; in the above case the event $X_1 = 2$ can happen if C_1 and C_2 or not C_1 and not C_2.
- But when you do break things down like this, make sure to neither expand or contract the scope of the probability. Say you write something like

$$P(A) = P(B) \tag{2.27}$$

  where B might be some complicated event expression such as in the right-hand side of (2.10). Make SURE that A and B are equivalent—meaning that in every notebook line in which A occurs, then B also occurs, and *vice versa*.

[6] The file is likely read in chunks called disk *sectors*, so there may be a slight pause between the uploading of chunks. Our model here is too coarse to reflect such things.

- Do not write/think nonsense. For example: the expression "P(A) or P(B)" is nonsense—do you see why? Probabilities are numbers, not boolean expressions, so "P(A) or P(B)" is like saying, "0.2 or 0.5"—meaningless!

  Similarly, say we have a random variable X. The "probability" P(X) is invalid. Say X is the number of dots we get when we roll a single die. Then P(X) would mean "the probability that the number of dots," which is nonsense English! P(X = 3) is valid, but P(X) is meaningless.

  Please note that = is not like a comma, or equivalent to the English word *therefore*. It needs a left side and a right side; "a = b" makes sense, but "= b" doesn't.

- Similarly, don't use "formulas" that you didn't learn and that are in fact false. For example, in an expression involving a random variable X, one can NOT replace X by its mean. (How would you like it if your professor were to lose your exam, and then tell you, "Well, I'll just assign you a score that is equal to the class mean"?)

- Adhere to convention! Use capital letters for random variables and names of events. Use P() notation, not p() (which will mean something else in this book).

- In the beginning of your learning probability methods, meticulously write down all your steps, with reasons, as in the computation of $P(X_1 = 2)$ in Equations (2.10)ff. After you gain more experience, you can start skipping steps, but not in the initial learning period.

- Solving probability problems—and even more so, building useful probability models—is like computer programming: It's a creative process.

  One can NOT—repeat, NOT—teach someone how to write programs. All one can do is show the person how the basic building blocks work, such as loops, if-else and arrays, then show a number of examples. But the actual writing of a program is a creative act, not formula-based. The programmer must creatively combined the various building blocks to produce the desired result. The teacher cannot teach the student how to do this.

  The same is true for solving probability problems. The basic building blocks were presented above in Section 2.5, and many more "mailing tubes" will be presented in the rest of this book. But it is up to the student to try using the various building blocks in a way that solves the problem. Sometimes use of one block may prove to be unfruitful, in which case one must try other blocks.

  For instance, in using probability formulas like P(A and B) = P(A) P(B|A), there is no magic rule as to how to choose A and B.

  Moreover, if you need P(B|A), there is no magic rule on how to find it. On the one hand, you might calculate it from (2.8), as we did in (2.18), but on the other hand you may be able to reason out the value of P(B|A), as we did following (2.16). Just try some cases until you find one that works, in the sense that you can evaluate both factors. It's the same as trying various programming ideas until you find one that works.

2.10 Example: Divisibility of Random Integers

Suppose at step i we generate a random integer between 1 and 1000, and check whether it's evenly divisible by i, i = 5,4,3,2,1. Let N denote the number of steps needed to reach an evenly divisible number.

Let's find P(N = 2). Let q(i) denote the fraction of numbers in 1...,1000 that are evenly divisible by i, so that for instance q(5) = 200/1000 = 1/5 while q(3) = 333/1000. Let's label the steps 5,4,..., so that the first step is number 5. Then since the random numbers are independent from step to step, we have

$$
\begin{aligned}
P(N=2) &= P(\text{fail in step 5 and succeed in step 4}) \quad (\text{“How can it happen?”}) && (2.28)\\
&= P(\text{fail in step 5}) \quad P(\text{succeed in step 4} \mid \text{fail in step 5}) \quad ((2.7)) && (2.29)\\
&= [1-q(5)]q(4) && (2.30)\\
&= \frac{4}{5} \cdot \frac{1}{4} && (2.31)\\
&= \frac{1}{5} && (2.32)
\end{aligned}
$$

But there's more.

First, note that q(i) is either equal or approximately equal to 1/i. Then following the derivation in (2.28), you'll find that

$$P(N=j) \approx \frac{1}{5} \tag{2.33}$$

for ALL j in 1,...,5.

That may seem counterintuitive. Yet the example here is in essence the same as one found as an exercise in many textbooks on probability:

> A man has five keys. He knows one of them opens a given lock, but he doesn't know which. So he tries the keys one at a time until he finds the right one. Find P(N = j), j = 1,...,5, where N is the number of keys he tries until he succeeds.

Here too the answer is 1/5 for all j. But this one makes intuitive sense: Each of the keys has chance 1/5 of being the right key, so each of the values 1,...,5 is equally likely for N.

This is then an example of the fact that sometimes we can gain insight into one problem by considering a mathematically equivalent problem in a quite different setting.

2.11 Example: A Simple Board Game

Consider a board game, which for simplicity we'll assume consists of two squares per side, on four sides. A player's token advances around the board. The squares are numbered 0-7, and play begins at square 0.

A token advances according to the roll of a single die. If a player lands on square 3, he/she gets a bonus turn. Let's find the probability that a player has yet to make a complete circuit of the board—i.e. has not yet reached or passed 0—after the first turn (including the bonus, if any). Let R denote his first roll, and let B be his bonus if there is one, with B being set to 0 if there is no bonus. Then (using commas as a shorthand notation for *and*)

$$
\begin{aligned}
P(\text{doesn't reach or pass 0}) &= P(R+B \leq 7) && (2.34)\\
&= P(R \leq 6, R \neq 3 \text{ or } R = 3, B \leq 4) && (2.35)\\
&= P(R \leq 6, R \neq 3) + P(R = 3, B \leq 4) && (2.36)\\
&= P(R \leq 6, R \neq 3) + P(R = 3)\ P(B \leq 4) && (2.37)\\
&= \frac{5}{6} + \frac{1}{6} \cdot \frac{4}{6} && (2.38)\\
&= \frac{17}{18} && (2.39)
\end{aligned}
$$

Now, here's a shorter way (there are always multiple ways to do a problem):

$$
\begin{aligned}
P(\text{don't reach or pass 0}) &= 1 - P(\text{do reach or pass 0}) && (2.40)\\
&= 1 - P(R+B > 7) && (2.41)\\
&= 1 - P(R = 3, B > 4) && (2.42)\\
&= 1 - \frac{1}{6} \cdot \frac{2}{6} && (2.43)\\
&= \frac{17}{18} && (2.44)
\end{aligned}
$$

Now suppose that, according to a telephone report of the game, you hear that on the player's first turn, his token ended up at square 4. Let's find the probability that he got there with the aid of a bonus roll.

Note that this a conditional probability—we're finding the probability that the player got a bonus roll, given that we know he ended up at square 4. The word *given* wasn't there in the statement of the problem, but it was implied.

A little thought reveals that we cannot end up at square 4 after making a complete circuit of the board, which simplifies the situation quite a bit. So, write

$$
\begin{aligned}
P(B > 0 \mid R + B = 4) &= \frac{P(R + B = 4, B > 0)}{P(R + B = 4)} && (2.45)\\
&= \frac{P(R + B = 4, B > 0)}{P(R + B = 4, B > 0 \text{ or } R + B = 4, B = 0)} && (2.46)\\
&= \frac{P(R + B = 4, B > 0)}{P(R + B = 4, B > 0) + P(R + B = 4, B = 0)} && (2.47)\\
&= \frac{P(R = 3, B = 1)}{P(R = 3, B = 1) + P(R = 4)} && (2.48)\\
&= \frac{\frac{1}{6} \cdot \frac{1}{6}}{\frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6}} && (2.49)\\
&= \frac{1}{7} && (2.50)
\end{aligned}
$$

We could have used Bayes' Rule to shorten the derivation a little here, but will prefer to derive everything, at least in this introductory chapter.

Pay special attention to that third equality above, as it is a frequent mode of attack in probability problems. In considering the probability P(R+B = 4, B > 0), we ask, what is a simpler—but still equivalent!—description of this event? Well, we see that R+B = 4, B > 0 boils down to R = 3, B = 1, so we replace the above probability with P(R = 3, B = 1).

Again, this is a very common approach. But be sure to take care that we are in an "if and only if" situation. Yes, R+B = 4, B > 0 implies R = 3, B = 1, but we must make sure that the converse is true as well. In other words, we must also confirm that R = 3, B = 1 implies R+B = 4, B > 0. That's trivial in this case, but one can make a subtle error in some problems if one is not careful; otherwise we will have replaced a higher-probability event by a lower-probability one.

2.12 Example: Bus Ridership

Consider the following analysis of bus ridership. (In order to keep things easy, it will be quite oversimplified, but the principles will be clear.) Here is the model:

- At each stop, each passsenger alights from the bus, independently, with probability 0.2 each.
- Either 0, 1 or 2 new passengers get on the bus, with probabilities 0.5, 0.4 and 0.1, respectively. Passengers at successive stops are independent.

- Assume the bus is so large that it never becomes full, so the new passengers can always get on.
- Suppose the bus is empty when it arrives at its first stop.

Let L_i denote the number of passengers on the bus as it *leaves* its ith stop, i = 1,2,3,... Let B_i denote the number of new passengers who board the bus at the ith stop.

First, let's find the probability that no passengers board the bus at the first three stops. That's easy:

$$P(B_1 = 0 \text{ and } B_2 = 0 \text{ and } B_3 = 0) = 0.5^3 \tag{2.51}$$

Now let's compute $P(L_2 = 0)$.

$$\begin{aligned} P(L_2 = 0) &= P(B_1 = 0 \text{ and } L_2 = 0 \text{ or } B_1 = 1 \text{ and } L_2 = 0 \text{ or } B_1 = 2 \text{ and } L_2 = 0) && (2.52)\\ &= \sum_{i=0}^{2} P(B_1 = i \text{ and } L_2 = 0) && (2.53)\\ &= \sum_{i=0}^{2} P(B_1 = i) P(L_2 = 0 | B_1 = i) && (2.54)\\ &= 0.5^2 + (0.4)(0.2)(0.5) + (0.1)(0.2^2)(0.5) && (2.55)\\ &= 0.292 && (2.56) \end{aligned}$$

For instance, where did that first term, 0.5^2, come from? Well, $P(B_1 = 0) = 0.5$, and what about $P(L_2 = 0|B_1 = 0)$? If $B_1 = 0$, then the bus approaches the second stop empty. For it to then *leave* that second stop empty, it must be the case that $B_2 = 0$, which has probability 0.5. In other words, $P(L_2 = 0|B_1 = 0) = 0.5$.

As another example, suppose we are told that the bus arrives empty at the third stop. What is the probability that exactly two people boarded the bus at the first stop?

Note first what is being asked for here: $P(B_1 = 2|L_2 = 0)$. Then we have

$$\begin{aligned} P(B_1 = 2 \mid L_2 = 0) &= \frac{P(B_1 = 2 \text{ and } L_2 = 0)}{P(L_2 = 0)} && (2.57)\\ &= P(B_1 = 2)\ P(L_2 = 0 \mid B_1 = 2)\ /\ 0.292 && (2.58)\\ &= 0.1 \cdot 0.2^2 \cdot 0.5/0.292 && (2.59) \end{aligned}$$

(the 0.292 had been previously calculated in (2.56)).

Now let's find the probability that fewer people board at the second stop than at the first:

$$\begin{aligned} P(B_2 < B_1) &= P(B_1 = 1 \text{ and } B_2 < B_1 \text{ or } B_1 = 2 \text{ and } B_2 < B_1) && (2.60) \\ &= 0.4 \cdot 0.5 + 0.1 \cdot (0.5 + 0.4) && (2.61) \end{aligned}$$

Also: Someone tells you that as she got off the bus at the second stop, she saw that the bus then left that stop empty. Let's find the probability that she was the only passenger when the bus left the first stop:

We are given that $L_2 = 0$. But we are *also* given that $L_1 > 0$. Then

$$\begin{aligned} P(L_1 = 1 | L_2 = 0 \text{ and } L_1 > 0) &= \frac{P(L_1 = 1 \text{ and } L_2 = 0)}{P(L_2 = 0 \text{ and } L_1 > 0)} && (2.62) \\ &= \frac{P(B_1 = 1 \text{ and } L_2 = 0)}{P(B_1 = 1 \text{ and } L_2 = 0 \text{ or } B_1 = 2 \text{ and } L_2 = 0)} && (2.63) \\ &= \frac{(0.4)(0.2)(0.5)}{(0.4)(0.2)(0.5) + (0.1)(0.2)^2(0.5)} && (2.64) \end{aligned}$$

Equation (2.63) requires some explanation. Let's first consider how we got the numerator from the preceding equation.

Ask the usual question: How can it happen? In this case, how can the event

$$L_1 = 1 \text{ and } L_2 = 0 \tag{2.65}$$

occur? Since we know a lot about the probabilistic behavior of the B_i, let's try to recast that event. A little thought shows that the event is equivalent to the event

$$B_1 = 0 \text{ and } L_2 = 0 \tag{2.66}$$

So, how did the denominator in (2.63) come from the preceding equation? In other words, how did we recast the event

$$L_2 = 0 \text{ and } L_1 > 0 \tag{2.67}$$

in terms of the B_i? Well, $L_1 > 0$ means that B_1 is either 1 or 2. Thus we broke things down accordingly in the denominator of (2.63).

The remainder of the computation is similar to what we did earlier in this example.

2.13 Random Graph Models

A *graph* consists of *vertices* and *edges*. To understand this, think of a social network. Here the vertices represent people and the edges represent friendships. For the time being, assume that friendship relations are mutual, i.e. if person i says he is friends with person j, then j will say the same about i.

For any graph, its *adjacency matrix* consists of 1 and 0 entries, with a 1 in row i, column j meaning there is an edge from vertex i to vertex j. For instance, say we have a simple tiny network of three people, with adjacency matrix

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \tag{2.68}$$

Row 1 of the matrix says that Person 1 is friends with persons 2 and 3, but we see from the other rows that Persons 2 and 3 are not friends with each other.

In any graph, the *degree* of a vertex is its number of edges. So, the degree of vertex i is the number of 1s in row i. In the little model above, vertex 1 has degree 2 but the other two vertices each have degree 1.

The assumption that friendships are mutual is described in graph theory as having a *undirected* graph. Note that that implies that the adjacency matrix is symmetric. However, we might model some other networks as *directed*, with adjacency matrices that are not necessarily symmetric. In a large extended family, for example, we could define edges in terms of being an elder sibling; there would be an edge from Person i to Person j if j is an older sibling of i.

Graphs need not represent people. They are used in myriad other settings, such as analysis of Web site relations, Internet traffic routing, genetics research and so on.

2.13.1 Example: Preferential Attachment Graph Model

A famous graph model is *Preferential Attachment*. Think of it again as an undirected social network, with each edge representing a "friend" relation. The number of vertices grows over time, one vertex per time step. At time 0, we have just two vertices, v_1 and v_2, with a link between them.

Thus at time 0, each of the two vertices has degree 1. Whenever a new vertex is added to the graph, it randomly chooses an existing vertex to *attach to*, creating a new edge with that existing vertex. In making that random choice, it follows probabilities in proportion to the degrees of the existing edges.

As an example of how the Preferential Attachment Model works, suppose that at time 2, when v_4 is added, the adjacency matrix for the graph is (2.68). Then there will be an edge created between v_4 with v_1, v_2 or v_3, with probability 2/4, 1/4 and 1/4, respectively. Let's find $P(v_4$ attaches to $v_1)$

Let N_i denote the node that v_i attaches to, i = 3,4,... Then, following the solution strategy "break big event down into small events," let's break this question about v_4 according to what happens with v_3:

$$\begin{aligned} P(N_4 = 1) &= P(N_3 = 1 \text{ and } N_4 = 1) + P(N_3 = 2 \text{ and } N_4 = 1) && (2.69) \\ &= (1/2)(2/4) + (1/2)(1/4) && (2.70) \\ &= 3/8 && (2.71) \end{aligned}$$

2.14 Simulation

To simulate whether a simple event occurs or not, we typically use R function **runif()**. This function generates random numbers from the interval (0,1), with all the points inside being equally likely. So for instance the probability that the function returns a value in (0,0.5) is 0.5. Thus here is code to simulate tossing a coin:

```
if (runif(1) < 0.5) heads <- TRUE else heads <- FALSE
```

The argument 1 means we wish to generate just one random number from the interval (0,1).

2.14.1 Example: Rolling Dice

If we roll three dice, what is the probability that their total is 8? We count all the possibilities, or we could get an approximate answer via simulation:

```
1  # roll d dice; find P(total = k)
2
3  probtotk <- function(d,k,nreps) {
4     count <- 0
5     # do the experiment nreps times
6     for (rep in 1:nreps) {
```

```
7        sum <- 0
8        # roll d dice and find their sum
9        for (j in 1:d) sum <- sum + roll()
10       if (sum == k) count <- count + 1
11    }
12    return(count/nreps)
13 }
14
15 # simulate roll of one die; the possible return values are 1,2,3,4,5,6,
16 # all equally likely
17 roll <- function() return(sample(1:6,1))
18
19 # example
20 probtotk(3,8,1000)
```

The call to the built-in R function **sample()** here says to take a sample of size 1 from the sequence of numbers 1,2,3,4,5,6. That's just what we want to simulate the rolling of a die. The code

```
for (j in 1:d) sum <- sum + roll()
```

then simulates the tossing of a die d times, and computing the sum.

2.14.2 First Improvement

Since applications of R often use large amounts of computer time, good R programmers are always looking for ways to speed things up. Here is an alternate version of the above program:

```
1  # roll d dice; find P(total = k)
2
3  probtotk <- function(d,k,nreps) {
4     count <- 0
5     # do the experiment nreps times
6     for (rep in 1:nreps)
7        total <- sum(sample(1:6,d,replace=TRUE))
8        if (total == k) count <- count + 1
9     }
10    return(count/nreps)
11 }
```

Let's first discuss the code.

```
sample(1:6,d,replace=TRUE)
```

The call to **sample()** here says, "Generate d random numbers, chosen randomly (i.e. with equal probability) from the integers 1 through 6, with replacement." Well, of course, that simulates

tossing the die d times. So, that call returns a d-element array, and we then call R's built-in function **sum()** to find the total of the d dice.

Note the call to R's **sum()** function, a nice convenience.

This second version of the code here eliminates one explicit loop, which is the key to writing fast code in R. But just as important, it is more compact and clearer in expressing what we are doing in this simulation.

2.14.3 Second Improvement

Further improvements are possible. Consider this code:

```
1  # roll d dice; find P(total = k)
2
3  # simulate roll of nd dice; the possible return values are 1,2,3,4,5,6,
4  # all equally likely
5  roll <- function(nd) return(sample(1:6,nd,replace=TRUE))
6
7  probtotk <- function(d,k,nreps) {
8     sums <- vector(length=nreps)
9     # do the experiment nreps times
10    for (rep in 1:nreps) sums[rep] <- sum(roll(d))
11    return(mean(sums==k))
12 }
```

There is quite a bit going on here.

We are storing the various "notebook lines" in a vector **sums**. We first call **vector()** to allocate space for it.

But the heart of the above code is the expression **sums==k**, which involves the very essence of the R idiom, **vectorization**. At first, the expression looks odd, in that we are comparing a vector (remember, this is what languages like C call an *array*), **sums**, to a scalar, **k**. But in R, every "scalar" is actually considered a one-element vector.

Fine, **k** is a vector, but wait! It has a different length than **sums**, so how can we compare the two vectors? Well, in R a vector is **recycled**—extended in length, by repeating its values—in order to conform to longer vectors it will be involved with. For instance:

```
> c(2,5) + 4:6
[1]  6 10  8
```

Here we added the vector (2,5) to (4,5,6). The former was first recycled to (2,5,2), resulting in a

sum of (6,10,8).[7]

So, in evaluating the expression **sums==k**, R will recycle **k** to a vector consisting of **nreps** copies of **k**, thus conforming to the length of **sums**. The result of the comparison will then be a vector of length **nreps**, consisting of TRUE and FALSE values. In numerical contexts, these are treated at 1s and 0s, respectively. R's **mean()** function will then average those values, resulting in the fraction of 1s! That's exactly what we want.

2.14.4 Third Improvement

Even better:

```
1  roll <- function(nd) return(sample(1:6,nd,replace=TRUE))
2
3  probtotk <- function(d,k,nreps) {
4     # do the experiment nreps times
5     sums <- replicate(nreps,sum(roll(d)))
6     return(mean(sums==k))
7  }
```

R's **replicate()** function does what its name implies, in this case executing the call **sum(roll(d))**. That produces a vector, which we then assign to **sums**. And note that we don't have to allocate space for **sums**; **replicate()** produces a vector, allocating space, and then we merely point **sums** to that vector.

The various improvements shown above compactify the code, and in many cases, make it much faster.[8] Note, though, that this comes at the expense of using more memory.

2.14.5 Simulation of Conditional Probability in Dice Problem

Suppose three fair dice are rolled. We wish to find the approximate probability that the first die shows fewer than 3 dots, given that the total number of dots for the 3 dice is more than 8, using simulation.

Here is the code:

```
1  dicesim <- function(nreps) {
2     count1 <- 0
3     count2 <- 0
```

[7] There was also a warning message, not shown here. The circumstances under which warnings are or are not generated are beyond our scope here, but recycling is a very common R operation.

[8] You can measure times using R's **system.time()** function, e.g. via the call **system.time(probtotk(3,7,10000))**.

```
4     for (i in 1:nreps) {
5        d <- sample(1:6,3,replace=T)
6        if (sum(d) > 8) {
7           count1 <- count1 + 1
8           if (d[1] < 3) count2 <- count2 + 1
9        }
10    }
11    return(count2 / count1)
12 }
```

Note carefully that we did NOT use (2.8). That would defeat the purpose of simulation, which is the model the actual process.

2.14.6 Simulation of the ALOHA Example

Following is a computation via simulation of the *approximate* values of $P(X_1 = 2)$, $P(X_2 = 2)$ and $P(X_2 = 2|X_1 = 1)$.

```
1  # finds P(X1 = 2), P(X2 = 2) and P(X2 = 2|X1 = 1) in ALOHA example
2  sim <- function(p,q,nreps) {
3     countx2eq2 <- 0
4     countx1eq1 <- 0
5     countx1eq2 <- 0
6     countx2eq2givx1eq1 <- 0
7     # simulate nreps repetitions of the experiment
8     for (i in 1:nreps) {
9        numsend <- 0  # no messages sent so far
10       # simulate A and B's decision on whether to send in epoch 1
11       for (j in 1:2)
12          if (runif(1) < p) numsend <- numsend + 1
13       if (numsend == 1)  X1 <- 1
14       else X1 <- 2
15       if (X1 == 2) countx1eq2 <- countx1eq2 + 1
16       # now simulate epoch 2
17       # if X1 = 1 then one node may generate a new message
18       numactive <- X1
19       if (X1 == 1 && runif(1) < q) numactive <- numactive + 1
20       # send?
21       if (numactive == 1)
22          if (runif(1) < p) X2 <- 0
23          else X2 <- 1
24       else {  # numactive = 2
25          numsend <- 0
26          for (i in 1:2)
27             if (runif(1) < p) numsend <- numsend + 1
28          if (numsend == 1) X2 <- 1
29          else X2 <- 2
30       }
31       if (X2 == 2) countx2eq2 <- countx2eq2 + 1
```

```
32        if (X1 == 1) {  # do tally for the cond. prob.
33           countx1eq1 <- countx1eq1 + 1
34           if (X2 == 2) countx2eq2givx1eq1 <- countx2eq2givx1eq1 + 1
35        }
36     }
37     # print results
38     cat("P(X1 = 2):",countx1eq2/nreps,"\n")
39     cat("P(X2 = 2):",countx2eq2/nreps,"\n")
40     cat("P(X2 = 2 | X1 = 1):",countx2eq2givx1eq1/countx1eq1,"\n")
41  }
```

Note that each of the **nreps** iterations of the main **for** loop is analogous to one line in our hypothetical notebook. So, to find (the approximate value of) $P(X_1 = 2)$, divide the count of the number of times $X_1 = 2$ occurred by the number of iterations.

Again, note especially that the way we calculated $P(X_2 = 2|X_1 = 1)$ was to count the number of times $X_2 = 2$, **among those times that** $X_1 = 1$, just like in the notebook case.

Also: Keep in mind that we did NOT use (2.25) or any other formula in our simulation. We stuck to basics, the "notebook" definition of probability. This is really important if you are using simulation to confirm something you derived mathematically. On the other hand, if you are using simulation because you CAN'T derive something mathematically (the usual situation), using some of the mailing tubes might speed up the computation.

2.14.7 Example: Bus Ridership, cont'd.

Consider the example in Section 2.12. Let's find the probability that after visiting the tenth stop, the bus is empty. This is too complicated to solve analytically, but can easily be simulated:

```
1  nreps <- 10000
2  nstops <- 10
3  count <- 0
4  for (i in 1:nreps) {
5     passengers <- 0
6     for (j in 1:nstops) {
7        if (passengers > 0)
8           for (k in 1:passengers)
9              if (runif(1) < 0.2)
10                passengers <- passengers - 1
11       newpass <- sample(0:2,1,prob=c(0.5,0.4,0.1))
12       passengers <- passengers + newpass
13    }
14    if (passengers == 0) count <- count + 1
15 }
16 print(count/nreps)
```

Note the different usage of the **sample()** function in the call

```
sample(0:2,1,prob=c(0.5,0.4,0.1))
```

Here we take a sample of size 1 from the set {0,1,2}, but with probabilities 0.5 and so on. Since the third argument for **sample()** is **replace**, not **prob**, we need to specify the latter in our call.

2.14.8 Example: Board Game con'd.

Recall the board game in Section 2.11. Below is simulation code to find the probability in (2.45):

```
1  boardsim <- function(nreps) {
2     count4 <- 0
3     countbonusgiven4 <- 0
4     for (i in 1:nreps) {
5        position <- sample(1:6,1)
6        if (position == 3) {
7           bonus <- TRUE
8           position <- (position + sample(1:6,1)) %% 8
9        } else bonus <- FALSE
10       if (position == 4) {
11          count4 <- count4 + 1
12          if (bonus) countbonusgiven4 <- countbonusgiven4 + 1
13       }
14    }
15    return(countbonusgiven4/count4)
16 }
```

2.14.9 Example: Broken Rod

Say a glass rod drops and breaks into 5 random pieces. Let's find the probability that the smallest piece has length below 0.02.

First, what does "random" mean here? Let's assume that the break points, treating the left end as 0 and the right end as 1, can be modeled with **runif()**. Here then is code to do the job:

```
# random breaks the rod in k pieces, returning the length of the
# shortest one
minpiece <- function(k) {
   breakpts <- sort(runif(k-1))
   lengths <- diff(c(0,breakpts,1))
   min(lengths)
}

# returns the approximate probability that the smallest of k pieces will
# have length less than q
```

```
bkrod <- function(nreps,k,q) {
   minpieces <- replicate(nreps,minpiece(k))
   mean(minpieces < q)
}

> bkrod(10000,5,0.02)
[1] 0.35
```

So, we generate the break points according to the model, then sort them in order to call R's **diff()** function. The latter finds differences between successvie values of its argument, which in our case will give us the lengths of the pieces. We then find the minimum length.

2.14.10 How Long Should We Run the Simulation?

Clearly, the large the value of **nreps** in our examples above, the more accurate our simulation results are likely to be. But how large should this value be? Or, more to the point, what measure is there for the degree of accuracy one can expect (whatever that means) for a given value of **nreps**? These questions will be addressed in Chapter 20.

2.15 Combinatorics-Based Probability Computation

And though the holes were rather small, they had to count them all—from the Beatles song, *A Day in the Life*

In some probability problems all the outcomes are equally likely. The probability computation is then simply a matter of counting all the outcomes of interest and dividing by the total number of possible outcomes. Of course, sometimes even such counting can be challenging, but it is simple in principle. We'll discuss two examples here.

2.15.1 Which Is More Likely in Five Cards, One King or Two Hearts?

Suppose we deal a 5-card hand from a regular 52-card deck. Which is larger, P(1 king) or P(2 hearts)? Before continuing, take a moment to guess which one is more likely.

Now, here is how we can compute the probabilities. **The key point is that all possible hands are equally likely, which implies that all we need to do is count them.** There are $\binom{52}{5}$ possible hands, so this is our denominator. For P(1 king), our numerator will be the number of hands consisting of one king and four non-kings. Since there are four kings in the deck, the number

of ways to choose one king is $\binom{4}{1} = 4$. There are 48 non-kings in the deck, so there are $\binom{48}{4}$ ways to choose them. Every choice of one king can be combined with every choice of four non-kings, so the number of hands consisting of one king and four non-kings is $4 \cdot \binom{48}{4}$. Thus

$$P(1 \text{ king}) = \frac{4 \cdot \binom{48}{4}}{\binom{52}{5}} = 0.299 \tag{2.72}$$

The same reasoning gives us

$$P(2 \text{ hearts}) = \frac{\binom{13}{2} \cdot \binom{39}{3}}{\binom{52}{5}} = 0.274 \tag{2.73}$$

So, the 1-king hand is just slightly more likely.

Note that an unstated assumption here was that all 5-card hands are equally likely. That *is* a realistic assumption, but it's important to understand that it plays a key role here.

By the way, I used the R function **choose()** to evaluate these quantities, running R in interactive mode, e.g.:

```
> choose(13,2) * choose(39,3) / choose(52,5)
[1] 0.2742797
```

R also has a very nice function **combn()** which will generate all the $\binom{n}{k}$ combinations of k things chosen from n, and also will at your option call a user-specified function on each combination. This allows you to save a lot of computational work. See the examples in R's online documentation.

Here's how we could do the 1-king problem via simulation:

```
1  # use simulation to find P(1 king) when deal a 5-card hand from a
2  # standard deck
3
4  # think of the 52 cards as being labeled 1-52, with the 4 kings having
5  # numbers 1-4
6
7  sim <- function(nreps) {
8     count1king <- 0  # count of number of hands with 1 king
9     for (rep in 1:nreps) {
10       hand <- sample(1:52,5,replace=FALSE)  # deal hand
11       kings <- intersect(1:4,hand)  # find which kings, if any, are in hand
12       if (length(kings) == 1) count1king <- count1king + 1
13    }
14    print(count1king/nreps)
15 }
```

Here the **intersect()** function performs set intersection, in this case the set 1,2,3,4 and the one in the variable **hand**. Applying the `length()` function then gets us number of kings.

2.15.2 Example: Random Groups of Students

A class has 68 students, 48 of which are CS majors. The 68 students will be randomly assigned to groups of 4. Find the probability that a random group of 4 has exactly 2 CS majors.

$$\frac{\binom{48}{2}\binom{20}{2}}{\binom{68}{4}}$$

2.15.3 Example: Lottery Tickets

Twenty tickets are sold in a lottery, numbered 1 to 20, inclusive. Five tickets are drawn for prizes. Let's find the probability that two of the five winning tickets are even-numbered.

Since there are 10 even-numbered tickets, there are $\binom{10}{2}$ sets of two such tickets. Continuing along these lines, we find the desired probability to be.

$$\frac{\binom{10}{2}\binom{10}{3}}{\binom{20}{5}} \tag{2.74}$$

Now let's find the probability that two of the five winning tickets are in the range 1 to 5, two are in 6 to 10, and one is in 11 to 20.

Picture yourself picking your tickets. Again there are $\binom{20}{5}$ ways to choose the five tickets. How many of those ways satisfy the stated condition?

Well, first, there are $\binom{5}{2}$ ways to choose two tickets from the range 1 to 5. Once you've done that, there are $\binom{5}{2}$ ways to choose two tickets from the range 6 to 10, and so on. So, The desired probability is then

$$\frac{\binom{5}{2}\binom{5}{2}\binom{10}{1}}{\binom{20}{5}} \tag{2.75}$$

2.15.4 "Association Rules" in Data Mining

The field of *data mining* is a branch of computer science, but it is largely an application of various statistical methods to really huge databases.

One of the applications of data mining is called the *market basket* problem. Here the data consists of records of sales transactions, say of books at Amazon.com. The business' goal is exemplified by Amazon's suggestion to customers that "Patrons who bought this book also tended to buy the following books."[9] The goal of the market basket problem is to sift through sales transaction records to produce *association rules*, patterns in which sales of some combinations of books imply likely sales of other related books.

The notation for association rules is $A, B \Rightarrow C, D, E$, meaning in the book sales example that customers who bought books A and B also tended to buy books C, D and E. Here A and B are called the **antecedents** of the rule, and C, D and E are called the **consequents**. Let's suppose here that we are only interested in rules with a single consequent.

We will present some methods for finding good rules in another chapter, but for now, let's look at how many possible rules there are. Obviously, it would be impractical to use rules with a large number of antecedents.[10]. Suppose the business has a total of 20 products available for sale. What percentage of potential rules have three or fewer antecedents?[11]

For each k = 1,...,19, there are $\binom{20}{k}$ possible sets of k antecedents, and for each such set there are $\binom{20-k}{1}$ possible consequents. The fraction of potential rules using three or fewer antecedents is then

$$\frac{\sum_{k=1}^{3} \binom{20}{k} \cdot \binom{20-k}{1}}{\sum_{k=1}^{19} \binom{20}{k} \cdot \binom{20-k}{1}} = \frac{23180}{10485740} = 0.0022 \tag{2.76}$$

So, this is just scratching the surface. And note that with only 20 products, there are already over ten million possible rules. With 50 products, this number is 2.81×10^{16}! Imagine what happens in a case like Amazon, with millions of products. These staggering numbers show what a tremendous challenge data miners face.

[9] Some customers appreciate such tips, while others view it as insulting or an invasion of privacy, but we'll not address such issues here.

[10] In addition, there are serious statistical problems that would arise, to be discussed in another chapter.

[11] Be sure to note that this is also a probability, namely the probability that a randomly chosen rule will have three or fewer antecedents.

2.15.5 Multinomial Coefficients

Question: We have a group consisting of 6 Democrats, 5 Republicans and 2 Independents, who will participate in a panel discussion. They will be sitting at a long table. How many seating arrangements are possible, with regard to political affiliation? (So we do not care, for instance, about permuting the individual Democrats within the seats assigned to Democrats.)

Well, there are $\binom{13}{6}$ ways to choose the Democratic seats. Once those are chosen, there are $\binom{7}{5}$ ways to choose the Republican seats. The Independent seats are then already determined, i.e. there will be only way at that point, but let's write it as $\binom{2}{2}$. Thus the total number of seating arrangements is

$$\frac{13!}{6!7!} \cdot \frac{7!}{5!2!} \cdot \frac{2!}{2!0!} \tag{2.77}$$

That reduces to

$$\frac{13!}{6!5!2!} \tag{2.78}$$

The same reasoning yields the following:

Multinomial Coefficients: Suppose we have c objects and r bins. Then the number of ways to choose c_1 of them to put in bin 1, c_2 of them to put in bin 2,..., and c_r of them to put in bin r is

$$\frac{c!}{c_1!...c_r!}, \quad c_1 + ... + c_r = c \tag{2.79}$$

Of course, the "bins" may just be metaphorical. In the political party example above, the "bins " were political parties, and "objects" were seats.

2.15.6 Example: Probability of Getting Four Aces in a Bridge Hand

A standard deck of 52 cards is dealt to four players, 13 cards each. One of the players is Millie. What is the probability that Millie is dealt all four aces?

Well, there are

$$\frac{52!}{13!13!13!13!} \tag{2.80}$$

possible deals. (the "objects" are the 52 cards, and the "bins" are the 4 players.) The number of deals in which Millie holds all four aces is the same as the number of deals of 48 cards, 9 of which go to Millie and 13 each to the other three players, i.e.

$$\frac{48!}{13!13!13!9!} \tag{2.81}$$

Thus the desired probability is

$$\frac{\frac{48!}{13!13!13!9!}}{\frac{52!}{13!13!13!13!}} = 0.00264 \tag{2.82}$$

Exercises

1. This problem concerns the ALOHA network model of Section 2.1. Feel free to use (but cite) computations already in the example.

(a) $P(X_1 = 2 \text{ and } X_2 = 1)$, for the same values of p and q in the examples.

(b) Find $P(X_2 = 0)$.

(c) Find $(P(X_1 = 1|X_2 = 1)$.

2. Urn I contains three blue marbles and three yellow ones, while Urn II contains five and seven of these colors. We draw a marble at random from Urn I and place it in Urn II. We then draw a marble at random from Urn II.

(a) Find P(second marble drawn is blue).

(b) Find P(first marble drawn is blue | second marble drawn is blue).

3. Consider the example of association rules in Section 2.15.4. How many two-antecedent, two-consequent rules are possible from 20 items? Express your answer in terms of combinatorial ("n choose k") symbols.

4. Suppose 20% of all C++ programs have at least one major bug. Out of five programs, what is the probability that exactly two of them have a major bug?

5. Assume the ALOHA network model as in Section 2.1, i.e. m = 2 and $X_0 = 2$, but with general values for p and q. Find the probability that a new message is created during epoch 2.

6. You bought three tickets in a lottery, for which 60 tickets were sold in all. There will be five prizes given. Find the probability that you win at least one prize, and the probability that you win exactly one prize.

7. Two five-person committees are to be formed from your group of 20 people. In order to foster communication, we set a requirement that the two committees have the same chair but no other overlap. Find the probability that you and your friend are both chosen for some committee.

8. Consider a device that lasts either one, two or three months, with probabilities 0.1, 0.7 and 0.2, respectively. We carry one spare. Find the probability that we have some device still working just before four months have elapsed.

9. A building has six floors, and is served by two freight elevators, named Mike and Ike. The destination floor of any order of freight is equally likely to be any of floors 2 through 6. Once an elevator reaches any of these floors, it stays there until summoned. When an order arrives to the building, whichever elevator is currently closer to floor 1 will be summoned, with elevator Ike being the one summoned in the case in which they are both on the same floor.

Find the probability that after the summons, elevator Mike is on floor 3. Assume that only one order of freight can fit in an elevator at a time. Also, suppose the average time between arrivals of freight to the building is much larger than the time for an elevator to travel between the bottom and top floors; this assumption allows us to neglect travel time.

10. Without resorting to using the fact that $\binom{n}{k} = n!/[k!(n-k!)]$, find c and d such that

$$\binom{n}{k} = \binom{n-1}{k} + \binom{c}{d} \tag{2.83}$$

11. Consider the ALOHA example from the text, for general p and q, and suppose that $X_0 = 0$, i.e. there are no active nodes at the beginning of our observation period. Find $P(X_1 = 0)$.

12. Consider a three-sided die, as opposed to the standard six-sided type. The die is cylinder-shaped, and gives equal probabilities to one, two and three dots. The game is to keep rolling the die until we get a total of at least 3. Let N denote the number of times we roll the die. For example, if we get a 3 on the first roll, N = 1. If we get a 2 on the first roll, then N will be 2 no matter what we get the second time. The largest N can be is 3. The rule is that one wins if one's final total is exactly 3.

(a) Find the probability of winning.

(b) Find P(our first roll was a 1 | we won).

(c) How could we construct such a die?

13. Consider the ALOHA simulation example in Section 2.14.6.

(a) Suppose we wish to find $P(X_2 = 1|X_1 = 1)$ instead of $P(X_2 = 2|X_1 = 1)$. What line(s) would we change, and how would we change them?

(b) In which line(s) are we in essence checking for a collision?

14. Jack and Jill keep rolling a four-sided and a three-sided die, respectively. The first player to get the face having just one dot wins, except that if they both get a 1, it's a tie, and play continues. Let N denote the number of turns needed. Find the following:

(a) P(N = 1), P(N = 2).

(b) P(the first turn resulted in a tie$|N = 2$)

15. In the ALOHA network example in Sec. 1.1, suppose $X_0 = 1$, i.e. we start out with just one active node. Find $P(X_2 = 0)$, as an expression in p and q.

16. Suppose a box contains two pennies, three nickels and five dimes. During transport, two coins fall out, unseen by the bearer. Assume each type of coin is equally likely to fall out. Find: P(at least \$0.10 worth of money is lost); P(both lost coins are of the same denomination)

17. Suppose we have the track record of a certain weather forecaster. Of the days for which he predicts rain, a fraction c actually do have rain. Among days for which he predicts no rain, he is correct a fraction d of the time. Among all days, he predicts rain g of the time, and predicts no rain 1-g of the time. Find P(he predicted rain | it does rain), P(he predicts wrong) and P(it does rain — he was wrong). Write R simulation code to verify. (Partial answer: For the case c = 0.8, d = 0.6 and g = 0.2, P(he predicted rain | it does rain) = 1/3.)

18. The Game of Pit is really fun because there are no turns. People shout out bids at random, chaotically. Here is a slightly simplified version of the game:

There are four suits, Wheat, Barley, Corn and Rye, with nine cards each, 36 cards in all. There are four players. At the opening, the cards are all dealt out, nine to each player. The players hide their cards from each other's sight.

Players then start trading. In computer science terms, trading is asynchronous, no turns; a player can bid at any time. The only rule is that a trade must be homogeneous in suit, e.g. all Rye. (The player trading Rye need not trade all the Rye he has, though.) The player bids by shouting out the number she wants to trade, say "2!" If another player wants to trade two cards (again, homogeneous in suit), she yells out, "OK, 2!" and they trade. When one player acquires all nine of a suit, he shouts "Corner!"

Consider the situation at the time the cards have just been dealt. Imagine that you are one of the players, and Jane is another. Find the following probabilities:

(a) P(you have no Wheats).

(b) P(you have seven Wheats).

(c) P(Jane has two Wheats — you have seven Wheats).

(d) P(you have a corner) (note: someone else might too; whoever shouts it out first wins).

19. In the board game example in Section 2.11, suppose that the telephone report is that A ended up at square 1 after his first turn. Find the probability that he got a bonus.

20. Consider the bus ridership example in Section 2.12 of the text. Suppose the bus is initially empty, and let X_n denote the number of passengers on the bus just after it has left the n^{th} stop, n = 1,2,... Find the following:

(a) $P(X_2 = 1)$

(b) P(at least one person boarded the bus at the first stop $|X_2 = 1$)

21. Suppose committees of sizes 3, 4 and 5 are to be chosen at random from 20 people, among whom are persons A and B. Find the probability that A and B are on the same committee.

22. Consider the ALOHA simulation in Section 28.

(a) On what line do we simulate the possible creation of a new message?

(b) Change line 10 so that it uses **sample()** instead of **runif()**.

Chapter 3

Discrete Random Variables

This chapter will introduce entities called *discrete random variables*. Some properties will be derived for means of such variables, with most of these properties actually holding for random variables in general. Well, all of that seems abstract to you at this point, so let's get started.

3.1 Random Variables

Definition 3 *A* **random variable** *is a numerical outcome of our experiment.*

For instance, consider our old example in which we roll two dice, with X and Y denoting the number of dots we get on the blue and yellow dice, respectively. Then X and Y are random variables, as they are numerical outcomes of the experiment. Moreover, X+Y, 2XY, sin(XY) and so on are also random variables.

In a more mathematical formulation, with a formal sample space defined, a random variable would be defined to be a real-valued function whose domain is the sample space.

3.2 Discrete Random Variables

In our dice example, the random variable X could take on six values in the set {1,2,3,4,5,6}. We say that the **support** of X is {1,2,3,4,5,6}. This is a finite set.

In the ALOHA example, X_1 and X_2 each have support {0,1,2}, again a finite set.[1]

[1] We could even say that X_1 takes on only values in the set {1,2}, but if we were to look at many epochs rather than just two, it would be easier not to make an exceptional case.

Now think of another experiment, in which we toss a coin until we get heads. Let N be the number of tosses needed. Then the support of N is the set {1,2,3,...} This is a countably infinite set.[2]

Now think of one more experiment, in which we throw a dart at the interval (0,1), and assume that the place that is hit, R, can take on any of the values between 0 and 1. Here the support is an uncountably infinite set.

We say that X, X_1, X_2 and N are **discrete** random variables, while R is **continuous**. We'll discuss continuous random variables in a later chapter.

3.3 Independent Random Variables

We already have a definition for the independence of events; what about independence of random variables? Here it is:

> Random variables X and Y are said to be **independent** if for any sets I and J, the events {X is in I} and {Y is in J} are independent, i.e. P(X is in I and Y is in J) = P(X is in I) P(Y is in J).

Sounds innocuous, but the notion of independent random variables is absolutely central to the field of probability and statistics, and will pervade this entire book.

3.4 Example: The Monty Hall Problem

This is an example of how the use of random variables in "translating" a probability problem to mathematical terms can simplify and clarify one's thinking. Imagine, **this simple device of introducing named random variables into our analysis makes a problem that has vexed famous mathematicians quite easy to solve!**

The Monty Hall Problem, which gets its name from a popular TV game show host, involves a contestant choosing one of three doors. Behind one door is a new automobile, while the other two doors lead to goats. The contestant chooses a door and receives the prize behind the door.

The host knows which door leads to the car. To make things interesting, after the contestant chooses, the host will open one of the other doors not chosen, showing that it leads to a goat.

[2]This is a concept from the fundamental theory of mathematics. Roughly speaking, it means that the set can be assigned an integer labeling, i.e. item number 1, item number 2 and so on. The set of positive even numbers is countable, as we can say 2 is item number 1, 4 is item number 2 and so on. It can be shown that even the set of all rational numbers is countable.

Should the contestant now change her choice to the remaining door, i.e. the one that she didn't choose and the host didn't open?

Many people answer No, reasoning that the two doors not opened yet each have probability 1/2 of leading to the car. But the correct answer is actually that the remaining door (not chosen by the contestant and not opened by the host) has probability 2/3, and thus the contestant should switch to it. Let's see why.

Let

- C = contestant's choice of door (1, 2 or 3)
- H = host's choice of door (1, 2 or 3)
- A = door that leads to the automobile

We can make things more concrete by considering the case $C = 1,\ H = 2$. The mathematical formulation of the problem is then to find

$$P(A = 3 \mid C = 1,\ H = 2) = \frac{P(A = 3,\ C = 1,\ H = 2)}{P(C = 1,\ H = 2)} \tag{3.1}$$

The key point, commonly missed even by mathematically sophisticated people, is the role of the host. Write the numerator above as

$$P(A = 3,\ C = 1)\ P(H = 2 \mid A = 3,\ C = 1) \tag{3.2}$$

Since C and A are independent random variables, the value of the first factor in (3.2) is

$$\frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9} \tag{3.3}$$

What about the second factor? Remember, the host knows that $A = 3$, and since the contestant has chosen door 1, the host will open the only remaining door that conceals a goat, i.e. door 2. In other words,

$$P(H = 2 \mid A = 3,\ C = 1) = 1 \tag{3.4}$$

On the other hand, if say $A = 1$, the host would randomly choose between doors 2 and 3, so that

$$P(H = 2 \mid A = 1,\ C = 1) = \frac{1}{2} \tag{3.5}$$

It is left to the reader to complete the analysis, calculating the denominator of (3.1), and then showing in the end that

$$P(A = 3 \mid C = 1,\ H = 2) = \frac{2}{3} \tag{3.6}$$

According to the "Monty Hall problem" entry in Wikipedia, even Paul Erdös, one of the most famous mathematicians in history, gave the wrong answer to this problem. Presumably he would have avoided this by writing out his analysis in terms of random variables, as above, rather than say, a wordy, imprecise and ultimately wrong solution.

3.5 Expected Value

3.5.1 Generality—Not Just for Discrete Random Variables

The concepts and properties introduced in this section form the very core of probability and statistics. **Except for some specific calculations, these apply to both discrete and continuous random variables.**

The properties developed for *variance*, defined later in this chapter, also hold for both discrete and continuous random variables.

3.5.1.1 What Is It?

The term "expected value" is one of the many misnomers one encounters in tech circles. The expected value is actually not something we "expect" to occur. On the contrary, it's often pretty unlikely.

For instance, let H denote the number of heads we get in tossing a coin 1000 times. The expected value, you'll see later, is 500. This is not surprising, given the symmetry of the situation, but P(H = 500) turns out to be about 0.025. In other words, we certainly should not "expect" H to be 500.

Of course, even worse is the example of the number of dots that come up when we roll a fair die. The expected value is 3.5, a value which not only rarely comes up, but in fact never does.

In spite of being misnamed, expected value plays an absolutely central role in probability and statistics.

3.5.2 Definition

Consider a repeatable experiment with random variable X. We say that the **expected value** of X is the long-run average value of X, as we repeat the experiment indefinitely.

In our notebook, there will be a column for X. Let X_i denote the value of X in the ith row of the notebook. Then the long-run average of X is

$$\lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} \tag{3.7}$$

Suppose for instance our experiment is to toss 10 coins. Let X denote the number of heads we get out of 10. We might get four heads in the first repetition of the experiment, i.e. $X_1 = 4$, seven heads in the second repetition, so $X_2 = 7$, and so on. Intuitively, the long-run average value of X will be 5. (This will be proven below.) Thus we say that the expected value of X is 5, and write E(X) = 5.

3.5.3 Existence of the Expected Value

The above defintion puts the cart before the horse, as it presumes that the limit exists. Theoretically speaking, this might not be the case. However, it does exist if the X_i have finite lower and upper bounds, which is always true in the real world. For instance, no person has height of 50 feet, say, and no one has negative height either.

For the remainder of this book, we will usually speak of "the" expected value of a random variable without adding the qualifier "if it exists."

3.5.4 Computation and Properties of Expected Value

Continuing the coin toss example above, let K_{in} be the number of times the value i occurs among $X_1, ..., X_n$, i = 0,...,10, n = 1,2,3,... For instance, $K_{4,20}$ is the number of times we get four heads, in the first 20 repetitions of our experiment. Then

$$\begin{aligned} E(X) &= \lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} && (3.8)\\ &= \lim_{n\to\infty} \frac{0 \cdot K_{0n} + 1 \cdot K_{1n} + 2 \cdot K_{2n} ... + 10 \cdot K_{10,n}}{n} && (3.9)\\ &= \sum_{i=0}^{10} i \cdot \lim_{n\to\infty} \frac{K_{in}}{n} && (3.10) \end{aligned}$$

To understand that second equation, suppose when n = 5 we have 2, 3, 1, 2 and 1 for our values of X_1, X_2, X_3, X_4, X_5,. Then we can group the 2s together and group the 1s together, and write

$$2 + 3 + 1 + 2 + 1 = 2 \times 2 + 2 \times 1 + 1 \times 3 \tag{3.11}$$

But $\lim_{n\to\infty} \frac{K_{in}}{n}$ is the long-run fraction of the time that X = i. In other words, it's P(X = i)! So,

$$E(X) = \sum_{i=0}^{10} i \cdot P(X = i) \tag{3.12}$$

So in general we have:

Property A:

The expected value of a discrete random variable X which takes values in the set A is

$$E(X) = \sum_{c \in A} c \; P(X = c) \tag{3.13}$$

Note that (3.13) is the formula we'll use. The preceding equations were derivation, to motivate the formula. Note too that (3.13) is not the *definition* of expected value; that was in 3.7. It is quite important to distinguish between all of these, in terms of goals.[3]

By the way, note the word *discrete* above. For the case of continuous random variables, the sum in (3.13) will become an integral.

It will be shown in Section 4.3 that in our example above in which X is the number of heads we get in 10 tosses of a coin,

$$P(X = i) = \binom{10}{i} 0.5^i (1 - 0.5)^{10-i} \tag{3.14}$$

So

$$E(X) = \sum_{i=0}^{10} i \binom{10}{i} 0.5^i (1 - 0.5)^{10-i} \tag{3.15}$$

It turns out that E(X) = 5.

[3]The matter is made a little more confusing by the fact that many books do in fact treat (3.13) as the definition, with (3.7) being the consequence.

For X in our dice example,

$$E(X) = \sum_{c=1}^{6} c \cdot \frac{1}{6} = 3.5 \tag{3.16}$$

It is customary to use capital letters for random variables, e.g. X here, and lower-case letters for values taken on by a random variable, e.g. c here. Please adhere to this convention.

By the way, it is also customary to write EX instead of E(X), whenever removal of the parentheses does not cause any ambiguity. An example in which it would produce ambiguity is $E(U^2)$. The expression EU^2 might be taken to mean either $E(U^2)$, which is what we want, or $(EU)^2$, which is not what we want.

For S = X+Y in the dice example,

$$E(S) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + ...12 \cdot \frac{1}{36} = 7 \tag{3.17}$$

In the case of N, tossing a coin until we get a head:

$$E(N) = \sum_{c=1}^{\infty} c \cdot \frac{1}{2^c} = 2 \tag{3.18}$$

(We will not go into the details here concerning how the sum of this particular infinite series is computed. See Section 4.2.)

Some people like to think of E(X) using a center of gravity analogy. Forget that analogy! Think notebook! **Intuitively, E(X) is the long-run average value of X among all the lines of the notebook.** So for instance in our dice example, E(X) = 3.5, where X was the number of dots on the blue die, means that if we do the experiment thousands of times, with thousands of lines in our notebook, the average value of X in those lines will be about 3.5. With S = X+Y, E(S) = 7. This means that in the long-run average in column S in Table 3.1 is 7.

Of course, by symmetry, E(Y) will be 3.5 too, where Y is the number of dots showing on the yellow die. That means we wasted our time calculating in Equation (3.17); we should have realized beforehand that E(S) is $2 \times 3.5 = 7$.

In other words:

Property B:

For any random variables U and V, the expected value of a new random variable D = U+V is the

notebook line	outcome	blue+yellow = 6?	S
1	blue 2, yellow 6	No	8
2	blue 3, yellow 1	No	4
3	blue 1, yellow 1	No	2
4	blue 4, yellow 2	Yes	6
5	blue 1, yellow 1	No	2
6	blue 3, yellow 4	No	7
7	blue 5, yellow 1	Yes	6
8	blue 3, yellow 6	No	9
9	blue 2, yellow 5	No	7

Table 3.1: Expanded Notebook for the Dice Problem

sum of the expected values of U and V:

$$E(U+V) = E(U) + E(V) \tag{3.19}$$

Note carefully that U and V do NOT need to be independent random variables for this relation to hold. You should convince yourself of this fact intuitively **by thinking about the notebook notion.** Say we look at 10000 lines of the notebook, which has columns for the values of U, V and U+V. It makes no difference whether we average U+V in that column, or average U and V in their columns and then add—either way, we'll get the same result.

While you are at it, use the notebook notion to convince yourself of the following:

Properties C:

- For any random variable U and constant a, then

$$E(aU) = aEU \tag{3.20}$$

- For random variables X and Y—not necessarily independent—and constants a and b, we have

$$E(aX+bY) = aEX + bEY \tag{3.21}$$

  This follows by taking U = aX and V = bY in (3.19), and then using (3.20).

By induction, for constants $a_1, ..., a_k$ and random variables $X_1, ..., X_k$, form the new random variable $a_1X_1 + ... + a_kX_k$. Then

$$E(a_1X_1 + ... + a_nX_k) = a_1EX_1 + ... + a_nEX_k \tag{3.22}$$

- For any constant b, we have

$$E(b) = b \tag{3.23}$$

For instance, say U is temperature in Celsius. Then the temperature in Fahrenheit is $W = \frac{9}{5}U + 32$. So, W is a new random variable, and we can get its expected value from that of U by using (3.21) with $a = \frac{9}{5}$ and b = 32.

If you combine (3.23) with (3.21), we have an important special case:

$$E(aX + b) = aEX + b \tag{3.24}$$

Another important point:

Property D: If U and V *are* independent, then

$$E(UV) = EU \cdot EV \tag{3.25}$$

In the dice example, for instance, let D denote the product of the numbers of blue dots and yellow dots, i.e. D = XY. Then

$$E(D) = 3.5^2 = 12.25 \tag{3.26}$$

Equation (3.25) doesn't have an easy "notebook proof." It is proved in Section 17.3.1.

Consider a function $g()$ of one variable, and let $W = g(X)$. W is then a random variable too. Say X has support A, as in (3.13). Then W has support $B = \{g(c) : c \epsilon A\}$. (

For instance, say *g()* is the squaring function, and X takes on the values -1, 0 and 1, with probability 0.5, 0.4 and 0.1. Then

$$A = \{-1, 0, 1\} \tag{3.27}$$

and

$$B = \{0, 1\} \tag{3.28}$$

Define

$$A_d = \{c : c \in A, g(c) = d\} \tag{3.29}$$

In our above squaring example, we will have

$$A_0 = \{0\}, \quad A_1 = \{-1, 1\} \tag{3.30}$$

Then

$$P(W = d) = P(X \in A_d) \tag{3.31}$$

so

$$\begin{aligned} E[g(X)] &= E(W) && (3.32)\\ &= \sum_{d \in B} dP(W = d) && (3.33)\\ &= \sum_{d \in B} d \sum_{c \in A_d} P(X = c) && (3.34)\\ &= \sum_{c \in A} g(c) P(X = c) && (3.35) \end{aligned}$$

(Going from the next-to-last equation here to the last one is rather tricky. Work through for the case of our squaring function example above in order to see why the final equation does follow.

Property E:

If E[g(X)] exists, then

$$E[g(X)] = \sum_{c \in A} g(c) \cdot P(X = c) \tag{3.36}$$

where the sum ranges over all values c that can be taken on by X.

For example, suppose for some odd reason we are interested in finding $E(\sqrt{X})$, where **X** is the number of dots we get when we roll one die. Let $W = \sqrt{X}$). Then **W** is another random variable, and is discrete, since it takes on only a finite number of values. (The fact that most of the values are not integers is irrelevant.) We want to find EW.

Well, W is a function of X, with $g(t) = \sqrt{t}$. So, (3.36) tells us to make a list of values in the support of W, i.e. $\sqrt{1}, \sqrt{2}, ..., \sqrt{6}$, and a list of the corresponding probabilities for **X**, which are all $\frac{1}{6}$. Substituting into (3.36), we find that

$$E(\sqrt{X}) = \frac{1}{6} \sum_{i=1}^{6} \sqrt{i} \tag{3.37}$$

What about a function of several variables? Say for instance you are finding E(UV), where U has support, say, 1,2 and V has support 5,12,13. In order to find E(UV), you need to know the support of UV, recognizing that it, the product UV, is a new random variable in its own right. Let's call it W. Then in this little example, W has support 5,12,13,10,24,26. Then compute

$$5P(W = 5) + 12P(W = 12) + ... = 5P(U = 1, V = 5) + 12P(U = 1, V = 12) + ...$$

Note: Equation (3.36) will be one of the most heavily used formulas in this book. Make sure you keep it in mind.

3.5.5 "Mailing Tubes"

The properties of expected value discussed above are key to the entire remainder of this book. You should notice immediately when you are in a setting in which they are applicable. For instance, if you see the expected value of the sum of two random variables, you should instinctively think of (3.19) right away.

As discussed in Section 2.4, these properties are "mailing tubes." For instance, (3.19) is a "mailing tube"–make a mental note to yourself saying, "If I ever need to find the expected value of the sum of two random variables, I can use (3.19)." Similarly, (3.36) is a mailing tube; tell yourself, "If I ever see a new random variable that is a function of one whose probabilities I already know, I can find the expected value of the new random variable using (3.36)."

You will encounter "mailing tubes" throughout this book. For instance, (3.49) below is a very important "mailing tube." Constantly remind yourself—"Remember the 'mailing tubes'!"

3.5.6 Casinos, Insurance Companies and "Sum Users," Compared to Others

The expected value is intended as a **measure of central tendency** (also called a **measure of location**, i.e. as some sort of definition of the probablistic "middle" in the range of a random variable. There are various other such measures one can use, such as the **median**, the halfway point of a distribution, and today they are recognized as being superior to the mean in certain senses. For historical reasons, the mean plays an absolutely central role in probability and statistics. Yet one should understand its limitations. (This discussion will be general, not limited to discrete random variables.)

(**Warning:** The concept of the mean is likely so ingrained in your consciousness that you simply take it for granted that you know what the mean means, no pun intended. But try to take a step back, and think of the mean afresh in what follows.)

First, the term *expected value* itself is a misnomer. We do not expect the number of dots D to be 3.5 in the die example in Section 3.5.1.1; in fact, it is impossible for W to take on that value.

Second, the expected value is what we call the **mean** in everyday life. And the mean is terribly overused. Consider, for example, an attempt to describe how wealthy (or not) people are in the city of Davis. If suddenly Bill Gates were to move into town, that would skew the value of the mean beyond recognition.

But even without Gates, there is a question as to whether the mean has that much meaning. After all, what is so meaningful about summing our data and dividing by the number of data points? The median has an easy intuitive meaning, but although the mean has familiarity, one would be hard pressed to justify it as a measure of central tendency.

What, for example, does Equation (3.7) mean in the context of people's heights in Davis? We would sample a person at random and record his/her height as X_1. Then we'd sample another person, to get X_2, and so on. Fine, but in that context, what would (3.7) mean? The answer is, not much. So the significance of the mean height of people in Davis would be hard to explain.

For a casino, though, (3.7) means plenty. Say X is the amount a gambler wins on a play of a roulette wheel, and suppose (3.7) is equal to \$1.88. Then after, say, 1000 plays of the wheel (not necessarily by the same gambler), the casino knows from 3.7 it will have paid out a total of about \$1,880. So if the casino charges, say \$1.95 per play, it will have made a profit of about \$70 over those 1000 plays. It might be a bit more or less than that amount, but the casino can be pretty sure that it will be around \$70, and they can plan their business accordingly.

The same principle holds for insurance companies, concerning how much they pay out in claims. With a large number of customers, they know ("expect"!) approximately how much they will pay out, and thus can set their premiums accordingly. Here the mean has a tangible, practical meaning.

The key point in the casino and insurance companies examples is that they are interested in *totals*,

such as *total* payouts on a blackjack table over a month's time, or *total* insurance claims paid in a year. Another example might be the number of defectives in a batch of computer chips; the manufacturer is interested in the *total* number of defectives chips produced, say in a month. Since the mean is by definition a *total* (divided by the number of data points), the mean will be of direct interest to casinos etc.

By contrast, in describing how wealthy people of a town are, the total height of all the residents is not relevant. Similarly, in describing how well students did on an exam, the sum of the scores of all the students doesn't tell us much. (Unless the professor gets $10 for each point in the exam scores of each of the students!) A better description for heights and exam scores might be the median height or score.

Nevertheless, the mean has certain mathematical properties, such as (3.19), that have allowed the rich development of the fields of probability and statistics over the years. The median, by contrast, does not have nice mathematical properties. In many cases, the mean won't be too different from the median anyway (barring Bill Gates moving into town), so you might think of the mean as a convenient substitute for the median. The mean has become entrenched in statistics, and we will use it often.

3.6 Variance

As in Section 3.5, the concepts and properties introduced in this section form the very core of probability and statistics. **Except for some specific calculations, these apply to both discrete and continuous random variables.**

3.6.1 Definition

While the expected value tells us the average value a random variable takes on, we also need a measure of the random variable's variability—how much does it wander from one line of the notebook to another? In other words, we want a measure of **dispersion**. The classical measure is **variance**, defined to be the mean squared difference between a random variable and its mean:

Definition 4 *For a random variable U for which the expected values written below exist, the* **variance** *of U is defined to be*

$$Var(U) = E[(U - EU)^2] \tag{3.38}$$

For X in the die example, this would be

$$Var(X) = E[(X - 3.5)^2] \tag{3.39}$$

Remember what this means: We have a random variable **X**, and we're creating a new random variable, $W = (X - 3.5)^2$, which is a function of the old one. We are then finding the expected value of that new random variable W.

In the notebook view, $E[(X - 3.5)^2]$ is the long-run average of the W column:

line	X	W
1	2	2.25
2	5	2.25
3	6	6.25
4	3	0.25
5	5	2.25
6	1	6.25

To evaluate this, apply (3.36) with $g(c) = (c - 3.5)^2$:

$$Var(X) = \sum_{c=1}^{6} (c - 3.5)^2 \cdot \frac{1}{6} = 2.92 \tag{3.40}$$

You can see that variance does indeed give us a measure of dispersion. In the expression $Var(U) = E[(U - EU)^2]$, if the values of U are mostly clustered near its mean, then $(U - EU)^2$ will usually be small, and thus the variance of U will be small; if there is wide variation in U, the variance will be large.

Property F:

$$Var(U) = E(U^2) - (EU)^2 \tag{3.41}$$

The term $E(U^2)$ is again evaluated using (3.36).

Thus for example, if X is the number of dots which come up when we roll a die. Then, from (3.41),

$$Var(X) = E(X^2) - (EX)^2 \tag{3.42}$$

Let's find that first term (we already know the second is 3.5^2). From (3.36),

$$E(X^2) = \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} = \frac{91}{6} \tag{3.43}$$

Thus $Var(X) = E(X^2) - (EX)^2 = \frac{91}{6} - 3.5^2$

Remember, though, that (3.41) is a shortcut formula for finding the variance, not the *definition* of variance.

Below is the derivation of (3.41). Keep in mind that EU is a constant.

$$\begin{aligned}
Var(U) &= E[(U - EU)^2] && (3.44)\\
&= E[U^2 - 2EU \cdot U + (EU)^2] \text{ (algebra)} && (3.45)\\
&= E(U^2) + E(-2EU \cdot U) + E[(EU)^2] \text{ (3.19)} && (3.46)\\
&= E(U^2) - 2EU \cdot EU + (EU)^2 \text{ (3.20), (3.23)} && (3.47)\\
&= E(U^2) - (EU)^2 && (3.48)
\end{aligned}$$

An important behavior of variance is:

Property G:

$$Var(cU) = c^2 Var(U) \tag{3.49}$$

for any random variable U and constant c. It should make sense to you: If we multiply a random variable by 5, say, then its average squared distance to its mean should increase by a factor of 25.

Let's prove (3.49). Define V = cU. Then

$$\begin{aligned}
Var(V) &= E[(V - EV)^2] \text{ (def.)} && (3.50)\\
&= E\{[cU - E(cU)]^2\} \text{ (subst.)} && (3.51)\\
&= E\{[cU - cEU]^2\} \text{ ((3.21))} && (3.52)\\
&= E\{c^2[U - EU]^2\} \text{ (algebra)} && (3.53)\\
&= c^2 E\{[U - EU]^2\} \text{ ((3.21))} && (3.54)\\
&= c^2 Var(U) \text{ (def.)} && (3.55)
\end{aligned}$$

Shifting data over by a constant does not change the amount of variation in them:

Property H:

$$Var(U + d) = Var(U) \tag{3.56}$$

for any constant d.

Intuitively, the variance of a constant is 0—after all, it never varies! You can show this formally using (3.41):

$$Var(c) = E(c^2) - [E(c)]^2 = c^2 - c^2 = 0 \tag{3.57}$$

The square root of the variance is called the **standard deviation**.

Again, we use variance as our main measure of dispersion for historical and mathematical reasons, not because it's the most meaningful measure. The squaring in the definition of variance produces some distortion, by exaggerating the importance of the larger differences. It would be more natural to use the **mean absolute deviation** (MAD), $E(|U - EU|)$. However, this is less tractable mathematically, so the statistical pioneers chose to use the mean squared difference, which lends itself to lots of powerful and beautiful math, in which the Pythagorean Theorem pops up in abstract vector spaces. (See Section 24.10.2 for details.)

As with expected values, the properties of variance discussed above, and also in Section 16.1.1 below, are key to the entire remainder of this book. You should notice immediately when you are in a setting in which they are applicable. For instance, if you see the variance of the sum of two random variables, you should instinctively think of (3.75) right away, and check whether they are independent.

3.6.2 More Practice with the Properties of Variance

Suppose X and Y are independent random variables, with $EX = 1$, $EY = 2$, $Var(X) = 3$ and $Var(Y) = 4$. Let's find $Var(XY)$. (The reader should make sure to supply the reasons for each step, citing equation numbers from the material above.)

$$\begin{aligned}
Var(XY) &= E(X^2Y^2) - [E(XY)]^2 && (3.58)\\
&= E(X^2) \cdot E(Y^2) - (EX \cdot EY)^2 && (3.59)\\
&= [Var(X) + (EX)^2] \cdot [Var(Y) + (EY)^2] - (EX \cdot EY)^2 && (3.60)\\
&= (3 + 1^2)(4 + 2^2) - (1 \cdot 2)^2 && (3.61)\\
&= 28 && (3.62)
\end{aligned}$$

3.6.3 Central Importance of the Concept of Variance

No one needs to be convinced that the mean is a fundamental descriptor of the nature of a random variable. But the variance is of central importance too, and will be used constantly throughout the remainder of this book.

The next section gives a quantitative look at our notion of variance as a measure of dispersion.

3.6.4 Intuition Regarding the Size of Var(X)

A billion here, a billion there, pretty soon, you're talking real money—attributed to the late Senator Everitt Dirksen, replying to a statement that some federal budget item cost "only" a billion dollars

Recall that the variance of a random variable X is supposed to be a measure of the dispersion of X, meaning the amount that X varies from one instance (one line in our notebook) to the next. But if Var(X) is, say, 2.5, is that a lot of variability or not? We will pursue this question here.

3.6.4.1 Chebychev's Inequality

This inequality states that for a random variable X with mean μ and variance σ^2,

$$P(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2} \tag{3.63}$$

In other words, X strays more than, say, 3 standard deviations from its mean at most only 1/9 of the time. This gives some concrete meaning to the concept of variance/standard deviation.

You've probably had exams in which the instructor says something like "An A grade is 1.5 standard deviations above the mean." Here c in (3.63) would be 1.5.

We'll prove the inequality in Section 3.14.

3.6.4.2 The Coefficient of Variation

Continuing our discussion of the magnitude of a variance, look at our remark following (3.63):

> In other words, X does not often stray more than, say, 3 standard deviations from its mean. This gives some concrete meaning to the concept of variance/standard deviation.

Or, think of the price of, say, widgets. If the price hovers around a $1 million, but the variation around that figure is only about a dollar, you'd say there is essentially no variation. But a variation of about a dollar in the price of a hamburger would be a lot.

These considerations suggest that any discussion of the size of Var(X) should relate to the size of E(X). Accordingly, one often looks at the **coefficient of variation**, defined to be the ratio of the standard deviation to the mean:

$$\text{coef. of var.} = \frac{\sqrt{Var(X)}}{EX} \tag{3.64}$$

This is a scale-free measure (e.g. inches divided by inches), and serves as a good way to judge whether a variance is large or not.

3.7 A Useful Fact

For a random variable X, consider the function

$$g(c) = E[(X - c)^2] \tag{3.65}$$

Remember, the quantity $E[(X - c)^2]$ is a number, so g(c) really is a function, mapping a real number c to some real output.

We can ask the question, What value of c minimizes g(c)? To answer that question, write:

$$g(c) = E[(X - c)^2] = E(X^2 - 2cX + c^2) = E(X^2) - 2cEX + c^2 \tag{3.66}$$

where we have used the various properties of expected value derived in recent sections.

To make this concrete, suppose we are guessing people's weights—without seeing them and without knowing anything about them at all. (This is a somewhat artificial question, but it will become highly practical in Chapter **??**.) Since we know nothing at all about these people, we will make the same guess for each of them.

What should that guess-in-common be? Your first inclination would be to guess everyone to be the mean weight of the population. If that value in our target population is, say, 142.8 pounds, then we'll guess everyone to be that weight. Actually, that guess turns out to be optimal in a certain sense, as follows.

Say X is a person's weight. It's a random variable, because these people are showing up at random from the population. Then $X - c$ is our prediction error. How well will do in our predictions? We can't measure that as

$$E(\text{error}) \tag{3.67}$$

because that quantity is 0! (What mailing tube is at work here?)

A reasonable measure would be

$$E(|X - c|) \tag{3.68}$$

However, due to tradition, we use

$$E[(X - c)^2] \tag{3.69}$$

Now differentiate with respect to c, and set the result to 0. Remembering that $E(X^2)$ and EX are constants, we have

$$0 = -2EX + 2c \tag{3.70}$$

so the minimizing c is c = EX!

In other words, the minimum value of $E[(X - c)^2]$ occurs at c = EX. Our intuition was right!

Moreover: Plugging c = EX into (3.66) shows that the minimum value of g(c) is $E(X - EX)^2]$, which is Var(X)!

In notebook terms, think of guessing many, many people, meaning many lines in the notebook, one per person. Then (3.69) is the long-run average squared error in our guesses, and we find that we minimize that by guessing everyone's weight to be the population mean weight.

But why look at average squared error? It accentuates the large errors. Instead, we could minimize (3.68). It turns out that the best c here is the population *median* weight.

3.8 Covariance

This is a topic we'll cover fully in Chapter 16, but at least introduce here.

A measure of the degree to which U and V vary together is their **covariance**,

$$Cov(U,V) = E[(U - EU)(V - EV)] \tag{3.71}$$

Except for a divisor, this is essentially **correlation**. If U is usually large (relative to its expectation) at the same time V is small (relative to its expectation), for instance, then you can see that the covariance between them will be negative. On the other hand, if they are usually large together or small together, the covariance will be positive.

For example, suppose U and V are the height and weight, respectively, of a person chosen at random from some population, and think in notebook terms. Each line shows the data for one person, and we'll have columns for U, V, U - EU, V - EV and (U - EU) (V - EV). Then (3.71) is the long-run average of that last column. Will it be positive or negative? Reason as follows:

Think of the lines in the notebook for people who are taller than average, i.e. for whom $U - EU > 0$. Most such people are also heavier than average, i.e. $V - EV > 0$, so that $(U - EU)(V - EV) > 0$. On the other hand, shorter people also tend to be lighter, so most lines with shorter people will have $U - EU < 0$ and $V - EV < 0$—but still $(U - EU)(V - EV) > 0$. In other words, the long-run average of the (U - EU) (V - EV) column will be positive.

The point is that, if two variables are positively related, e.g. height and weight, their covariance should be positive. This is the intuitive underlying defining covariance as in (3.71).

Again, one can use the properties of E() to show that

$$Cov(U,V) = E(UV) - EU \cdot EV \tag{3.72}$$

Again, this will be derived fully in Chapter **??**, but you think about how to derive it yourself. Just use our old mailing tubes, e.g. E(X+Y) = EX + EY, E(cX) for a constant c, etc. Note that EU and EV are constants!

Also

$$Var(U+V) = Var(U) + Var(V) + 2Cov(U,V) \tag{3.73}$$

and more generally,

$$Var(aU+bV) = a^2 Var(U) + b^2 Var(V) + 2abCov(U,V) \tag{3.74}$$

for any constants a and b.

(3.72) imply that Cov(U,V) = 0. In that case,

$$Var(U+V) = Var(U) + Var(V) \tag{3.75}$$

By the way, (3.75) is actually the Pythagorean Theorem in a certain esoteric, infinite-dimesional vector space (related to a similar remark made earlier). This is pursued in Section 24.10.2 for the mathematically inclined.

Generalizing (3.74), for constants $a_1, ..., a_k$ and random variables $X_1, ..., X_k$, form the new random variable $a_1X_1 + ... + a_kX_k$. Then

$$Var(a_1X_1 + ... + a_kX_k) = \sum_{i=1}^{k} a_i^2 Var(X_i) + 2 \sum_{1 \leq i < j \leq k} a_i a_j Cov(X_i, X_j) \tag{3.76}$$

If the X_i are independent, then we have the special case

$$Var(a_1X_1 + ... + a_kX_k) = \sum_{i=1}^{k} a_i^2 Var(X_i) \tag{3.77}$$

3.9 Indicator Random Variables, and Their Means and Variances

Definition 5 *A random variable that has the value 1 or 0, according to whether a specified event occurs or not is called an* **indicator random variable** *for that event.*

You'll often see later in this book that the notion of an indicator random variable is a very handy device in certain derivations. But for now, let's establish its properties in terms of mean and variance.

Handy facts: Suppose X is an indicator random variable for the event A. Let p denote P(A). Then

$$E(X) = p \tag{3.78}$$

$$Var(X) = p(1-p) \tag{3.79}$$

These two facts are easily derived. In the first case we have, using our properties for expected value,

$$EX = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = P(X = 1) = P(A) = p \tag{3.80}$$

The derivation for Var(X) is similar (use (3.41)).

For example, say Coin A has probability 0.6 of heads, Coin B is fair, and Coin C has probability 0.2 of heads. I toss A once, getting X heads, then toss B once, getting Y heads, then toss C once, getting Z heads. Let W = X + Y + Z, i.e. the total number of heads from the three tosses (W ranges from 0 to 3). Let's find P(W = 1) and Var(W).

The first one uses old methods:

$$\begin{aligned} P(W = 1) &= P(X = 1 \text{ and } Y = 0 \text{ and } Z = 0 \text{ or } ...) && (3.81) \\ &= 0.6 \cdot 0.5 \cdot 0.8 + 0.4 \cdot 0.5 \cdot 0.8 + 0.4 \cdot 0.5 \cdot 0.2 && (3.82) \end{aligned}$$

For Var(W), let's use what we just learned about indicator random variables; each of X, Y and Z are such variables. Var(W) = Var(X) + Var(Y) + Var(Z), by independence and (3.75). Since X is an indicator random variable, $Var(X) = 0.6 \cdot 0.4$, etc. The answer is then

$$0.6 \cdot 0.4 + 0.5 \cdot 0.5 + 0.2 \cdot 0.8 \tag{3.83}$$

3.9.1 Example: Return Time for Library Books

Suppose at some public library, patrons return books exactly 7 days after borrowing them, never early or late. However, they are allowed to return their books to another branch, rather than the branch where they borrowed their books. In that situation, it takes 9 days for a book to return to its proper library, as opposed to the normal 7. Suppose 50% of patrons return their books to a "foreign" library. Find Var(T), where T is the time, either 7 or 9 days, for a book to come back to its proper location.

Note that

$$T = 7 + 2I, \tag{3.84}$$

where I is an indicator random variable for the event that the book is returned to a "foreign"

branch. Then

$$Var(T) = Var(7 + 2I) = 4Var(I) = 4 \cdot 0.5(1 - 0.5) \tag{3.85}$$

Now let's look at a somewhat more general model. Here we will assume that borrowers return books after 4, 5, 6 or 7 days, with probabilities 0.1, 0.2, 0.3, 0.4, respectively. As before, 50% of patrons return their books to a "foreign" branch, resulting in an extra 2-day delay before the book arrives back to its proper location. The library is open 7 days a week.

Suppose you wish to borrow a certain book, and inquire at the library near the close of business on Monday. Assume too that no one else is waiting for the book. You are told that it had been checked out the previous Thursday. Find the probability that you will need to wait until Wednesday evening to get the book. (You check every evening.)

Let B denote the time needed for the book to arrive back at its home branch, and define I as before. Then

$$\begin{aligned}
P(B = 6 \mid B > 4) &= \frac{P(B = 6 \text{ and } B > 4)}{P(B > 4)} && (3.86)\\
&= \frac{P(B = 6)}{P(B > 4)} && (3.87)\\
&= \frac{P(B = 6 \text{ and I} = 0 \text{ or B} = 6 \text{ and I} = 1)}{1 - P(B = 4)} && (3.88)\\
&= \frac{0.5 \cdot 0.3 + 0.5 \cdot 0.1}{1 - 0.5 \cdot 0.1} && (3.89)\\
&= \frac{4}{19} && (3.90)
\end{aligned}$$

Here is a simulation check:

```
libsim <- function(nreps) {
   # patron return time
   prt <- sample(c(4,5,6,7),nreps,replace=T,prob=c(0.1,0.2,0.3,0.4))
   # indicator for foreign branch
   i <- sample(c(0,1),nreps,replace=T)
   b <- prt + 2*i
   x <- cbind(prt,i,b)
   # look only at the relevant notebook lines
   bgt4 <- x[b > 4,]
   # among those lines, what proportion have B = 6?
```

```
    mean(bgt4[,3] == 6)
}
```

3.9.2 Example: Indicator Variables in a Committee Problem

A committee of four people is drawn at random from a set of six men and three women. Suppose we are concerned that there may be quite a gender imbalance in the membership of the committee. Toward that end, let M and W denote the numbers of men and women in our committee, and let D = M-W. Let's find E(D), in two different ways.

D has support consisting of the values 4-0, 3-1, 2-2 and 1-3, i.e. 4, 2, 0 and -2. So from (3.13)

$$ED = -2 \cdot P(D = -2) + 0 \cdot P(D = 0) + 2 \cdot P(D = 2) + 4 \cdot P(D = 4) \tag{3.91}$$

Now, using reasoning along the lines in Section 2.15, we have

$$P(D = -2) = P(M = 1 \text{ and } W = 3) = \frac{\binom{6}{1}\binom{3}{3}}{\binom{9}{4}} \tag{3.92}$$

After similar calculations for the other probabilities in (3.91), we find the $ED = \frac{4}{3}$.

Note what this means: If we were to perform this experiment many times, i.e. choose committees again and again, on average we would have a little more than one more man than women on the committee.

Now let's use our "mailing tubes" to derive ED a different way:

$$\begin{aligned} ED &= E(M - W) && (3.93)\\ &= E[M - (4 - M)] && (3.94)\\ &= E(2M - 4) && (3.95)\\ &= 2EM - 4 \text{ (from (3.21))} && (3.96) \end{aligned}$$

Now, let's find EM by using indicator random variables. Let G_i denote the indicator random variable for the event that the i^{th} person we pick is male, i = 1,2,3,4. Then

$$M = G_1 + G_2 + G_3 + G_4 \tag{3.97}$$

so

$$\begin{aligned} EM &= E(G_1 + G_2 + G_3 + G_4) && (3.98) \\ &= EG_1 + EG_2 + EG_3 + EG_4 \ \ [\text{ from } (3.19)] && (3.99) \\ &= P(G_1 = 1) + P(G_2 = 1) + P(G_3 = 1) + P(G_4 = 1) \ \ [\text{ from } (3.78)] && (3.100) \end{aligned}$$

Note carefully that the second equality here, which uses (3.19), is true in spite of the fact that the G_i are not independent. Equation (3.19) does not require independence.

Another key point is that, due to symmetry, $P(G_i = 1)$ is the same for all i. Note that we did not write a *conditional* probability here! Once again, think of the notebook view: **By definition**, ($P(G_2 = 1)$ is the long-run proportion of the number of notebook lines in which $G_2 = 1$—regardless of the value of G_1 in those lines.

Now, to see that $P(G_i = 1)$ is the same for all i, suppose the six men that are available for the committee are named Alex, Bo, Carlo, David, Eduardo and Frank. When we select our first person, any of these men has the same chance of being chosen (1/9). *But that is also true for the second pick.* Think of a notebook, with a column named "second pick." In some lines, that column will say Alex, in some it will say Bo, and so on, and in some lines there will be women's names. But in that column, Bo will appear the same fraction of the time as Alex, due to symmetry, and that will be the same fraction as for, say, Alice, again 1/9.

Now,

$$P(G_1 = 1) = \frac{6}{9} = \frac{2}{3} \tag{3.101}$$

Thus

$$ED = 2 \cdot (4 \cdot \frac{2}{3}) - 4 = \frac{4}{3} \tag{3.102}$$

3.9.3 Example: Spinner Game

In a certain game, Person A spins a spinner and wins S dollars, with mean 10 and variance 5. Person B flips a coin. If it comes up heads, Person A must give B whatever A won, but if it comes up tails, B wins nothing. Let T denote the amount B wins. Let's find $Var(T)$.

We can use (3.60), in this case with X = I, where I is an indicator variable for the event that B

gets a head, and with Y = S. Then $T = I \cdot S$, and I and S are independent, so

$$Var(T) = Var(IS) = [Var(I) + (EI)^2] \cdot [Var(S) + (ES)^2] - (EI \cdot ES)^2 \tag{3.103}$$

Then use the facts that I has mean 0.5 and variance 0.5(1-0.5) (Equations (3.78) and (3.79), with S having the mean 10 and variance 5, as given in the problem.

3.10 Expected Value, Etc. in the ALOHA Example

Finding expected values etc. in the ALOHA example is straightforward. For instance,

$$EX_1 = 0 \cdot P(X_1 = 0) + 1 \cdot P(X_1 = 1) + 2 \cdot P(X_1 = 2) = 1 \cdot 0.48 + 2 \cdot 0.52 = 1.52 \tag{3.104}$$

Here is R code to find various values approximately by simulation:

```
# finds E(X1), E(X2), Var(X2), Cov(X1,X2)
sim <- function(p,q,nreps) {
   sumx1 <- 0
   sumx2 <- 0
   sumx2sq <- 0
   sumx1x2 <- 0
   for (i in 1:nreps) {
      numtrysend <-
         sum(sample(0:1,2,replace=TRUE,prob=c(1-p,p)))
      if (numtrysend == 1)  X1 <- 1
      else X1 <- 2
      numactive <- X1
      if (X1 == 1 && runif(1) < q) numactive <- numactive + 1
      if (numactive == 1)
         if (runif(1) < p) X2 <- 0
         else X2 <- 1
      else {  # numactive = 2
         numtrysend <- 0
         for (i in 1:2)
            if (runif(1) < p) numtrysend <- numtrysend + 1
         if (numtrysend == 1) X2 <- 1
         else X2 <- 2
      }
      sumx1 <- sumx1 + X1
      sumx2 <- sumx2 + X2
      sumx2sq <- sumx2sq + X2^2
      sumx1x2 <- sumx1x2 + X1*X2
   }
   # print results
   meanx1 <- sumx1 /nreps
   cat("E(X1):",meanx1,"\n")
   meanx2 <- sumx2 /nreps
   cat("E(X2):",meanx2,"\n")
```

```
34      cat("Var(X2):",sumx2sq/nreps - meanx2^2,"\n")
35      cat("Cov(X1,X2):",sumx1x2/nreps - meanx1*meanx2,"\n")
36  }
```

As a check on your understanding so far, you should find at least one of these values by hand, and see if it jibes with the simulation output.

3.11 Example: Measurements at Different Ages

Say a large research program measures boys' heights at age 10 and age 15. Call the two heights X and Y. So, each boy has an X and a Y. Each boy is a "notebook line", and the notebook has two columns, for X and Y. We are interested in Var(Y-X). Which of the following is true?

(i) $Var(Y-X) = Var(Y) + Var(X)$

(ii) $Var(Y-X) = Var(Y) - Var(X)$

(iii) $Var(Y-X) < Var(Y) + Var(X)$

(iv) $Var(Y-X) < Var(Y) - Var(X)$

(v) $Var(Y-X) > Var(Y) + Var(X)$

(vi) $Var(Y-X) > Var(Y) - Var(X)$

(vii) None of the above.

Use the mailing tube (3.74):

$$Var(Y-X) = Var[Y + (-X)] = Var(Y) + Var(X) - 2Cov(X,Y) \tag{3.105}$$

Since X and Y are positively correlated, their covariance is positive, so the answer is (iii).

3.12 Example: Bus Ridership Model

In the bus ridership model, Section 2.12, let's find $Var(L_1)$:

$$Var(L_1) = E(L_1^2) - (EL_1)^2 \tag{3.106}$$

$$EL_1 = EB_1 = 0 \cdot 0.5 + 1 \cdot 0.4 + 2 \cdot 0.1 \tag{3.107}$$

$$E(L_1^2) = 0^2 \cdot 0.5 + 1^2 \cdot 0.4 + 2^2 \cdot 0.1 \tag{3.108}$$

Then put it all together.

3.13 Distributions

The idea of the **distribution** of a random variable is central to probability and statistics.

Definition 6 *Let U be a discrete random variable. Then the distribution of U is simply the support of U, together with the associated probabilities.*

Example: Let X denote the number of dots one gets in rolling a die. Then the values X can take on are 1,2,3,4,5,6, each with probability 1/6. So

$$\text{distribution of } X = \{(1, \frac{1}{6}), (2, \frac{1}{6}), (3, \frac{1}{6}), (4, \frac{1}{6}), (5, \frac{1}{6}), (6, \frac{1}{6})\} \tag{3.109}$$

Example: Recall the ALOHA example. There X_1 took on the values 1 and 2, with probabilities 0.48 and 0.52, respectively (the case of 0 was impossible). So,

$$\text{distribution of } X_1 = \{(0, 0.00), (1, 0.48), (2, 0.52)\} \tag{3.110}$$

Example: Recall our example in which N is the number of tosses of a coin needed to get the first head. N has support 1,2,3,..., the probabilities of which we found earlier to be 1/2, 1/4, 1/8,... So,

$$\text{distribution of } N = \{(1, \frac{1}{2}), (2, \frac{1}{4}), (3, \frac{1}{8}), ...\} \tag{3.111}$$

It is common to express this in functional notation:

Definition 7 *The* **probability mass function** *(pmf) of a discrete random variable V, denoted p_V, as*

$$p_V(k) = P(V = k) \tag{3.112}$$

for any value k in the support of V.

(Please keep in mind the notation. It is customary to use the lower-case p, with a subscript consisting of the name of the random variable.)

Note that $p_V()$ is just a function, like any function (with integer domain) you've had in your previous math courses. For each input value, there is an output value.

3.13.1 Example: Toss Coin Until First Head

In (3.111),

$$p_N(k) = \frac{1}{2^k}, k = 1, 2, ... \tag{3.113}$$

3.13.2 Example: Sum of Two Dice

In the dice example, in which S = X+Y,

$$p_S(k) = \begin{cases} \frac{1}{36}, & k = 2 \\ \frac{2}{36}, & k = 3 \\ \frac{3}{36}, & k = 4 \\ ... & \\ \frac{1}{36}, & k = 12 \end{cases} \tag{3.114}$$

It is important to note that there may not be some nice closed-form expression for p_V like that of (3.113). There was no such form in (3.114), nor is there in our ALOHA example for p_{X_1} and p_{X_2}.

3.13.3 Example: Watts-Strogatz Random Graph Model

Random graph models are used to analyze many types of link systems, such as power grids, social networks and even movie stars. We saw our first example in Section 2.13.1, and here is another, a variation on a famous model of that type, due to Duncan Watts and Steven Strogatz.

3.13.3.1 The Model

We have a graph of n nodes, e.g. in which each node is a person).[4] Think of them as being linked in a circle—we're just talking about relations here, not physical locations—so we already have n links. One can thus reach any node in the graph from any other, by following the links of the circle. (We'll assume all links are bidirectional.)

We now randomly add k more links (k is thus a parameter of the model), which will serve as "shortcuts." There are $\binom{n}{2} = n(n-1)/2$ possible links between nodes, but remember, we already have n of those in the graph, so there are only $n(n-1)/2 - n = n^2/2 - 3n/2$ possibilities left. We'll be forming k new links, chosen at random from those $n^2/2 - 3n/2$ possibilities.

Let M denote the number of links attached to a particular node, known as the **degree** of a node. M is a random variable (we are choosing the shortcut links randomly), so we can talk of its pmf, p_M, termed the **degree distribution** of M, which we'll calculate now.

Well, $p_M(r)$ is the probability that this node has r links. Since the node already had 2 links before the shortcuts were constructed, $p_M(r)$ is the probability that r-2 of the k shortcuts attach to this node.

This problem is similar in spirit to (though admittedly more difficult to think about than) kings-and-hearts example of Section 2.15.1. Other than the two neighboring links in the original circle and the "link" of a node to itself, there are n-3 possible shortcut links to attach to our given node. We're interested in the probability that r-2 of them are chosen, and that k-(r-2) are chosen from the other possible links. Thus our probability is:

$$p_M(r) = \frac{\binom{n-3}{r-2}\binom{n^2/2-3n/2-(n-3)}{k-(r-2)}}{\binom{n^2/2-3n/2}{k}} = \frac{\binom{n-3}{r-2}\binom{n^2/2-5n/2+3}{k-(r-2)}}{\binom{n^2/2-3n/2}{k}} \tag{3.115}$$

3.13.3.2 Further Reading

UCD professor Raissa D'Souza specializes in random graph models. See for instance Beyond Friendship: Modeling User activity Graphs on Social Network-Based Gifting Applications, A. Nazir, A. Waagen, V. Vijayaraghavan, C.-N. Chuah, R. M. D'Souza, B. Krishnamurthy, *ACM Internet Measurement Conference (IMC 2012)*, Nov 2012.

[4] The word *graph* here doesn't mean "graph" in the sense of a picture. Here we are using the computer science sense of the word, meaning a system of vertices and edges. It's common to call those *nodes* and *links*.

notebook line	Y	dZ	$Y \geq dZ$?
1	0.36	0	yes
2	3.6	3	yes
3	2.6	0	yes

Table 3.2: Illustration of Y and Z

3.14 Proof of Chebychev's Inequality (optional section)

To prove (3.63), let's first state and prove Markov's Inequality: For any nonnegative random variable Y and positive constant d,

$$P(Y \geq d) \leq \frac{EY}{d} \tag{3.116}$$

To prove (3.116), let Z be the indicator random variable for the event $Y \geq d$ (Section 3.9).

Now note that

$$Y \geq dZ \tag{3.117}$$

To see this, just think of a notebook, say with d = 3. Then the notebook might look like Table 3.2.

So

$$EY \geq dEZ \tag{3.118}$$

(Again think of the notebook. The long-run average in the Y column will be $\geq$ the corresponding average for the dZ column.)

The right-hand side of (3.118) is $dP(Y \geq d)$, so (3.116) follows.

Now to prove (3.63), define

$$Y = (X - \mu)^2 \tag{3.119}$$

and set $d = c^2\sigma^2$. Then (3.116) says

$$P[(X-\mu)^2 \geq c^2\sigma^2] \leq \frac{E[(X-\mu)^2]}{c^2\sigma^2} \tag{3.120}$$

Since

$$(X-\mu)^2 \geq c^2\sigma^2 \text{ if and only if } |X-\mu| \geq c\sigma \tag{3.121}$$

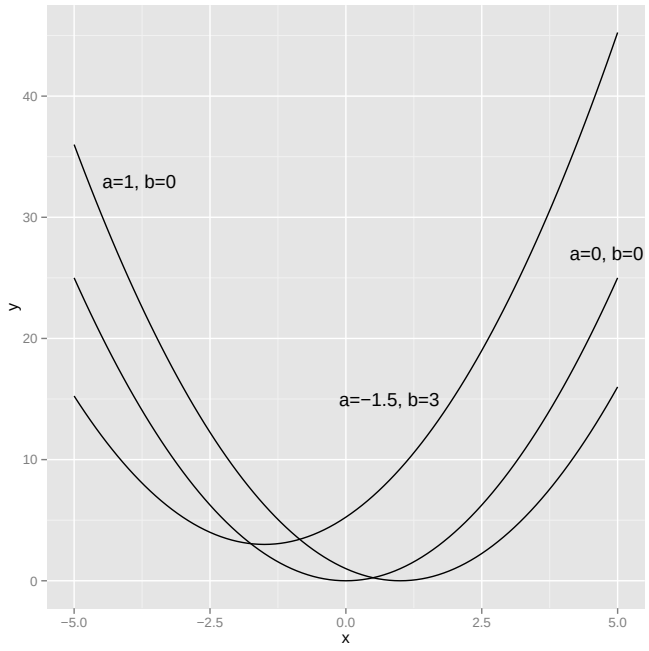
the left-hand side of (3.120) is the same as the left-hand side of (3.63). The numerator of the right-hand size of (3.120) is simply Var(X), i.e. σ^2, so we are done.

Chapter 4

Discrete Parametric Distribution Families

The notion of a *parametric family* of distributions is a key concept that will recur throughout the book.

Consider plotting the curves $g_{a,b}(t) = (t-a)^2 + b$. For each a and b, we get a different parabola, as seen in this plot of three of the curves:



This is a family of curves, thus a family of functions. We say the numbers a and b are the **parameters** of the family. Note carefully that t is not a parameter, but rather just an argument of each function. The point is that a and b are indexing the curves.

4.1 The Case of Importance to Us: Parameteric Families of pmfs

Probability mass functions are still functions.[1] Thus they too can come in parametric families, indexed by one or more parameters. We had an example in Section 3.13.3. Since we get a different function p_M for each different values of k and n, that was a parametric family of pmfs, indexed by k and n.

Some parametric families of pmfs have been found to be so useful over the years that they've been given names. We will discuss some of those families here. But remember, they are famous just because they have been found useful, i.e. that they fit real data well in various settings. **Do not jump to the conclusion that we always "must" use pmfs from some family.**

4.2 The Geometric Family of Distributions

To explain our first parametric family of pmfs, recall our example of tossing a coin until we get the first head, with N denoting the number of tosses needed. In order for this to take k tosses, we need k-1 tails and then a head. Thus

$$p_N(k) = (1 - \frac{1}{2})^{k-1} \cdot \frac{1}{2}, k = 1, 2, ... \tag{4.1}$$

We might call getting a head a "success," and refer to a tail as a "failure." Of course, these words don't mean anything; we simply refer to the outcome of interest (which of course we ourselves choose) as "success."

Define M to be the number of rolls of a die needed until the number 5 shows up. Then

$$p_M(k) = \left(1 - \frac{1}{6}\right)^{k-1} \frac{1}{6}, k = 1, 2, ... \tag{4.2}$$

reflecting the fact that the event {M = k} occurs if we get k-1 non-5s and then a 5. Here "success" is getting a 5.

[1] The domains of these functions are typically the integers, but that is irrelevant; a function is a function.

The tosses of the coin and the rolls of the die are known as **Bernoulli trials**, which is a sequence of independent events. We call the occurrence of the event **success** and the nonoccurrence **failure** (just convenient terms, not value judgments). The associated indicator random variable are denoted B_i, i = 1,2,3,... So B_i is 1 for success on the ith trial, 0 for failure, with success probability p. For instance, p is 1/2 in the coin case, and 1/6 in the die example.

In general, suppose the random variable W is defined to be the number of trials needed to get a success in a sequence of Bernoulli trials. Then

$$p_W(k) = (1-p)^{k-1}p, k = 1, 2, ... \tag{4.3}$$

Note that there is a different distribution for each value of p, so we call this a **parametric family** of distributions, indexed by the parameter p. We say that W is **geometrically distributed** with parameter p.[2]

It should make good intuitive sense to you that

$$E(W) = \frac{1}{p} \tag{4.4}$$

This is indeed true, which we will now derive. First we'll need some facts (which you should file mentally for future use as well):

Properties of Geometric Series:

(a) For any $t \neq 1$ and any nonnegative integers $r \leq s$,

$$\sum_{i=r}^{s} t^i = t^r \frac{1-t^{s-r+1}}{1-t} \tag{4.5}$$

This is easy to derive for the case r = 0, using mathematical induction. For the general case, just factor out t^r.

(b) For $|t| < 1$,

$$\sum_{i=0}^{\infty} t^i = \frac{1}{1-t} \tag{4.6}$$

To prove this, just take r = 0 and let $s \to \infty$ in (4.5).

[2]Unfortunately, we have overloaded the letter p here, using it to denote the probability mass function on the left side, and the unrelated parameter p, our success probability on the right side. It's not a problem as long as you are aware of it, though.

(c) For $|t| < 1$,

$$\sum_{i=1}^{\infty} it^{i-1} = \frac{1}{(1-t)^2} \tag{4.7}$$

This is derived by applying $\frac{d}{dt}$ to (4.6).[3]

Deriving (4.4) is then easy, using (4.7):

$$\begin{aligned} EW &= \sum_{i=1}^{\infty} i(1-p)^{i-1}p && (4.8)\\ &= p\sum_{i=1}^{\infty} i(1-p)^{i-1} && (4.9)\\ &= p \cdot \frac{1}{[1-(1-p)]^2} && (4.10)\\ &= \frac{1}{p} && (4.11) \end{aligned}$$

Using similar computations, one can show that

$$Var(W) = \frac{1-p}{p^2} \tag{4.12}$$

We can also find a closed-form expression for the quantities $P(W \leq m)$, m = 1,2,... (This has a formal name $F_W(m)$, as will be seen later in Section 7.3.) For any positive integer m we have

$$\begin{aligned} F_W(m) &= P(W \leq m) && (4.13)\\ &= 1 - P(W > m) && (4.14)\\ &= 1 - P(\text{the first m trials are all failures}) && (4.15)\\ &= 1 - (1-p)^m && (4.16) \end{aligned}$$

By the way, if we were to think of an experiment involving a geometric distribution in terms of our notebook idea, the notebook would have an infinite number of columns, one for each B_i. Within each row of the notebook, the B_i entries would be 0 until the first 1, then NA ("not applicable") after that.

[3]To be more careful, we should differentiate (4.5) and take limits.

4.2.1 R Functions

You can simulate geometrically distributed random variables via R's **rgeom()** function. Its first argument specifies the number of such random variables you wish to generate, and the second is the success probability p.

For example, if you run

```
> y <- rgeom(2,0.5)
```

then it's simulating tossing a coin until you get a head (**y[1]**) and then tossing the coin until a head again (**y[2]**). Of course, you could simulate on your own, say using **sample()** and **while()**, but R makes it convenient for you.

Here's the full set of functions for a geometrically distributed random variable X with success probability p:

- **dgeom(i,p)**, to find $P(X = i)$
- **pgeom(i,p)**, to find $P(X \leq i)$
- **qgeom(q,p)**, to find c such that $P(X \leq c) = q$
- **rgeom(n,p)**, to generate n variates from this geometric distribution

Important note: Some books define geometric distributions slightly differently, as the number of failures before the first success, rather than the number of trials to the first success. The same is true for software—both R and Python define it this way. Thus for example in calling **dgeom()**, subtract 1 from the value used in our definition.

For example, here is P(N = 3) for a geometric distribution under our defintion, with p = 0.4:

```
> dgeom(2,0.4)
[1] 0.144
> # check
> (1-0.4)^(3-1) * 0.4
[1] 0.144
```

Note that this also means one must *add* 1 to the result of **rgeom()**.

4.2.2 Example: a Parking Space Problem

Suppose there are 10 parking spaces per block on a certain street. You turn onto the street at the start of one block, and your destination is at the start of the next block. You take the first parking space you encounter. Let D denote the distance of the parking place you find from your destination, measured in parking spaces. Suppose each space is open with probability 0.15, with the spaces being independent. Find ED.

To solve this problem, you might at first think that D follows a geometric distribution. **But don't jump to conclusions!** Actually this is not the case; D is a somewhat complicated distance. But clearly D is a function of N, where the latter denotes the number of parking spaces you see until you find an empty one—and N *is* geometrically distributed.

As noted, D is a function of N:

$$D = \begin{cases} 11 - N, & N \leq 10 \\ N - 11, & N > 10 \end{cases} \tag{4.17}$$

Since D is a function of N, we can use (3.36) with g(t) as in (4.17):

$$ED = \sum_{i=1}^{10} (11 - i)(1 - 0.15)^{i-1} 0.15 + \sum_{i=11}^{\infty} (i - 11) 0.85^{i-1} 0.15 \tag{4.18}$$

This can now be evaluated using the properties of geometric series presented above.

Alternatively, here's how we could find the result by simulation:

```
1 parksim <- function(nreps) {
2    # do the experiment nreps times, recording the values of N
3    nvals <- rgeom(nreps,0.15) + 1
4    # now find the values of D
5    dvals <- ifelse(nvals <= 10,11-nvals,nvals-11)
6    # return ED
7    mean(dvals)
8 }
```

Note the vectorized addition and recycling (Section 2.14.2) in the line

```
nvals <- rgeom(nreps,0.15) + 1
```

The call to **ifelse()** is another instance of R's vectorization, a vectorized if-then-else. The first argument evaluates to a vector of TRUE and FALSE values. For each TRUE, the corresponding

element of **dvals** will be set to the corresponding element of the vector **11-nvals** (again involving vectorized addition and recycling), and for each false, the element of **dvals** will be set to the element of **nvals-11**.

Let's find some more, first $p_N(3)$:

$$p_N(3) = P(N = 3) = (1 - 0.15)^{3-1}0.15 \tag{4.19}$$

Next, find P(D = 1):

$$\begin{aligned} P(D = 1) &= P(N = 10 \text{ or } N = 12) &(4.20)\\ &= (1 - 0.15)^{10-1}0.15 + (1 - 0.15)^{12-1}0.15 &(4.21)\end{aligned}$$

Say Joe is the one looking for the parking place. Paul is watching from a side street at the end of the first block (the one before the destination), and Martha is watching from an alley situated right after the sixth parking space in the second block. Martha calls Paul and reports that Joe never went past the alley, and Paul replies that he did see Joe go past the first block. They are interested in the probability that Joe parked in the second space in the second block. In mathematical terms, what probability is that? Make sure you understand that it is $P(N = 12 \mid N > 10 \text{ and } N \leq 16)$. It can be evaluated as above.

Also: Good news! I found a parking place just one space away from the destination. Find the probability that I am parked in the same block as the destination.

$$\begin{aligned} P(N = 12 \mid N = 10 \text{ or } N = 12) &= \frac{P(N = 12)}{P(N = 10 \text{ or } N = 12)} &(4.22)\\ &= \frac{(1 - 0.15)^{11}\ 0.15}{(1 - 0.15)^{9}\ 0.15 + (1 - 0.15)^{11}\ 0.15} &(4.23)\end{aligned}$$

4.3 The Binomial Family of Distributions

A geometric distribution arises when we have Bernoulli trials with parameter p, with a variable number of trials (N) but a fixed number of successes (1). A **binomial distribution** arises when we have the opposite—a fixed number of Bernoulli trials (n) but a variable number of successes (say X).[4]

[4]Note again the custom of using capital letters for random variables, and lower-case letters for constants.

For example, say we toss a coin five times, and let X be the number of heads we get. We say that X is binomially distributed with parameters n = 5 and p = 1/2. Let's find P(X = 2). There are many orders in which that could occur, such as HHTTT, TTHHT, HTTHT and so on. Each order has probability $0.5^2(1-0.5)^3$, and there are $\binom{5}{2}$ orders. Thus

$$P(X=2) = \binom{5}{2} 0.5^2 (1-0.5)^3 = \binom{5}{2}/32 = 5/16 \tag{4.24}$$

For general n and p,

$$p_X(k) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.25}$$

So again we have a parametric family of distributions, in this case a family having two parameters, n and p.

Let's write X as a sum of those 0-1 Bernoulli variables we used in the discussion of the geometric distribution above:

$$X = \sum_{i=1}^{n} B_i \tag{4.26}$$

where B_i is 1 or 0, depending on whether there is success on the i^{th} trial or not. Note again that the B_i are indicator random variables (Section 3.9), so

$$EB_i = p \tag{4.27}$$

and

$$Var(B_i) = p(1-p) \tag{4.28}$$

Then the reader should use our earlier properties of E() and Var() in Sections 3.5 and 3.6 to fill in the details in the following derivations of the expected value and variance of a binomial random variable:

$$EX = E(B_1 + ..., +B_n) = EB_1 + ... + EB_n = np \tag{4.29}$$