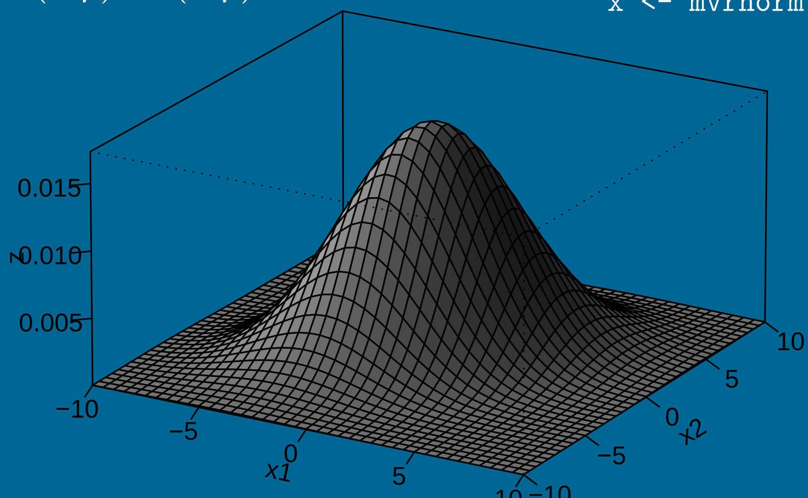
From Algorithms to Z-Scores:

Probabilistic and Statistical Modeling in Computer Science

Norm Matloff, University of California, Davis

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

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



The author has striven to minimize the number of errors, but no guarantee is made as to accuracy of the contents of this book.

Author's Biographical Sketch

Dr. Norm Matloff is a professor of computer science at the University of California at Davis, and was formerly a professor of statistics at that university. He is a former database software developer in Silicon Valley, and has been a statistical consultant for firms such as the Kaiser Permanente Health Plan.

Dr. Matloff was born in Los Angeles, and grew up in East Los Angeles and the San Gabriel Valley. He has a PhD in pure mathematics from UCLA, specializing in probability theory and statistics. He has published numerous papers in computer science and statistics, with current research interests in machine learning, parallel processing, statistical computing, and regression methodology.

Prof. Matloff is a former appointed member of IFIP Working Group 11.3, an international committee concerned with database software security, established under the United Nations. He was a founding member of the UC Davis Department of Statistics, and participated in the formation of the UCD Computer Science Department as well. He is a recipient of the campuswide Distinguished Teaching Award and Distinguished Public Service Award at UC Davis.

Dr. Matloff is currently Editor-in-Chief of the *R Journal*, and is on the editorial board of the *Journal of Statistical Software*.

Dr. Matloff is the author of several published textbooks. His book *Statistical Regression and Classification: from Linear Models to Machine Learning*, won the Ziegel Award in 2017.

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, 5, 6, 7, **??**, **??**, **??**, 9, 10, **??** and 11. 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 **??**, **??**, **??**, **??**, **??**, **??**, **??**, 11, **??**, **??** and 12.

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 A “Notebook” View: the 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

??.

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. In the dice example, 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 $A \cup B$.
[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 ∩ 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.

Let's say the two nodes consist of terminals, at which John and Mary are typing. Call the nodes Node John and Node Mary.

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 John would send and Node B wouldn't, or John would refrain from sending and Node 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 Example: Dice and Conditional Probability

Note in particular that $P(B|A)$ and $P(A|B)$ are completely different quantities. The first restricts attention to lines of the notebook in which A occurs, while for the second we look at lines in which B occurs.

Suppose two dice are rolled, resulting in the random variables X and Y. Let S be the sum X+Y, and let T denote the number of dice having an even number of dots, 0, 1 or 2. Suppose we hear that S = 12, and we know nothing else. Then we know that T must be 2. In other words,

$$P(T = 2 \mid S = 12) = 1 \tag{2.22}$$

On the other hand, if we hear instead that T = 2, that does *not* imply that S is 12, i.e.

$$P(S = 12 \mid T = 2) < 1 \tag{2.23}$$

In other words

$$P(S = 12 \mid T = 2) \neq P(T = 2 \mid S = 12) \tag{2.24}$$

(In fact the reader should show that $P(S = 12 \mid T = 2) = 1/9$.

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

notebook line	$X_1 = 2$	$X_2 = 2$	$X_1 = 2$ and $X_2 = 2$	$X_2 = 2 \vert X_1 = 2$
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

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, 2/7 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 $X_2 = 2|X_1 = 2$ 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:

> 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

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

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.

$$P(X_1 = 2) = P(\underbrace{C_1 \text{ and } C_2}\ \text{ or }\ \underbrace{\text{not } C_1 \text{ and } \text{ not } C_2}) \tag{2.25}$$

  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

$$P(A) = P(B) \tag{2.26}$$

  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: 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. The token advances according to the roll of a single die. If a player lands on square 3, he/she gets a bonus turn.

In most problems like this, **it is greatly helpful to first name the quantities or events involved**. Toward that end, let R denote the player's first roll, and let B be his bonus if there is one, with B being set to 0 if there is no bonus.

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).

$$
\begin{aligned}
P(\text{doesn't reach or pass 0}) &= P(R + B \leq 7) && (2.27)\\
&= P(R \leq 6, R \neq 3 \text{ or } R = 3, B \leq 4) && (2.28)\\
&= P(R \leq 6, R \neq 3) + P(R = 3, B \leq 4) && (2.29)\\
&= P(R \leq 6, R \neq 3) + P(R = 3)\ P(B \leq 4 \mid R = 3) && (2.30)\\
&= \frac{5}{6} + \frac{1}{6} \cdot \frac{4}{6} && (2.31)\\
&= \frac{17}{18} && (2.32)
\end{aligned}
$$

(Above we have written commas as a shorthand notation for *and*, a common abbreviation.)

Now here is a very subtle issue. The events $R + B = 3$ and $B \leq 4$ are *not* independent. At first it would seem that they're independent, as we are rolling the die twice, and successful rolls are independent. That's true, they are independent, but remember, B can be 0 (no bonus), and the event $B \leq 4$ includes that 0 case. And if we know $R = 3$, then we know for sure that B cannot be 0, so we see the two events are not independent.

Note that we first had to "translate" the English specification of the problem ("doesn't reach or...") to math symbols. As mentioned, this is quite helpful.

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

$$
\begin{aligned}
P(\text{don't reach or pass 0}) &= 1 - P(\text{do reach or pass 0}) && (2.33)\\
&= 1 - P(R+B > 7) && (2.34)\\
&= 1 - P(R=3, B>4) && (2.35)\\
&= 1 - \frac{1}{6} \cdot \frac{2}{6} && (2.36)\\
&= \frac{17}{18} && (2.37)
\end{aligned}
$$

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

$$
\begin{aligned}
P(B > 0 \mid R+B=4) &= \frac{P(R+B=4, B>0)}{P(R+B=4)} && (2.38)\\
&= \frac{P(R+B=4, B>0)}{P(R+B=4, B>0 \text{ or } R+B=4, B=0)} && (2.39)\\
&= \frac{P(R+B=4, B>0)}{P(R+B=4, B>0) + P(R+B=4, B=0)} && (2.40)\\
&= \frac{P(R=3, B=1)}{P(R=3, B=1) + P(R=4)} && (2.41)\\
&= \frac{\frac{1}{6} \cdot \frac{1}{6}}{\frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6}} && (2.42)\\
&= \frac{1}{7} && (2.43)
\end{aligned}
$$

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.

2.11 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.

Once again, **it is greatly helpful to first name the quantities or events involved**. 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:

$$P(B_1 = 0 \text{ and } B_2 = 0 \text{ and } B_3 = 0) = 0.5^3 \tag{2.44}$$

Now let's compute the probability that the bus leaves the second stop empty. Again, we must translate this to math first, $P(L_2 = 0)$. Now as usual, "break big events into small events":

$$
\begin{aligned}
P(L_2 = 0) &= P(B_1 = 0 \text{ and } L_2 = 0 \text{ or } B_1 = 1 \text{ and } L_2 = 0 \text{ or } B_1 = 2 \text{ and } L_2 = 0) && (2.45)\\
&= \sum_{i=0}^{2} P(B_1 = i \text{ and } L_2 = 0) && (2.46)\\
&= \sum_{i=0}^{2} P(B_1 = i) P(L_2 = 0 | B_1 = i) && (2.47)\\
&= 0.5^2 + (0.4)(0.2)(0.5) + (0.1)(0.2^2)(0.5) && (2.48)\\
&= 0.292 && (2.49)
\end{aligned}
$$

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

$$
\begin{aligned}
P(B_1 = 2 \mid L_2 = 0) &= \frac{P(B_1 = 2 \text{ and } L_2 = 0)}{P(L_2 = 0)} && (2.50)\\
&= P(B_1 = 2)\ P(L_2 = 0 \mid B_1 = 2)\ /\ 0.292 && (2.51)\\
&= 0.1 \cdot 0.2^2 \cdot 0.5/0.292 && (2.52)
\end{aligned}
$$

(the 0.292 had been previously calculated in (2.49)).

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

$$
\begin{aligned}
P(B_2 < B_1) &= P(B_1 = 1 \text{ and } B_2 < B_1 \text{ or } B_1 = 2 \text{ and } B_2 < B_1) && (2.53)\\
&= 0.4 \cdot 0.5 + 0.1 \cdot (0.5 + 0.4) && (2.54)
\end{aligned}
$$

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

$$
\begin{aligned}
P(L_1 = 1 | L_2 = 0 \text{ and } L_1 > 0) &= \frac{P(L_1 = 1 \text{ and } L_2 = 0)}{P(L_2 = 0 \text{ and } L_1 > 0)} && (2.55)\\
&= \frac{P(B_1 = 1 \text{ and } L_2 = 0)}{P(B_1 = 1 \text{ and } L_2 = 0 \text{ or } B_1 = 2 \text{ and } L_2 = 0)} && (2.56)\\
&= \frac{(0.4)(0.2)(0.5)}{(0.4)(0.2)(0.5) + (0.1)(0.2)^2(0.5)} && (2.57)
\end{aligned}
$$

Equation (2.56) 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

$$L_1 = 1 \text{ and } L_2 = 0 \tag{2.58}$$

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

$$B_1 = 0 \text{ and } L_2 = 0 \tag{2.59}$$

So, how did the denominator in (2.56) come from the preceding equation? In other words, how did we recast the event

$$L_2 = 0 \text{ and } L_1 > 0 \tag{2.60}$$

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 (2.56).

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

Here is another computation: An observer at the second stop notices that no one alights there, but it is dark and the observer couldn't see whether anyone was on the bus. Find the probability that there was one passenger on the bus at the time.

Let A_i denote the number of passengers alighting at stop i.

$$P(L_1 = 1|A_2 = 0) = \frac{P(L_1 = 1 \text{ and } A_2 = 0)}{P(A_2 = 0)} \tag{2.61}$$
$$= \frac{P(L_1 = 1 \text{ and } A_2 = 0)}{\sum_{j=0}^{2} P(L_1 = j \text{ and } A_2 = 0)} \tag{2.62}$$
$$= \frac{0.4 \cdot 0.8}{0.5 \cdot 1 + 0.4 \cdot 0.8 + 0.1 \cdot 0.8^2} \tag{2.63}$$

2.12 Bayes' Rule

(This section should not be confused with Section **??**. 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

$$P(X_1 = 1|X_2 = 2) = \frac{P(X_1 = 1 \text{ and } X_2 = 2)}{P(X_2 = 2)} \tag{2.64}$$
$$= \frac{P(X_1 = 1 \text{ and } X_2 = 2)}{P(X_1 = 1 \text{ and } X_2 = 2) + P(X_1 = 2 \text{ and } X_2 = 2)} \tag{2.65}$$
$$= \frac{P(X_1 = 1)P(X_2 = 2|X_1 = 1)}{P(X_1 = 1)P(X_2 = 2|X_1 = 1) + P(X_1 = 2)P(X_2 = 2|X_1 = 2)} \tag{2.66}$$

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

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)} \tag{2.67}$$

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.

2.13 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

$$\left(\begin{array}{ccc} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{array} \right) \tag{2.68}$$

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 just before 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 $P(v_4 \text{ attaches to } v_1)$ (in general, no longer assuming (2.68)). 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

Computer simulation essentially does in actual code what one does conceptually in our "notebook" view of probability (Section 2.2).

2.14.1 Example: Rolling Dice

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

```
# roll d dice; find P(total = k)

probtotk <- function(d,k,nreps) {
   count <- 0
   # do the experiment nreps times -- like doing nreps notebook lines
   for (rep in 1:nreps) {
      sum <- 0
      # roll d dice and find their sum
      for (j in 1:d) sum <- sum + roll()
      if (sum == k) count <- count + 1
   }
   return(count/nreps)
}

# simulate roll of one die; the possible return values are 1,2,3,4,5,6,
# all equally likely
roll <- function() return(sample(1:6,1))

# example
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:

```
# roll d dice; find P(total = k)

probtotk <- function(d,k,nreps) {
   count <- 0
   # do the experiment nreps times
   for (rep in 1:nreps)
      total <- sum(sample(1:6,d,replace=TRUE))
      if (total == k) count <- count + 1
   }
   return(count/nreps)
}
```

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:

```
# roll d dice; find P(total = k)

# simulate roll of nd dice; the possible return values are 1,2,3,4,5,6,
# all equally likely
roll <- function(nd) return(sample(1:6,nd,replace=TRUE))

```

```
probtotk <- function(d,k,nreps) {
   sums <- vector(length=nreps)
   # do the experiment nreps times
   for (rep in 1:nreps) sums[rep] <- sum(roll(d))
   return(mean(sums==k))
}
```

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:

```
roll <- function(nd) return(sample(1:6,nd,replace=TRUE))

probtotk <- function(d,k,nreps) {
   # do the experiment nreps times
```

[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.

```
   sums <- replicate(nreps,sum(roll(d)))
   return(mean(sums==k))
}
```

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 Example: 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.

Again, simulation is writing code the implements our "notebook" view of probability. In this case, we are working with a conditional probability, which our notebook view defined as follows. $P(B \mid A)$ is the long-run proportion of the time B occurs, *among those lines in which A occurs*. Here is the code:

```
dicesim <- function(nreps) {
   count1 <- 0
   count2 <- 0
   for (i in 1:nreps) {
      d <- sample(1:6,3,replace=T)
      if (sum(d) > 8) {  # "among those lines in which A occurs"
         count1 <- count1 + 1
         if (d[1] < 3) count2 <- count2 + 1
      }
   }
   return(count2 / count1)
}
```

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

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

2.14.6 Use of runif() for Simulating Events

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).

2.14.7 Example: ALOHA Network (cont'd.)

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

```
# finds P(X1 = 2), P(X2 = 2) and P(X2 = 2|X1 = 1) in ALOHA example
sim <- function(p,q,nreps) {
   countx2eq2 <- 0
   countx1eq1 <- 0
   countx1eq2 <- 0
   countx2eq2givx1eq1 <- 0
   # simulate nreps repetitions of the experiment
   for (i in 1:nreps) {
      numsend <- 0  # no messages sent so far
      # simulate A and B's decision on whether to send in epoch 1
      for (j in 1:2)
         if (runif(1) < p) numsend <- numsend + 1
      if (numsend == 1)  X1 <- 1
      else X1 <- 2
      if (X1 == 2) countx1eq2 <- countx1eq2 + 1
      # now simulate epoch 2
      # if X1 = 1 then one node may generate a new message
      numactive <- X1
      if (X1 == 1 && runif(1) < q) numactive <- numactive + 1
      # send?
      if (numactive == 1)
         if (runif(1) < p) X2 <- 0
         else X2 <- 1
      else {  # numactive = 2
         numsend <- 0
         for (i in 1:2)
            if (runif(1) < p) numsend <- numsend + 1
         if (numsend == 1) X2 <- 1
         else X2 <- 2
      }
      if (X2 == 2) countx2eq2 <- countx2eq2 + 1
      if (X1 == 1) {  # do tally for the cond. prob.
         countx1eq1 <- countx1eq1 + 1
```

```
         if (X2 == 2) countx2eq2givx1eq1 <- countx2eq2givx1eq1 + 1
      }
   }
   # print results
   cat("P(X1 = 2):",countx1eq2/nreps,"\n")
   cat("P(X2 = 2):",countx2eq2/nreps,"\n")
   cat("P(X2 = 2 | X1 = 1):",countx2eq2givx1eq1/countx1eq1,"\n")
}
```

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 (2.67) 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.

2.14.8 Example: Bus Ridership (cont'd.)

Consider the example in Section 2.11. 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:

```
nreps <- 10000
nstops <- 10
count <- 0
for (i in 1:nreps) {
   passengers <- 0
   for (j in 1:nstops) {
      if (passengers > 0)
         for (k in 1:passengers)
            if (runif(1) < 0.2)
               passengers <- passengers - 1
      newpass <- sample(0:2,1,prob=c(0.5,0.4,0.1))
      passengers <- passengers + newpass
   }
   if (passengers == 0) count <- count + 1
}
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.

2.14.9 Example: Board Game (cont'd.)

Recall the board game in Section 2.10. Below is simulation code to find the probability in (2.38):

```
boardsim <- function(nreps) {
   count4 <- 0
   countbonusgiven4 <- 0
   for (i in 1:nreps) {
      position <- sample(1:6,1)
      if (position == 3) {
         bonus <- TRUE
         position <- (position + sample(1:6,1)) %% 8
      } else bonus <- FALSE
      if (position == 4) {
         count4 <- count4 + 1
         if (bonus) countbonusgiven4 <- countbonusgiven4 + 1
      }
   }
   return(countbonusgiven4/count4)
}
```

2.14.10 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.

2.14.11 Example: Toss a Coin Until k Consecutive Heads

We toss a coin until we get k heads in a row. Let N denote the number of tosses needed, so that for instance the pattern HTHHH gives $N = 5$ for $k = 3$. Here is code that finds the approximate probability that $N > m$:

```
ngtm <- function(k,m,nreps) {
   count <- 0
   for (rep in 1:nreps) {
      consech <- 0
      for (i in 1:m) {
         toss <- sample(0:1,1)
         if (toss) {
            consech <- consech + 1
            if (consech == k) break
         } else consech <- 0
      }
      if (consech < k) count <- count + 1
   }
   return(count/nreps)
}
```

2.14.12 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 9.

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:

```
# use simulation to find P(1 king) when deal a 5-card hand from a
# standard deck

# think of the 52 cards as being labeled 1-52, with the 4 kings having
# numbers 1-4

sim <- function(nreps) {
   count1king <- 0  # count of number of hands with 1 king
   for (rep in 1:nreps) {
      hand <- sample(1:52,5,replace=FALSE)  # deal hand
      kings <- intersect(1:4,hand)  # find which kings, if any, are in hand
      if (length(kings) == 1) count1king <- count1king + 1
   }
   print(count1king/nreps)
}
```

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

[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.

$\binom{20-k}{1}$ 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.

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

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

2.16 Computational Complements

2.16.1 More on the replicate() Function

The call form of **replicate()** is

```
replicate(numberOfReplications,codeBlock
```

In our example in Section 2.14.4,

```
sums <- replicate(nreps,sum(roll(d)))
```

codeBlock was just a single statement, a call to R's **sum()** function. If more than one statement is to be executed, it must be done so in a *block*, a set of statements enclosed by braces, such as

```
f <- function()
{
   replicate(3,
      {
```

```
        x <- sample(1:10,5,replace=TRUE)
        c(mean(x),sd(x))
      }
    )
}
```

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 Misnomer

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, i.e. the long-run average in the X column of the notebook, is

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

To make this more explicit, look at the partial notebook example in Table 3.1. Here we roll two dice, and let S denote their sum. E(S) is then the long-run average of the values in the "S" column.

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

Here we will derive a handy computational formula for the expected value of a discrete random variable.

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.

Now 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]

So, we see that E(X) is a weighted average of the values in the support of X, with the weights being the probabilities of those values. You might wonder how to reconcile this with the fact that (3.7) is an *unweighted* average. But in fact it *is* weighted, because some values of X will appear more often than others in the X column of the notebook; indeed, the relative frequencies of those values will be given by P(X = c).

[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.

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.

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.)

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

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 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

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

  This should make sense. If the "random" variable X has the constant value 3, say, then the "X" column in the notebook will consist entirely of 3s. Thus the long-run average value in that column will be 3, so EX = 3.

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 **??**.

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

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

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 income 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.

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

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 **??** for details.)

As with expected values, the properties of variance discussed above, and also in Section ?? 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 **??**, 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,

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

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 **??** 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

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

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, Version I

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

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

3.9.2 Example: Return Time for Library Books, Version II

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)
}
```

Note that in this simulation. all **nreps** values of I, B and the patron return time are generated first. This uses more memory space (though not an issue in this small problem), but makes things easier to code, as we can exploit R's vector operations. Those not only are more convenient, but also faster running.

3.9.3 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 ith 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 <u>not</u> 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

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

3.9.4 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 the reasoning in (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")
   cat("Var(X2):",sumx2sq/nreps - meanx2^2,"\n")
   cat("Cov(X1,X2):",sumx1x2/nreps - meanx1*meanx2,"\n")
}
```

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 2.11, 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

[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*.

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.

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).

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

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

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

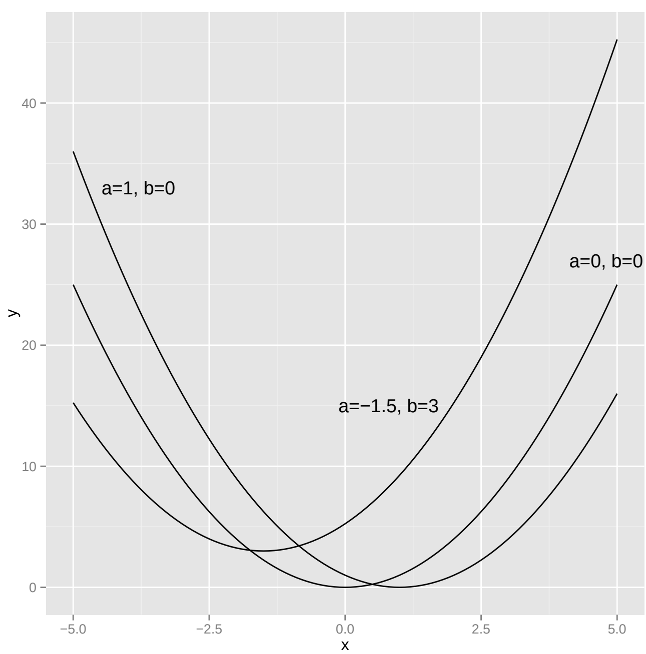
the left-hand side of (3.120) is the same as the left-hand side of (3.63). The numerator of the right-hand size of (3.120) is simply Var(X), i.e. σ^2, so we are done.

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 $|t| < 1$,

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

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 6.3.) 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:

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

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.

Or consider a different question: 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:

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

and from (3.75),

$$Var(X) = Var(B_1 + ... + B_n) = Var(B_1) + ... + Var(B_n) = np(1-p) \tag{4.30}$$

Again, (4.29) should make good intuitive sense to you.

4.3.1 R Functions

Relevant functions for a binomially distributed random variable X for k trials and with success probability p are:

- **dbinom(i,k,p)**, to find $P(X = i)$
- **pbinom(i,k,p)**, to find $P(X \leq i)$
- **qbinom(q,k,p)**, to find c such that $P(X \leq c) = q$
- **rbinom(n,k,p)**, to generate n independent values of X

Our definition above of **qbinom()** is not quite tight, though. Consider a random variable X which has a binomial distribution with n = 2 and p = 0.5. Then

$$F_X(0) = 0.25,\ F_X(1) = 0.75 \tag{4.31}$$

So if q is, say, 0.33, there is no c such that $P(X \leq c) = q$. For that reason, the actual definition of **qbinom()** is the smallest c satisfying $P(X \leq c) \geq q$.

4.3.2 Example: Parking Space Model

Recall Section 4.2.2. Let's find the probability that there are three open spaces in the first block.

Let M denote the number of open spaces in the first block. This fits the definition of binomially-distributed random variables: We have a fixed number (10) of independent Bernoulli trials, and we are interested in the number of successes. So, for instance,

$$p_M(3) = \binom{10}{3} 0.15^3 (1-0.15)^{10-3} \tag{4.32}$$

4.4 The Negative Binomial Family of Distributions

Recall that a typical example of the geometric distribution family (Section 4.2) arises as N, the number of tosses of a coin needed to get our first head. Now generalize that, with N now being the number of tosses needed to get our r^{th} head, where r is a fixed value. Let's find P(N = k), k = r, r+1, ... For concreteness, look at the case r = 3, k = 5. In other words, we are finding the probability that it will take us 5 tosses to accumulate 3 heads.

First note the equivalence of two events:

$$\{N = 5\} = \{2 \text{ heads in the first 4 tosses and head on the } 5^{th} \text{ toss}\} \tag{4.33}$$

That event described before the "and" corresponds to a binomial probability:

$$P(2 \text{ heads in the first 4 tosses}) = \binom{4}{2} \left(\frac{1}{2}\right)^4 \tag{4.34}$$

Since the probability of a head on the k^{th} toss is 1/2 and the tosses are independent, we find that

$$P(N = 5) = \binom{4}{2} \left(\frac{1}{2}\right)^5 = \frac{3}{16} \tag{4.35}$$

The negative binomial distribution family, indexed by parameters r and p, corresponds to random variables that count the number of independent trials with success probability p needed until we get r successes. The pmf is

$$p_N(k) = P(N = k) = \binom{k-1}{r-1} (1-p)^{k-r} p^r, k = r, r+1, ... \tag{4.36}$$

We can write

$$N = G_1 + ... + G_r \tag{4.37}$$

where G_i is the number of tosses between the successes numbers i-1 and i. But each G_i has a geometric distribution! Since the mean of that distribution is 1/p, we have that

$$E(N) = r \cdot \frac{1}{p} \tag{4.38}$$

In fact, those r geometric variables are also independent, so we know the variance of N is the sum of their variances:

$$Var(N) = r \cdot \frac{1-p}{p^2} \tag{4.39}$$

4.4.1 R Functions

Relevant functions for a negative binomial distributed random variable X with success parameter p are:

- **dnbinom(i,size=1,prob=p)**, to find $P(X = i)$
- **pnbinom(i,size=1,prob=p)**, to find $P(X <= i)$
- **qnbinom(q,sixe=1,prob=p)**, to find c such that $P(X \leq c) = q$
- **rnbinom(n,size=1,prob=p)**, to generate n independent values of X

Here **size** is our r. Note, though, that as with the **geom()** family, R defines the distribution in terms of number of failures. So, in **dbinom()**, the argument **i** is the number of failures, and **i + r** is our X.

4.4.2 Example: Backup Batteries

A machine contains one active battery and two spares. Each battery has a 0.1 chance of failure each month. Let L denote the lifetime of the machine, i.e. the time in months until the third battery failure. Find P(L = 12).

The number of months until the third failure has a negative binomial distribution, with r = 3 and p = 0.1. Thus the answer is obtained by (4.36), with k = 12:

$$P(L = 12) = \binom{11}{2}(1-0.1)^9 0.1^3 \tag{4.40}$$

4.5 The Poisson Family of Distributions

Another famous parametric family of distributions is the set of **Poisson Distributions**.

This family is a little different from the geometric, binomial and negative binomial families, in the sense that in those cases there were qualitative descriptions of the settings in which such distributions arise. Geometrically distributed random variables, for example occur as the number of Bernoulli trials needed to get the first success.

By contrast, the Poisson family does not really have this kind of qualitative description.[5] It is merely something that people have found to be a reasonably accurate model of actual data in many cases. We might be interested, say, in the number of disk drive failures in periods of a specified length of time. If we have data on this, we might graph it, and if it looks like the pmf form below, then we might adopt it as our model.

The pmf is

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, ... \tag{4.41}$$

It turns out that

$$EX = \lambda \tag{4.42}$$

$$Var(X) = \lambda \tag{4.43}$$

The derivations of these facts are similar to those for the geometric family in Section 4.2. One starts with the Maclaurin series expansion for e^t:

$$e^t = \sum_{i=0}^{\infty} \frac{t^i}{i!} \tag{4.44}$$

and finds its derivative with respect to t, and so on. The details are left to the reader.

The Poisson family is very often used to model count data. For example, if you go to a certain bank every day and count the number of customers who arrive between 11:00 and 11:15 a.m., you will probably find that that distribution is well approximated by a Poisson distribution for some λ.

There is a lot more to the Poisson story than we see in this short section. We'll return to this distribution family in Section 8.1.

[5] Some such descriptions are possible in the Poisson case, but they are complicated and difficult to verify.

4.5.1 R Functions

Relevant functions for a Poisson distributed random variable X with parameter lambda are:

- **dpois(i,lambda)**, to find $P(X = i)$
- **ppois(i,lambda)**, to find $P(X \leq i)$
- **qpois(q,lambda)**, to find c such that $P(X \leq c) = q$
- **rpois(n,lambda)**, to generate n independent values of X

4.5.2 Example: Broken Rod

Recall the example of a broken glass rod in Section 2.14.10. Suppose now that the number of breaks is random, not just the break points. A reasonable model to try would be Poisson. However, the latter's support starts at 0, and we cannot have 0 pieces, so we need to model the number of pieces minus 1 (the number of break points) as Poisson.

The code is similar to that in Section 2.14.10, but we must first generate the number of break points:

```
minpiecepois <- function(lambda) {
   nbreaks <- rpois(1,lambda)
   breakpts <- sort(runif(nbreaks))
   lengths <- diff(c(0,breakpts,1))
   min(lengths)
}

bkrodpois <- function(nreps,lambda,q) {
   minpieces <- replicate(nreps,minpiecepois(lambda))
   mean(minpieces < q)
}

> bkrodpois(10000,5,0.02)
[1] 0.4655
```

Note that in each call to **minpiecepois()**, there will be a different number of breakpoints.

4.6 The Power Law Family of Distributions

This family has attracted quite a bit of attention in recent years, due to its use in random graph models.

4.6.1 The Model

Here

$$p_X(k) = ck^{-\gamma},\ \ k = 1, 2, 3, ... \tag{4.45}$$

It is required that $\gamma > 1$, as otherwise the sum of probabilities will be infinite. For γ satisfying that condition, the value c is chosen so that that sum is 1.0:

$$1.0 = \sum_{k=1}^{\infty} ck^{-\gamma} \approx c \int_1^{\infty} k^{-\gamma}\ dk = c/(\gamma - 1) \tag{4.46}$$

so $c \approx \gamma - 1$.

Here again we have a parametric family of distributions, indexed by the parameter γ.

The power law family is an old-fashioned model (an old-fashioned term for *distribution* is *law*), but there has been a resurgence of interest in it in recent years. Analysts have found that many types of social networks in the real world exhibit approximately power law behavior in their degree distributions.

For instance, in a famous study of the Web (A. Barabasi and R. Albert, Emergence of Scaling in Random Networks, *Science*, 1999, 509-512), degree distribution on the Web (a directed graph, with incoming links being the ones of interest here) it was found that the number of links leading to a Web page has an approximate power law distribution with $\gamma = 2.1$. The number of links leading out of a Web page was found to be approximately power-law distributed, with $\gamma = 2.7$.

Much of the interest in power laws stems from their **fat tails**, a term meaning that values far from the mean are more likely under a power law than they would be under a normal distribution with the same mean. In recent popular literature, values far from the mean have often been called **black swans**. The financial crash of 2008, for example, is blamed by some on the ignorance by **quants** (people who develop probabilistic models for guiding investment) in underestimating the probabilities of values far from the mean.

Some examples of real data that are, or are not, fit well by power law models are given in the paper *Power-Law Distributions in Empirical Data*, by A. Clauset, C. Shalizi and M. Newman, at `http://arxiv.org/abs/0706.1062`. Methods for estimating the parameter γ are discussed and evaluated.

A variant of the power law model is the **power law with exponential cutoff**, which essentially consists of a blend of the power law and a geometric distribution. Here

$$p_X(k) = ck^{-\gamma}q^k \tag{4.47}$$

This now is a two-parameter family, the parameters being γ and q. Again c is chosen so that the pmf sums to 1.0.

This model is said to work better than a pure power law for some types of data. Note, though, that this version does not really have the fat tail property, as the tail decays exponentially now.

4.6.2 Further Reading

There is nice paper on fitting (or not fitting) power law models:

Power-Law Distributions in Empirical Data, *SIAM Review*, A. Clauset, C.R. Shalizi, and M.E.J. Newman, 51(4), 661-703, 2009.

4.7 Recognizing Some Parametric Distributions When You See Them

Three of the discrete distribution families we've considered here arise in settings with very definite structure, all dealing with independent trials:

- the binomial family gives the distribution of the number of successes in a fixed number of trials
- the geometric family gives the distribution of the number of trials needed to obtain the first success
- the negative binomial family gives the distribution of the number of trials needed to obtain the kth success

Such situations arise often, hence the fame of these distribution families.

By contrast, the Poisson and power law distributions have no underlying structure. They are famous for a different reason, that it has been found empirically that they provide a good fit to many real data sets.

In other words, the Poisson and power law distributions are typically fit to data, in an attempt to find a good model, whereas in the binomial, geometric and negative binomial cases, the fundamental nature of the setting implies one of those distributions.

You should make a strong effort to get to the point at which you automatically recognize such settings when your encounter them.

4.8 Example: a Coin Game

Life is unfair—former President Jimmy Carter

Consider a game played by Jack and Jill. Each of them tosses a coin many times, but Jack gets a head start of two tosses. So by the time Jack has had, for instance, 8 tosses, Jill has had only 6; when Jack tosses for the 15^{th} time, Jill has her 13^{th} toss; etc.

Let X_k denote the number of heads Jack has gotten through his k^{th} toss, and let Y_k be the head count for Jill at that same time, i.e. among only k-2 tosses for her. (So, $Y_1 = Y_2 = 0$.) Let's find the probability that Jill is winning after the 6^{th} toss, i.e. $P(Y_6 > X_6)$.

Your first reaction might be, "Aha, binomial distribution!" You would be on the right track, but the problem is that you would not be thinking precisely enough. Just WHAT has a binomial distribution? The answer is that both X_6 and Y_6 have binomial distributions, both with p = 0.5, but n = 6 for X_6 while n = 4 for Y_6.

Now, as usual, ask the famous question, "How can it happen?" How can it happen that $Y_6 > X_6$? Well, we could have, for example, $Y_6 = 3$ and $X_6 = 1$, as well as many other possibilities. Let's write it mathematically:

$$P(Y_6 > X_6) = \sum_{i=1}^{4} \sum_{j=0}^{i-1} P(Y_6 = i \text{ and } X_6 = j) \tag{4.48}$$

Make SURE your understand this equation.

Now, to evaluate $P(Y_6 = i \text{ and } X_6 = j)$, we see the "and" so we ask whether Y_6 and X_6 are independent. They in fact are; Jill's coin tosses certainly don't affect Jack's. So,

$$P(Y_6 = i \text{ and } X_6 = j) = P(Y_6 = i) \cdot P(X_6 = j) \tag{4.49}$$

It is at this point that we finally use the fact that X_6 and Y_6 have binomial distributions. We have

$$P(Y_6 = i) = \binom{4}{i} 0.5^i (1 - 0.5)^{4-i} \tag{4.50}$$

and

$$P(X_6 = j) = \binom{6}{j} 0.5^j (1 - 0.5)^{6-j} \tag{4.51}$$

We would then substitute (4.50) and (4.51) in (4.48). We could then evaluate it by hand, but it would be more convenient to use R's **dbinom()** function:

```
prob <- 0
for (i in 1:4)
   for (j in 0:(i-1))
      prob <- prob + dbinom(i,4,0.5) * dbinom(j,6,0.5)
prob
```

We get an answer of about 0.17. If Jack and Jill were to play this game repeatedly, stopping each time after the 6^{th} toss, then Jill would win about 17% of the time.

4.9 Example: Tossing a Set of Four Coins

Consider a game in which we have a set of four coins. We keep tossing the set of four until we have a situation in which exactly two of them come up heads. Let N denote the number of times we must toss the set of four coins.

For instance, on the first toss of the set of four, the outcome might be HTHH. The second might be TTTH, and the third could be THHT. In the situation, N = 3.

Let's find P(N = 5). Here we recognize that N has a geometric distribution, with "success" defined as getting two heads in our set of four coins. What value does the parameter p have here?

Well, p is P(X = 2), where X is the number of heads we get from a toss of the set of four coins. We recognize that X is binomial! Thus

$$p = \binom{4}{2} 0.5^4 = \frac{3}{8} \tag{4.52}$$

Thus using the fact that N has a geometric distribution,

$$P(N=5) = (1-p)^4 p = 0.057 \tag{4.53}$$

4.10 Example: the ALOHA Example Again

As an illustration of how commonly these parametric families arise, let's again look at the ALOHA example. Consider the general case, with transmission probability p, message creation probability q, and m network nodes. We will not restrict our observation to just two epochs.

Suppose $X_i = m$, i.e. at the end of epoch i all nodes have a message to send. Then the number which attempt to send during epoch i+1 will be binomially distributed, with parameters m and p.[6] For instance, the probability that there is a successful transmission is equal to the probability that exactly one of the m nodes attempts to send,

$$\binom{m}{1} p(1-p)^{m-1} = mp(1-p)^{m-1} \tag{4.54}$$

Now in that same setting, $X_i = m$, let K be the number of epochs it will take before some message actually gets through. In other words, we will have $X_i = m$, $X_{i+1} = m$, $X_{i+2} = m$,... but finally $X_{i+K-1} = m-1$. Then K will be geometrically distributed, with success probability equal to (4.54).

There is no Poisson distribution in this example, but it is central to the analysis of Ethernet, and almost any other network. We will discuss this at various points in later chapters.

4.11 Example: the Bus Ridership Problem Again

Recall the bus ridership example of Section 2.11. Let's calculate some expected values, for instance $E(B_1)$:

$$E(B_1) = 0 \cdot P(B_1 = 0) + 1 \cdot P(B_1 = 1) + 2 \cdot P(B_1 = 2) = 0.4 + 2 \cdot 0.1 \tag{4.55}$$

Now suppose the company charges \$3 for passengers who board at the first stop, but charges \$2 for those who join at the second stop. (The latter passengers get a possibly shorter ride, thus pay

[6] Note that this is a conditional distribution, given $X_i = m$.

less.) So, the total revenue from the first two stops is $T = 3B_1 + 2B_2$. Let's find E(T). We'd write

$$E(T) = 3E(B_1) + 2E(B_2) \tag{4.56}$$

making use of (3.21). We'd then compute the terms as in 4.55.

Suppose the bus driver has the habit of exclaiming, "What? No new passengers?!" every time he comes to a stop at which $B_i = 0$. Let N denote the number of the stop (1,2,...) at which this first occurs. Find P(N = 3):

N has a geometric distribution, with p equal to the probability that there 0 new passengers at a stop, i.e. 0.5. Thus $p_N(3) = (1-0.5)^2 0.5$, by (4.3).

Let S denote the number of stops, out of the first 6, at which 2 new passengers board. For example, S would be 3 if $B_1 = 2$, $B_2 = 2$, $B_3 = 0$, $B_4 = 1$, $B_5 = 0$, and $B_6 = 2$. Find $p_S(4)$:

S has a binomial distribution, with n = 6 and p = probability of 2 new passengers at a stop = 0.1. Then

$$p_S(4) = \binom{6}{4} 0.1^4 (1-0.1)^{6-4} \tag{4.57}$$

By the way, we can exploit our knowledge of binomial distributions to simplify the simulation code in Section 2.14.8. The lines

```
for (k in 1:passengers)
   if (runif(1) < 0.2)
      passengers <- passengers - 1
```

simulate finding that number of passengers that alight at that stop. But that number is binomially distributed, so the above code can be compactified (and speeded up in execution) as

```
passengers <- passengers - rbinom(1,passengers,0.2)
```

4.12 Example: Flipping Coins with Bonuses

A game involves flipping a coin k times. Each time you get a head, you get a bonus flip, not counted among the k. (But if you get a head from a bonus flip, that does not give you its own bonus flip.) Let X denote the number of heads you get among all flips, bonus or not. Let's find the distribution of X.

As with the parking space example above, we should be careful not to come to hasty conclusions. The situation here "sounds" binomial, but X, based on a variable number of trials, doesn't fit the definition of binomial.

But let Y denote the number of heads you obtain through nonbonus flips. Y then has a binomial distribution with parameters k and 0.5. To find the distribution of X, we'll condition on Y.

We will as usual ask, "How can it happen?", but we need to take extra care in forming our sums, recognizing constraints on Y:

- $Y \geq X/2$
- $Y \leq X$
- $Y \leq k$

Keeping those points in mind, we have

$$
\begin{aligned}
p_X(m) &= P(X = m) && (4.58)\\
&= \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} P(X = m \text{ and } Y = i) && (4.59)\\
&= \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} P(X = m|Y = i)\ P(Y = i) && (4.60)\\
&= \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} \binom{i}{m-i} 0.5^i \binom{k}{i} 0.5^k && (4.61)\\
&= 0.5^k \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} \frac{k!}{(m-i)!(2i-m)!(k-i)!} 0.5^i && (4.62)
\end{aligned}
$$

There doesn't seem to be much further simplification possible here.

4.13 Example: Analysis of Social Networks

Let's continue our earlier discussion from Section 3.13.3.

One of the earliest—and now the simplest—models of social networks is due to Erdös and Renyi. Say we have n people (or n Web sites, etc.), with $\binom{n}{2}$ potential links between pairs. (We are assuming an undirected graph here.) In this model, each potential link is an actual link with probability p, and a nonlink with probability 1-p, with all the potential links being independent.

Recall the notion of degree distribution from Section 3.13.3. Clearly the degree distribution D_i here for a single node i is binomial with parameters n-1 and p.

But consider k nodes, say 1 through k, among the n total nodes, and let T denote the number of links involving these nodes. Let's find the distribution of T. That distribution is again binomial, but the number of trials must be carefully calculated. We cannot simply say that, since each of the k vertices can have as many as n-1 links, there are k(n-1) potential links, because there is overlap; two nodes among the k have a potential link with each other, but we can't count it twice. So, let's reason this out.

Say n = 9 and k = 4. Among the four special nodes, there are $\binom{4}{2} = 6$ potential links, each on or off with probability p, independently. Also each of the four special nodes has $9 - 4 = 5$ potential links with the "outside world," i.e. the five non-special nodes. So there are $4 \times 5 = 20$ potential links here, for a total of 26.

So, the distribution of T is binomial with

$$k(n-k) + \binom{k}{2} \tag{4.63}$$

trials and success probability p.

4.14 Multivariate Distributions

(I am borrowing some material here from Section **??**, for instructors or readers who skip Chapter **??**. It is important to know that multivariate distributions exist, even if one doesn't know the details.)

Recall that for a single discrete random variable X, the distribution of X was defined to be a list of all the values of X, together with the probabilities of those values. The same is done for a pair (or more than a pair) of discrete random variables U and V.

Suppose we have a bag containing two yellow marbles, three blue ones and four green ones. We choose four marbles from the bag at random, without replacement. Let Y and B denote the number of yellow and blue marbles that we get. Then define the *two-dimensional* pmf of Y and B to be

$$p_{Y,B}(i,j) = P(Y = i \text{ and } B = j) = \frac{\binom{2}{i}\binom{3}{j}\binom{4}{4-i-j}}{\binom{9}{4}} \tag{4.64}$$

Here is a table displaying all the values of P(Y = i and B = j):

i ↓, j →	0	1	2	3
0	0.008	0.095	0.143	0.032
1	0.063	0.286	0.190	0.016
2	0.048	0.095	0.024	0.000

So this table is the distribution of the pair (Y,B).

Recall further that in the discrete case, we introduced a symbolic notation for the distribution of a random variable X, defined as $p_X(i) = P(X = i)$, where i ranged over the support of X. We do the same thing for a pair of random variables:

Definition 8 *For discrete random variables U and V, their probability mass function is defined to be*

$$p_{U,V}(i,j) = P(U = i \text{ and } V = j) \tag{4.65}$$

where (i,j) ranges over all values taken on by (U,V). Higher-dimensional pmfs are defined similarly, e.g.

$$p_{U,V,W}(i,j,k) = P(U = i \text{ and } V = j \text{ and } W = k) \tag{4.66}$$

So in our marble example above, $p_{Y,B}(1,2) = 0.286$, $p_{Y,B}(2,0) = 0.048$ and so on.

4.15 Iterated Expectations

This section has an abstract title, but the contents are quite useful.

4.15.1 Conditional Distributions

Just as we can define bivariate pmfs, we can also speak of conditional pmfs. Suppose we have random variables U and V.

In our bus ridership example, for instance, we can talk about

$$E(L_2 \mid B_1 = 0) \tag{4.67}$$

In notebook terms, think of many replications of watching the bus during times 1 and 2. Then (4.67) is defined to be the long-run average of values in the L_2 column, **among those rows** in which the B_1 column is 0. (And by the way, make sure you understand why (4.67) works out to be 0.6.)

4.15.2 The Theorem

The key relation says, in essence,

> The overall mean of V is a weighted average of the conditional means of V given U. The weights are the pmf of U.

Note again that $E(V \mid U = c)$ is defined in "notebook" terms as the long-run average of V, *among those lines in which* $U = c$.

Here is the formal version:

Suppose we have random variables U and V, with U discrete and with V having an expected value. Then

$$E(V) = \sum_c P(U = c)\ E(V \mid U = c) \tag{4.68}$$

where c ranges through the support of U.

So, just as EX was a weighted average in (3.13), with weights being probabilities, we see here that the unconditional mean is a weighted average of the conditional mean.

In spite of its intimidating form, (4.68) makes good intuitive sense, as follows: Suppose we want to find the average height of all students at a university. Each department measures the heights of its majors, then reports the mean height among them. Then (4.68) says that to get the overall mean in the entire school, we should take a *weighted* average of all the within-department means, with the weights being the proportions of each department's student numbers among the entire school. Clearly, we would not want to take an unweighted average, as that would count tiny departments just as much as large majors.

Here is the derivation (reader: supply the reasons!).

$$
\begin{aligned}
EV &= \sum_d d \ P(V = d) && (4.69)\\
&= \sum_d d \ \sum_c P(U = c \text{ and } V = d) && (4.70)\\
&= \sum_d d \ \sum_c P(U = c) \ P(V = d \mid U = c) && (4.71)\\
&= \sum_d \sum_c d \ P(U = c) \ P(V = d \mid U = c) && (4.72)\\
&= \sum_c \sum_d d \ P(U = c) \ P(V = d \mid U = c) && (4.73)\\
&= \sum_c P(U = c) \sum_d d \ P(V = d \mid U = c) && (4.74)\\
&= \sum_c P(U = c) \ E(V \mid U = c) && (4.75)
\end{aligned}
$$

4.15.3 Example: Coin and Die Game

You roll a die until it comes up 5, taking M rolls to do so. You then toss a coin M times, winning one dollar for each head. Find the expected winnings, EW.

Solution: Given M = k, the number of heads has a binomial distribution with n = k and p = 0.5. So

$$E(W|M = k) = 0.5k. \tag{4.76}$$

So, from (4.68), we have

$$EW = \sum_{k=1}^{\infty} P(M = k) \ 0.5k = 0.5 \sum_{k=1}^{\infty} P(M = k)k = 0.5 \ EM \tag{4.77}$$

from (3.13). And from (4.4), we know EM = 6. So, EW = 3.

Note especially that we didn't need calculate P(M = k).

4.15.4 Example: Flipping Coins with Bonuses

Recall the example in Section 4.12. The principle of iterated expectation can easily get us EX:

$$
\begin{aligned}
EX &= \sum_{i=1}^{k} P(Y=i)\ E(X|Y=i) && (4.78)\\
&= \sum_{i=1}^{k} P(Y=i)\ (i+0.5i) && (4.79)\\
&= E(1.5Y) && (4.80)\\
&= 1.5\cdot k/2 && (4.81)\\
&= 0.75k && (4.82)
\end{aligned}
$$

To understand that second equality, note that if $Y = i$, X will already include those i heads, the nonbonus heads, and since there will be i bonus flips, the expected value of the number of bonus heads will be $0.5i$. The third equality uses (3.36) (in a reverse manner).

Chapter 5

Introduction to Discrete Markov Chains

(From this point onward, we will be making heavy use of linear algebra. The reader may find it helpful to review Appendix **??**.)

Here we introduce Markov chains, a topic covered in much more detail in Chapter **??**.

The basic idea is that we have random variables $X_1, X_2, ...$, with the index representing time. Each one can take on any value in a given set, called the **state space**; X_n is then the **state** of the system at time n. The state space is assumed either finite or **countably infinite**.[1]

We sometimes also consider an initial state, X_0, which might be modeled as either fixed or random. However, this seldom comes into play.

The key assumption is the **Markov property**, which in rough terms can be described as:

> The probabilities of future states, given the present state and the past states, depends only on the present state; the past is irrelevant.

In formal terms:

$$P(X_{t+1} = s_{t+1} | X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) = P(X_{t+1} = s_{t+1} | X_t = s_t) \tag{5.1}$$

[1]The latter is a mathematical term meaning, in essence, that it is possible to denote the space using integer subscripts. It can be shown that the set of all real numbers is not countably infinite, though perhaps surprisingly the set of all rational numbers *is* countably infinite.

Note that in (5.1), the two sides of the equation are equal but their common value may depend on t. We assume that this is not the case; nondependence on t is known as **stationarity**.[2] For instance, the probability of going from state 2 to state 5 at time 29 is assumed to be the same as the corresponding probability at time 333.

5.1 Matrix Formulation

We define p_{ij} to be the probability of going from state i to state j in one time step; note that this is a *conditional* probability, i.e. $P(X_{n+1} = j \mid X_n = i)$. These quantities form a matrix P,[3] whose row i, column j element is p_{ij}, which is called the **transition matrix**. Each row of P must sum to 1 (do you see why?).

For example, consider a three-state Markov chain with transition matrix

$$P = \left(\begin{array}{ccc} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 1 & 0 & 0 \end{array} \right) \tag{5.2}$$

This means that if we are now at state 1, the probabilities of going to states 1, 2 and 3 are 1/2, 0 and 1/2, respectively. Note that each row's probabilities must sum to 1—after all, from any particular state, we must go *somewhere*.

Actually, the mth power, P^m, of the transition matrix gives the probabilities for m-step transitions. In other words, the (i,j) element of P^m is $P(X_{t+m} = j | X_t = i)$. This is clear for the case m = 2 (after which one can use mathematical induction), as follows.

As usual, "break big events down into small events." How can it happen that $X_{t+2} = j$? Well, break things down according to where we might go first after leaving i. We might go from i to 1, say, then from 1 to j. So,

$$P(X_{t+2} = j | X_t = i) = \sum_k p_{ik} \ p_{kj} \tag{5.3}$$

In view of the rule for multiplying matrices, the expression on the right-hand side is simply the (i,j) element of P^2!

[2] Not to be confused with the notion of a stationary distribution, coming below.

[3] Unfortunately, we have some overloading of symbols here. Both in this book and in the field in general, we usually use the letter P to denote this matrix, yet we continue to denote probabilities by P( ). However, it will be clear from context which we mean. The same is true for our transition probabilities p_{ij}, which use a subscripted letter p, which is also the case for probably mass functions.

5.2 Example: Die Game

As our first example of Markov chains, consider the following game. One repeatedly rolls a die, keeping a running total. Each time the total exceeds 10 (not equals 10), we receive one dollar, and continue playing, resuming where we left off, mod 10. Say for instance we have a total of 8, then roll a 5. We receive a dollar, and now our total is 3.

It will simplify things if we assume that the player starts with one free point.

This process clearly satisfies the Markov property, with our state being our current total, 1, 2, ..., 10. If our current total is 6, for instance, then the probability that we next have a total of 9 is 1/6, *regardless of what happened our previous rolls.* We have p_{25}, p_{72} and so on all equal to 1/6, while for instance $p_{29} = 0$. Here's the code to find the transition matrix P:

```
# 10 states, so 10x10 matrix
# since most elements will be 0s, set them all to 0 first,
# then replace those that should be nonzero
p <- matrix(rep(0,100),nrow=10)
onesixth <- 1/6
for (i in 1:10) {  # look at each row
   # since we are rolling a die, there are only 6 possible states we
   # can go to from i, so check these
   for (j in 1:6) {
      k <- i + j  # new total, but did we win?
      if (k > 10) k <- k - 10
      p[i,k] <- onesixth
   }
}
```

Note that since we knew that many entries in the matrix would be zero, it was easier just to make them all 0 first, and then fill in the nonzero ones.

5.3 Long-Run State Probabilities

Let N_{it} denote the number of times we have visited state i during times 1,...,t. For instance, in the die game, $N_{8,22}$ would be the number of rolls among the first 22 that resulted in our having a total of 8.

In typical applications we have that

$$\pi_i = \lim_{t \to \infty} \frac{N_{it}}{t} \tag{5.4}$$

exists for each state i, regardless of where we start.[4] Under a couple more conditions,[5] we have the stronger result,

$$\lim_{t \to \infty} P(X_t = i) = \pi_i \tag{5.5}$$

These quantities π_i are typically the focus of analyses of Markov chains.

We will use the symbol π to name the column vector of all the π_i:

$$\pi = (\pi_1, \pi_2, ...)' \tag{5.6}$$

where $'$ means matrix transpose.

5.3.1 Stationary Distribution

The π_i are called **stationary probabilities**, because if the initial state X_0 is a random variable with that distribution, then all X_i will have that distribution. Here's why:

Using (5.5), we have

$$\begin{aligned} \pi_i &= \lim_{n \to \infty} P(X_n = i) && (5.7) \\ &= \lim_{n \to \infty} \sum_k P(X_{n-1} = k)\ p_{ki} && (5.8) \\ &= \sum_k \pi_k\ p_{ki} && (5.9) \end{aligned}$$

[4] A more mathematically rigorous statement of (5.4) would include the qualifier "with probability 1." What does does mean?

The infinite sequeunce $X_1, X_2, ...$ is called a **sample path**. In our notebook, we have a column for each X_i, so one row is one sample path. So, what (**??**) says, with the phrase "with probability 1" added, is that the limit holds *for each row.*

[5] Basically, we need the chain to not be **periodic**. Consider a random walk, for instance: We start at position 0 on the number line, at time 0. The states are the integers. (So, this chain has an infinite state space.) At each time, we flip a coin to determine whether to move right (heads) or left (tails) 1 unit. A little thought shows that if we start at 0, the only times we can return to 0 are even-number times, i.e. $P(X_n = 0\ |X_0 = 0)$ for all odd numbers n. This is a periodic chain. By the way, (5.4) turns out to be 0 for this chain.

So, if $P(X_0 = i) = \pi_i$, then

$$P(X_1 = i) = \sum_k P(X_0 = k)\ p_{ki} \tag{5.10}$$

$$= \sum_k \pi_k\ p_{ki} \tag{5.11}$$

$$= \pi_i \tag{5.12}$$

this last using (5.9). So, if X_0 has distribution π, then the same will be true for X_1, and contiuing in this manner we see that $X_2, X_3, ...$ will all have that distribution, thus demostrating the claimed stationary property for π. (This will be illustrated more concretely in Section 5.4.2.)

Of course, (5.7) holds for all states i. So in matrix terms, (5.7) says

$$\pi' = \pi' P \tag{5.13}$$

5.3.2 Calculation of π

Equation (5.13) then shows us how to find the π_i, at least in the case of finite state spaces, the subject of this section here, as follows.

First, rewrite (5.13)

$$(I - P')\pi = 0 \tag{5.14}$$

Here I is the $n \times n$ identity matrix (for a chain with n states), and again $'$ denotes matrix transpose.

This equation has infinitely many solutions; if π is a solution, then so for example is 8π. Moreover, the equation shows that the matrix there, $I - P'$, cannot have an inverse; if it did, we could multiply both sides by the inverse, and find that the unique solution is $\pi = 0$, which can't be right. This says in turn that the rows of $I - P'$ are not linearly independent.

The latter fact is quite important, for the following reason. Recall the close connection of matrix inverse and systems of linear equations. (5.14), a matrix equation, represents a system of n linear equations in n unknowns, the latter being $\pi_1, \pi_2, .., \pi_n$. So, the lack of linear independence of the rows of $I - P'$ means, in plain English, that at least one of those equations is redundant.

But we need n independent equations, and fortunately one is available:

$$\sum_i \pi_i = 1 \tag{5.15}$$

Note that (5.15) can be written as

$$O'\pi = 1 \tag{5.16}$$

where O is a vector of n 1s. Excellent, let's use it!

So, again, thinking of (5.14) as a system of linear equations, let's replace the last equation by (5.16). Switching back to the matrix view, that means that we replace the last row in the matrix $I - P'$ in (5.14) by O', and correspondingly replace the last element of the right-side vector by 1. Now we have a nonzero vector on the right side, and a full-rank (i.e. invertible) matrix on the left side. This is the basis for the following code, which we will use for finding π.

```
findpi1 <- function(p) {
   n <- nrow(p)
   # find I-P'
   imp <- diag(n) - t(p)
   # replace the last row of I-P' as discussed
   imp[n,] <- rep(1,n)
   # replace the corresponding element of the
   # right side by (the scalar) 1
   rhs <- c(rep(0,n-1),1)
   # now use R's built-in solve()
   solve(imp,rhs)
}
```

5.3.2.1 Example: π in Die Game

Consider the die game example above. Guess what! Applying the above code, all the π_i turn out to be 1/10. In retrospect, this should be obvious. If we were to draw the states 1 through 10 as a ring, with 1 following 10, it should be clear that all the states are completely symmetric.

5.3.2.2 Another Way to Find π

Here is another way to compute π. It is not commonly used, but **it will also help illustrate some of the concepts.**

Suppose (5.5) holds. Recall that P^m is the m-step transition matrix, so that for instance row 1 of that matrix is the set of probabilities of going from state 1 to the various states in m steps. The same will be true for row 2 and so on. Putting that together with (5.5), we have that

$$\lim_{n \to \infty} P^n = \Pi \tag{5.17}$$

where the $n \times n$ matrix Π has each of its rows equal to π'.

We can use this to find π. We take P to a large power m, and then each of the rows will approximate π. In fact, we can get an even better appoximation by averaging the rows.

Moreover, we can save a lot of computation by noting the following. Say we want the 16^{th} power of P. We could set up a loop with 15 iterations, building up a product. But actually we can do it with just 4 iterations. We first square P, yielding P^2. But then we square *that*, yielding P^4. Square twice more, yielding P^8 and finally P^{16}. This is especially fast on a GPU (graphics processing unit).

```
# finds stationary probabilities of a Markov chain using matrix powers

altfindpi <- function(p,k) {
   niters <- ceiling(log2(k))
   prd <- p
   for (i in 1:niters) {
      prd <- prd %*% prd
   }
   colMeans(prd)
}
```

This approach has the advantage of being easy to parallelize, unlike matrix inversion.

5.4 Example: 3-Heads-in-a-Row Game

How about the following game? We keep tossing a coin until we get three consecutive heads. What is the expected value of the number of tosses we need?

We can model this as a Markov chain with states 0, 1, 2 and 3, where state i means that we have accumulated i consecutive heads so far. If we simply stop playing the game when we reach state 3, that state would be known as an **absorbing state**, one that we never leave.

We could proceed on this basis, but to keep things elementary, let's just model the game as being played repeatedly, as in the die game above. You'll see that that will still allow us to answer the original question. Note that now that we are taking that approach, it will suffice to have just three states, 0, 1 and 2; there is no state 3, because as soon as we win, we immediately start a new game, in state 0.

5.4.1 Markov Analysis

Clearly we have transition probabilities such as p_{01}, p_{12}, p_{10} and so on all equal to 1/2. Note from state 2 we can only go to state 0, so $p_{20} = 1$.

Here's the code below. Of course, since R subscripts start at 1 instead of 0, we must recode our states as 1, 2 an 3.

```
p <- matrix(rep(0,9),nrow=3)
p[1,1] <- 0.5
p[1,2] <- 0.5
p[2,3] <- 0.5
p[2,1] <- 0.5
p[3,1] <- 1
findpi1(p)
```

It turns out that

$$\pi = (0.5714286, 0.2857143, 0.1428571) \tag{5.18}$$

So, in the long run, about 57.1% of our tosses will be done while in state 0, 28.6% while in state 1, and 14.3% in state 2.

Now, look at that latter figure. Of the tosses we do while in state 2, half will be heads, so half will be wins. In other words, about 0.071 of our tosses will be wins. And THAT figure answers our original question, through the following reasoning:

Think of, say, 10000 tosses. There will be about 710 wins sprinkled among those 10000 tosses. Thus the average number of tosses between wins will be about 10000/710 = 14.1. In other words, the expected time until we get three consecutive heads is about 14.1 tosses.

5.4.2 Back to the word "Stationary"

Let's look at the term *stationary distribution* in the context of this game.

At time 0, we haven't done any coin flips, so $X_0 = 0$. But consider a modified version of the game, in which at time 0 you are given "credit" or a "head start. For instance, you might be allowed to start in state 1, meaning that you already have one consecutive head, even though you actually haven't done any flips.

Now, suppose your head start is random, so that you start in state 0, 1 or 2 with certain probabilities. And suppose those probabilities are given by π! In other words, you start in state 0, 1 or 2, with probabilities 0.57, 0.29 and 0.14, as in (5.18). What, then, will be the distribution of X_1?

$$
\begin{aligned}
p_{X_1}(i) &= P(X_1 = i) && (5.19)\\
&= \sum_{j=0}^{2} P(X_0 = j)\ P(X_1 = i\ |X_0 = j) && (5.20)\\
&= \sum_{j=0}^{2} \pi_j\ p(j,i) && (5.21)
\end{aligned}
$$

This is exactly what we saw in (5.12) and the equations leading up to it. So we know that (5.21) will work out to π_i. In other words, $P(X_1 = 0) = 0.57$ and so on. (Work it out numerically if you are in doubt.) So, we see that

> If X_0 has distribution π, then the same will be true for X_1. We say that the distribution of the chain is *stationary*.

5.5 A Modified Notebook Analysis

Our previous notebook analysis (and most of our future ones, other than for Markov chains), relied on imagining performing many independent replications of the same experiment.

5.5.1 A Markov-Chain Notebook

Consider Table 2.3, for instance. There our experiment was to watch the network during epochs 1 and 2. So, on the first line of the notebook, we would watch the network during epochs 1 and 2 and record the result. On the second line, we watch a new, independent replication of the network during epochs 1 and 2, and record the results.

But instead of imagining a notebook recording infinitely many replications of the two epochs, we could imagine watching just *one* replication but over infinitely many epochs. We'd watch the network during epoch 1, epoch 2, epoch 3 and so on. Now one line of the notebook would record one epoch.

For general Markov chains, each line would record one time step. We would have columns of the notebook labeled n and X_n. The reason this approach would be natural is (5.4). In that context, π_i would be the long-run proportion of notebook lines in which $X_n = i$.

5.5.2 Example: 3-Heads-in-a-Row Game

For instance, consider the 3-heads-in-a-row game. Then (5.18) says that about 57% of the notebook lines would have a 0 in the X_n column, with about 29% and 14% of the lines showing 1 and 2, respectively.

Moreover, suppose we also have a notebook column labeled *Win*, with an entry Yes in a certain line meaning, yes, that coin flip resulted in a win, with a No entry meaning no win. Then the mean time until a win, which we found to be 14.1 above, would be described in notebook terms as the long-run average number of lines between Yes entries in the *Win* column.

5.6 Simulation of Markov Chains

Following up on Section 5.5, recall that our previous simulations have basically put into code form our notebook concept. Our simulations up until now have been based on the definition of probability, which we had way back in Section 2.3. Our simulation code modeled the notebook independent replications notion. We can do a similar thing now, based on the ideas in Section 5.5.

In a time series kind of situation such as Markov chains, since we are interested in long-run behavior in the sense of time, our simulation is based on (5.5). In other words, we simulate the evolution of X_n as n increases, and take long-run averages of whatever is of interest.

Here is simulation code for the example in Section 5.4, caculating the approximate value of the long-run time between wins (found to be about 14.1 by mathematical means above):

```
# simulation of 3-in-a-row coin toss gaem

threeinrow <- function(ntimesteps) {
   consec <- 0  # number of consec Hs
   nwins <- 0  # number of wins
   wintimes <- 0  # total of times to win
   # startplay will be the time at which we started the current game
   startplay <- 0  # time step 0
   for (i in 1:ntimesteps) {
      if (toss() == 'H') {
         consec <- consec + 1
```

```
         if (consec == 3) {
            nwins <- nwins + 1
            wintimes <-
               wintimes + i - startplay
            consec <- 0
            startplay <- i
         }
      } else consec <- 0
   }
   wintimes / nwins
}

toss <- function()
   if (runif(1) < 0.5) 'H' else 'T'
```

l

5.7 Example: ALOHA

Consider our old friend, the ALOHA network model. (You may wish to review the statement of the model in Section 2.5 before continuing.) The key point in that system is that it was "memoryless," in that the probability of what happens at time k+1 depends only on the state of the system at time k.

For instance, consider what might happen at time 6 if $X_5 = 2$. Recall that the latter means that at the end of epoch 5, both of our two network nodes were active. The possibilities for X_6 are then

- X_6 will be 2 again, with probability $p^2 + (1-p)^2$
- X_6 will be 1, with probability $2p(1-p)$

The central point here is that the past history of the system—i.e. the values of $X_1, X_2, X_3,$ and X_4—don't have any impact. We can state that precisely:

> The quantity
>
> $$P(X_6 = j | X_1 = i_1, X_2 = i_2, X_3 = i_3, X_4 = i_4, X_5 = i) \tag{5.22}$$
>
> does not depend on $i_m, m = 1, ..., 4$. Thus we can write (5.22) simply as $P(X_6 = j | X_5 = i)$.

Furthermore, that probability is the same as $P(X_9 = j|X_8 = i)$ and in general $P(X_{k+1} = j|X_k = i)$. So, we do have a Markov chain.

Since this is a three-state chain, the p_{ij} form a 3x3 matrix:

$$P = \begin{pmatrix} (1-q)^2 + 2q(1-q)p & 2q(1-q)(1-p) + 2q^2p(1-p) & q^2[p^2 + (1-p)^2] \\ (1-q)p & 2qp(1-p) + (1-q)(1-p) & q[p^2 + (1-p)^2] \\ 0 & 2p(1-p) & p^2 + (1-p)^2 \end{pmatrix} \tag{5.23}$$

For instance, the element in row 0, column 2, p_{02}, is $q^2[p^2 + (1-p)^2]$, reflecting the fact that to go from state 0 to state 2 would require that both inactive nodes become active (which has probability q^2, and then either both try to send or both refrain from sending (probability $p^2 + (1-p)^2$.

For the ALOHA example here, with p = 0.4 and q = 0.3, the solution is $\pi_0 = 0.37$, $\pi_1 = 0.45$ and $\pi_2 = 0.18$.

So we know that in the long run, about 47% of the epochs will have no active nodes, 43% will have one, and 10% will have two. From this we see that the long-run average number of active nodes is

$$0 \cdot 0.37 + 1 \cdot 0.45 + 2 \cdot 0.18 = 0.81 \tag{5.24}$$

By the way, note that every row in a transition matrix must sum to 1. (The probability that we go from state i to *somewhere* is 1, after all, so we must have $\sum_j p_{ij} = 1$.) That implies that we can save some work in writing R code; the last column must be 1 minus the others. In our example above, we would write

```
transmat <- matrix(rep(0,9),nrow=3)
p1 <- 1 - p
q1 <- 1 - q
transmat[1,1] <- q1^2 + 2 * q * q1 *p
transmat[1,2] <- 2 * q * q1 * p1 + 2 * q^2 * p * p1
transmat[2,1] <- q1 * p
transmat[2,2] <- 2 * q * p * p1 + q1 * p1
transmat[3,1] <- 0
transmat[3,2] <- 2 * p * p1
transmat[,3] <- 1 - transmat[,1] - transmat[,2]
findpi1(transmat)
```

Note the vectorized addition and recycling (Section 2.14.2).

5.8 Example: Bus Ridership Problem

Consider the bus ridership problem in Section 2.11. Make the same assumptions now, but add a new one: There is a maximum capacity of 20 passengers on the bus.

The random variables L_i, i = 1,2,3,... form a Markov chain. Let's look at some of the transition probabilities:

$$p_{00} = 0.5 \tag{5.25}$$

$$p_{01} = 0.4 \tag{5.26}$$

$$p_{11} = (1 - 0.2) \cdot 0.5 + 0.2 \cdot 0.4 \tag{5.27}$$

$$p_{20} = (0.2)^2 (0.5) = 0.02 \tag{5.28}$$

$$p_{20,20} = (0.8)^{20} (0.5 + 0.4 + 0.1) + \binom{20}{1} (0.2)^1 (0.8)^{20-1} (0.4 + 0.1) + \binom{20}{2} (0.2)^2 (0.8)^{18} (0.1) \tag{5.29}$$

(Note that for clarity, there is a comma in $p_{20,20}$, as p_{2020} would be confusing and in some other examples even ambiguous. A comma is not necessary in p_{11}, since there must be two subscripts; the 11 here can't be eleven.)

After finding the π vector as above, we can find quantities such as the long-run average number of passengers on the bus,

$$\sum_{i=0}^{20} \pi_i i \tag{5.30}$$

We can also compute the long-run average number of would-be passengers who fail to board the bus. Denote by A_i denote the number of passengers on the bus as it *arrives* at stop i. The key point is that since $A_i = L_{i-1}$, then (5.4) and (5.5) will give the same result, no matter whether we look at the L_j chain or the A_j chain.

Now, armed with that knowledge, let D_j denote the number of disappointed people at stop j. Then

$$ED_j = 1 \cdot P(D_j = 1) + 2 \cdot P(D_j = 2). \tag{5.31}$$

That latter probability, for instance, is

$$\begin{aligned} P(D_j = 2) &= P(A_j = 20 \text{ and } B_j = 2 \text{ and } G_j = 0) && (5.32) \\ &= P(A_j = 20)\ P(B_j = 2)\ P(G_j = 0 \mid A_j = 20) && (5.33) \\ &= P(A_j = 20) \cdot 0.1 \cdot 0.8^{20} && (5.34) \end{aligned}$$

where as before B_j is the number who wish to board at stop j, and G_j is the number who get off the bus at that stop. Using the same reasoning, one can find $P(D_j = 1)$. (A number of cases to consider, left as an exercise for the reader.)

Taking the limits as $j \to \infty$, we have the long-run average number of disappointed customers on the left, and on the right, the term $P(A_j = 20)$ goes to π_{20}, which we have and thus can obtain the value regarding disappointed customers.

Let's find the long-run average number of customers who alight from the bus. This can be done by considering all the various cases, but (4.68) really shortens our work:

$$EG_n = \sum_{i=0}^{20} P(A_n = i) E(G_n | A_n = i) \tag{5.35}$$

Given $A_n = i$, G_n has a binomial distribution with i trials and success probability 0.2, so

$$E(G_n | A_n = i) = i \cdot 0.2 \tag{5.36}$$

So, the right-hand side of 5.35 converges to

$$\sum_{i=0}^{20} \pi_i i \cdot 0.2 \tag{5.37}$$

In other words, the long-run average number alighting is 0.2 times (5.30).

5.9 Example: an Inventory Model

Consider the following simple inventory model. A store has 1 or 2 customers for a certain item each day, with probabilities v and w (v+w = 1). Each customer is allowed to buy only 1 item.

When the stock on hand reaches 0 on a day, it is replenished to r items immediately after the store closes that day.

If at the start of a day the stock is only 1 item and 2 customers wish to buy the item, only one customer will complete the purchase, and the other customer will leave emptyhanded.

Let X_n be the stock on hand at the end of day n (*after* replenishment, if any). Then $X_1, X_2, ...$ form a Markov chain, with state space 1,2,...,r.

The transition probabilities are easy to find. Take p_{21}, for instance. If there is a stock of 2 items at the end of one day, what is the (conditional) probability that there is only 1 item at the end of the next day? Well, for this to happen, there would have to be just 1 customer coming in, not 2, and that has probability v. So, $p_{21} = v$. The same reasoning shows that $p_{2r} = w$.

Let's write a function **inventory(v,w,r)** that returns the π vector for this Markov chain. It will call **findpi1()**, similarly to the two code snippets on page 104. For convenience, let's assume **r** is at least 3.[6]

```
inventory <- function(v,w,r) {
   tm <- matrix(rep(0,r^2),nrow=r)
   for (i in 3:r) {
      tm[i,i-1] <- v
      tm[i,i-2] <- w
   }
   tm[2,1] <- v
   tm[2,r] <- w
   tm[1,r] <- 1
   findpi1(tm)
}
```

5.10 Expected Hitting Times

Consider an n-state Markov chain that is *irreducible*, meaning that it is possible to get from any state i to any other state j in some number of steps. Define the random variable T_{ij} to be the time

[6] If **r** is 2, then the expression 3:2 in the code evaluates to the vector (3,2), which would not be what we want in this case.

needed to go from state i to state j. (Note that T_{ii} is NOT 0, though it can be 1 if $p_{ii} > 0$.)

In many applications, we are interested in the mean time to reach a certain state, given that we start at some specified state. In other words, we wish to calculate ET_{ij}.

The material in Section 4.15 is useful here. In (4.68), take $V = T_{ij}$ and take U to be the first state we visit after leaving state i (which could be i again). Then

$$E(T_{ij}) = \sum_k p_{ik} E(T_{ij} \mid U = k), \quad 1 \leq i, j \leq n \tag{5.38}$$

Consider what happens if $U = k \neq j$. Then we will have already used up 1 time step, and *still* will need an average of ET_{kj} more steps to get to j. In other words,

$$E(T_{ij} \mid U = k \neq j) = 1 + ET_{kj} \tag{5.39}$$

Did you notice the Markov property being invoked?

On the other hand, the case $U = j$ is simple:

$$E(T_{ij} \mid U = j) = 1 \tag{5.40}$$

This then implies that

$$E(T_{ij}) = 1 + \sum_{k \neq j} p_{ik} E(T_{kj}), \quad 1 \leq i, j \leq n \tag{5.41}$$

We'll focus on the case $j = n$, i.e. look at how long it takes to get to state n. (We could of course do the same analysis for any other destination state.) Let η_i denote $E(T_{in})$, and define $\eta = (\eta_1, \eta_2, ..., \eta_{n-1})'$. (Note that η has only $n - 1$ components!) So,

$$\eta_i = 1 + \sum_{k<n} p_{ik} \eta_k, \quad 1 \leq i \leq n - 1 \tag{5.42}$$

Equation (5.42) is a system of linear equations, which we can write in matrix form. Then the code to solve the system is (remember, we have only an $(n - 1) \times (n - 1)$ system)

```
findeta <- function(p) {
   n <- nrow(p)
   q <- diag(n) - p
   q <- q[1:(n-1),1:(n-1)]
```

```
    ones <- rep(1,n-1)
    solve(q,ones)
}
```

Chapter 6

Continuous Probability Models

There are other types of random variables besides the discrete ones you studied in Chapter 3. This chapter will cover another major class, *continuous random variables*, which form the heart of statistics and are used extensively in applied probability as well. It is for such random variables that the calculus prerequisite for this book is needed.

6.1 Running Example: a Random Dart

Imagine that we throw a dart at random at the interval (0,1). Let D denote the spot we hit. By "at random" we mean that all subintervals of equal length are equally likely to get hit. For instance, the probability of the dart landing in (0.7,0.8) is the same as for (0.2,0.3), (0.537,0.637) and so on.

Because of that randomness,

$$P(u \leq D \leq v) = v - u \tag{6.1}$$

for any case of $0 \leq u < v \leq 1$.

We call D a **continuous** random variable, because its support is a continuum of points, in this case, the entire interval (0,1).

6.2 Individual Values Now Have Probability Zero

The first crucial point to note is that

$$P(D = c) = 0 \tag{6.2}$$

for any individual point c. This may seem counterintuitive, but it can be seen in a couple of ways:

- Take for example the case c = 0.3. Then

  $$P(D = 0.3) \leq P(0.29 \leq D \leq 0.31) = 0.02 \tag{6.3}$$

  the last equality coming from (6.1).

  So, $P(D = 0.3) \leq 0.02$. But we can replace 0.29 and 0.31 in (6.3) by 0.299 and 0.301, say, and get $P(D = 0.3) \leq 0.002$. So, $P(D = 0.3)$ must be smaller than any positive number, and thus it's actually 0.

- Reason that there are infinitely many points, and if they all had some nonzero probability w, say, then the probabilities would sum to infinity instead of to 1; thus they must have probability 0.

Similarly, one will see that (6.2) will hold for any continuous random variable.

Remember, we have been looking at probability as being the long-run fraction of the time an event occurs, in infinitely many repetitions of our experiment—the "notebook" view. So (6.2) doesn't say that D = c can't occur; it merely says that it happens so rarely that the long-run fraction of occurrence is 0.

6.3 But Now We Have a Problem

But Equation (6.2) presents a problem. In the case of discrete random variables M, we defined their distribution via their probability mass function, p_M. Recall that Section 3.13 defined this as a list of the values M takes on, together with their probabilities. But that would be impossible in the continuous case—all the probabilities of individual values here are 0.

So our goal will be to develop another kind of function, which is similar to probability mass functions in spirit, but circumvents the problem of individual values having probability 0. To do this, we first must define another key function:

6.3.1 Our Way Out of the Problem: Cumulative Distribution Functions

Definition 9 *For any random variable W (including discrete ones), its* **cumulative distribution function** *(cdf), F_W, is defined by*

$$F_W(t) = P(W \leq t), -\infty < t < \infty \tag{6.4}$$

(Please keep in mind the notation. It is customary to use capital F to denote a cdf, with a subscript consisting of the name of the random variable.)

What is t here? It's simply an argument to a function. The function here has domain $(-\infty, \infty)$, and we must thus define that function for every value of t. This is a simple point, but a crucial one.

For an example of a cdf, consider our "random dart" example above. We know that, for example for t = 0.23,

$$F_D(0.23) = P(D \leq 0.23) = P(0 \leq D \leq 0.23) = 0.23 \tag{6.5}$$

Also,

$$F_D(-10.23) = P(D \leq -10.23) = 0 \tag{6.6}$$

and

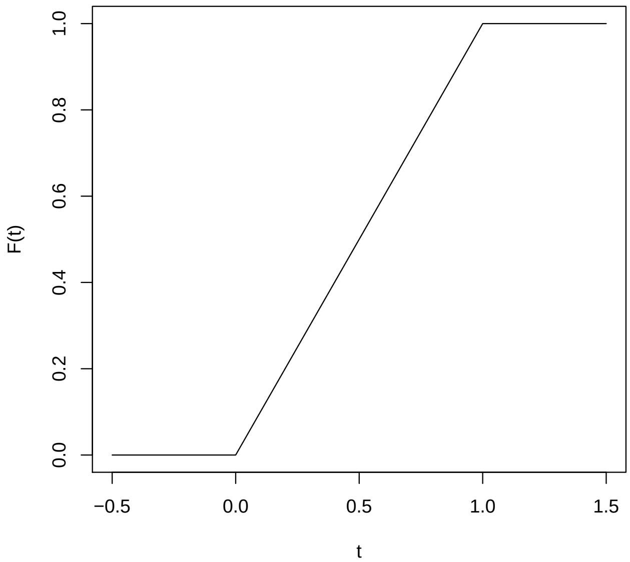
$$F_D(10.23) = P(D \leq 10.23) = 1 \tag{6.7}$$

Note that *the fact that D can never be equal to 10.23 or anywhere near it is irrelevant.* $F_D(t)$ is defined for *all* t in $(-\infty, \infty)$, including 10.23! The definition of $F_D(10.23)$ is $P(D \leq 10.23)$), and that probability is 1! Yes, D is *always* less than or equal to 10.23, right?

In general for our dart,

$$F_D(t) = \begin{cases} 0, & \text{if } t \leq 0 \\ t, & \text{if } 0 < t < 1 \\ 1, & \text{if } t \geq 1 \end{cases} \tag{6.8}$$

Here is the graph of F_D:



The cdf of a discrete random variable is defined as in Equation (6.4) too. For example, say Z is the number of heads we get from two tosses of a coin. Then

$$F_Z(t) = \begin{cases} 0, & \text{if } t < 0 \\ 0.25, & \text{if } 0 \leq t < 1 \\ 0.75, & \text{if } 1 \leq t < 2 \\ 1, & \text{if } t \geq 2 \end{cases} \tag{6.9}$$

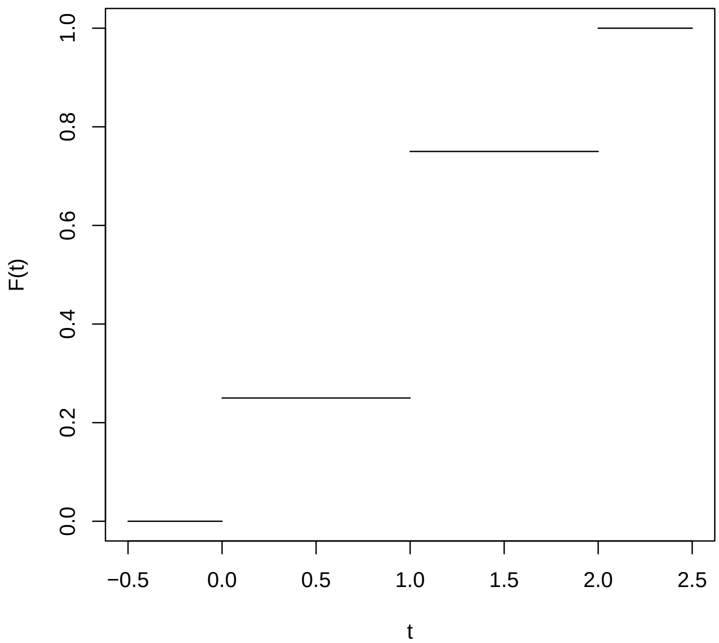
For instance,

$$\begin{aligned} F_Z(1.2) &= P(Z \leq 1.2) && (6.10) \\ &= P(Z = 0 \text{ or } Z = 1) && (6.11) \\ &= 0.25 + 0.50 && (6.12) \\ &= 0.75 && (6.13) \end{aligned}$$

Note that (6.11) is simply a matter of asking our famous question, "How can it happen?" Here we are asking how it can happen that $Z \leq 1.2$. The answer is simple: That can happen if Z is 0 or 1. *The fact that Z cannot equal 1.2 is irrelevant.*

(6.12) uses the fact that Z has a binomial distribution with n = 2 and p = 0.5.

F_Z is graphed below.



The fact that one cannot get a noninteger number of heads is what makes the cdf of Z flat between consecutive integers.

In the graphs you see that F_D in (6.8) is continuous while F_Z in (6.9) has jumps. This is another reason we call random variables such as D **continuous random variables**.

At this level of study of probability, random variables are either discrete or continuous. But some exist that are neither. We won't see any random variables from the "neither" case here, and they occur rather rarely in practice.

Armed with cdfs, let's turn to the original goal, which was to find something for continuous random variables that is similar in spirit to probability mass functions for discrete random variables.

6.3.2 Density Functions

Intuition is key here. Make SURE you develop a good intuitive understanding of density functions, as it is vital in being able to apply probability well. We will use it a lot in our course.

(The reader may wish to review pmfs in Section 3.13.)

Think as follows. From (6.4) we can see that for a discrete random variable, its cdf can be calculated by summing its pmf. Recall that in the continuous world, we integrate instead of sum. So, our continuous-case analog of the pmf should be something that integrates to the cdf. That of course is the derivative of the cdf, which is called the **density**:

Definition 10 *Consider a continuous random variable W. Define*

$$f_W(t) = \frac{d}{dt} F_W(t), -\infty < t < \infty \tag{6.14}$$

wherever the derivative exists. The function f_W is called the **probability density function** *(pdf), or just the* **density** *of W.*

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

But what *is* a density function? First and foremost, it is a tool for finding probabilities involving continuous random variables:

6.3.3 Properties of Densities

Equation (6.14) implies

Property A:

$$P(a < W \le b) = F_W(b) - F_W(a) \tag{6.15}$$

$$= \int_a^b f_W(t) \, dt \tag{6.16}$$

Where does (6.15) come from? Well, $F_W(b)$ is all the probability accumulated from $-\infty$ to b, while $F_W(a)$ is all the probability accumulated from $-\infty$ to a. The difference is the probability that X is *between* a and b.

(6.16) is just the Fundamental Theorem of Calculus: Integrate the derivative of a function, and you get the original function back again.

Since P(W = c) = 0 for any single point c, Property A also means:

Property B:

$$P(a < W \le b) = P(a \le W \le b) = P(a \le W < b) = P(a < W < b) = \int_a^b f_W(t)\ dt \tag{6.17}$$

This in turn implies:

Property C:

$$\int_{-\infty}^{\infty} f_W(t)\ dt = 1 \tag{6.18}$$

Note that in the above integral, $f_W(t)$ will be 0 in various ranges of t corresponding to values W cannot take on. For the dart example, for instance, this will be the case for $t < 0$ and $t > 1$.

Any nonnegative function that integrates to 1 is a density. A density could be increasing, decreasing or mixed. Note too that a density can have values larger than 1 at some points, even though it must integrate to 1.

6.3.4 Intuitive Meaning of Densities

Suppose we have some continuous random variable X, with density f_X, graphed in Figure 6.1.

Let's think about probabilities of the form

$$P(s - 0.1 < X < s + 0.1) \tag{6.19}$$

Let's first consider the case of s = 1.3.

The rectangular strip in the picture should remind you of your early days in calculus. What the picture says is that the area under f_X from 1.2 to 1.4 (i.e. 1.3 ± 0.1) is approximately equal to the area of the rectangle. In other words,

$$2(0.1) f_X(1.3) \approx \int_{1.2}^{1.4} f_X(t)\ dt \tag{6.20}$$

But from our Properties above, we can write this as

$$P(1.2 < X < 1.4) \approx 2(0.1) f_X(1.3) \tag{6.21}$$

Similarly, for s = 0.4,

$$P(0.3 < X < 0.5) \approx 2(0.1) f_X(0.4) \tag{6.22}$$

and in general

$$P(s - 0.1 < X < s + 0.1) \approx 2(0.1) f_X(s) \tag{6.23}$$

This reasoning shows that:

> Regions in the number line (X-axis in the picture) with low density have low probabilities while regions with high density have high probabilities.

So, although densities themselves are not probabilities, they do tell us which regions will occur often or rarely. For the random variable X in our picture, there will be many lines in the notebook in which X is near 1.3, but many fewer in which X is near 0.4.

6.3.5 Expected Values

What about E(W)? Recall that if W were discrete, we'd have

$$E(W) = \sum_c c p_W(c) \tag{6.24}$$

where the sum ranges overall all values c that W can take on. If for example W is the number of dots we get in rolling two dice, c will range over the values 2,3,...,12.

So, the analog for continuous W is:

Property D:

$$E(W) = \int_t t f_W(t)\ dt \tag{6.25}$$

where here t ranges over the values W can take on, such as the interval (0,1) in the dart case. Again, we can also write this as

$$E(W) = \int_{-\infty}^{\infty} t f_W(t)\ dt \tag{6.26}$$

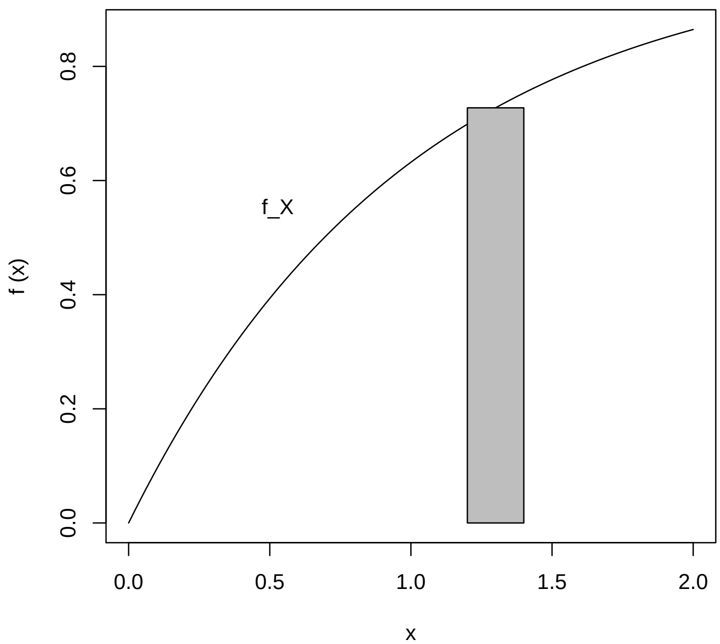


Figure 6.1: Approximation of Probability by a Rectangle

in view of the previous comment that $f_W(t)$ might be 0 for various ranges of t.

And of course,

$$E(W^2) = \int_t t^2 f_W(t) \ dt \tag{6.27}$$

and in general, similarly to (3.36):

Property E:

$$E[g(W)] = \int_t g(t) f_W(t) \ dt \tag{6.28}$$

Most of the properties of expected value and variance stated previously for discrete random variables hold for continuous ones too:

Property F:

Equations (3.19), (3.21), (3.25), (3.41) and (3.49) still hold in the continuous case.

6.4 A First Example

Consider the density function equal to 2t/15 on the interval (1,4), 0 elsewhere. Say X has this density. Here are some computations we can do:

$$EX = \int_1^4 t \cdot 2t/15 \ dt = 2.8 \tag{6.29}$$

$$P(X > 2.5) = \int_{2.5}^4 2t/15 \ dt = 0.65 \tag{6.30}$$

$$F_X(s) = \int_1^s 2t/15 \ dt = \frac{s^2-1}{15} \quad \text{for s in (1,4) (cdf is 0 for t < 1, and 1 for t > 4)} \tag{6.31}$$

$$\begin{aligned} Var(X) &= E(X^2) - (EX)^2 \quad \text{(from (3.41)} && (6.32) \\ &= \int_1^4 t^2 2t/15 \ dt - 2.8^2 \quad \text{(from (6.29))} && (6.33) \\ &= 0.66 && (6.34) \end{aligned}$$

Suppose L is the lifetime of a light bulb (say in years), with the density that X has above. Let's find some quantities in that context:

Proportion of bulbs with lifetime less than the mean lifetime:

$$P(L < 2.8) = \int_1^{2.8} 2t/15 \; dt = (2.8^2 - 1)/15 \tag{6.35}$$

Mean of 1/L:

$$E(1/L) = \int_1^4 \frac{1}{t} \cdot 2t/15 \; dt = \frac{2}{5} \tag{6.36}$$

In testing many bulbs, mean number of bulbs that it takes to find two that have lifetimes longer than 2.5:

Use (4.38) with r = 2 and p = 0.65.

6.5 The Notion of *Support* in the Continuous Case

Recall from Section 3.2 that the *support* of a discrete distribution is its "domain." If for instance X is the number of heads I get from 3 tosses of a coin, X can only take on the values 0, 1, 2 and 3. We say that that set is the support of this distribution; 8, for example, is not in the support.

The notion extends to continuous random variables. In Section 6.4, the support of the density there is the interval (1,4).

6.6 Famous Parametric Families of Continuous Distributions

6.6.1 The Uniform Distributions

6.6.1.1 Density and Properties

In our dart example, we can imagine throwing the dart at the interval (q,r) (so this will be a two-parameter family). Then to be a uniform distribution, i.e. with all the points being "equally likely," the density must be constant in that interval. But it also must integrate to 1 [see (6.18).

So, that constant must be 1 divided by the length of the interval:

$$f_D(t) = \frac{1}{r-q} \tag{6.37}$$

for t in (q,r), 0 elsewhere.

It easily shown that $E(D) = \frac{q+r}{2}$ and $Var(D) = \frac{1}{12}(r-q)^2$.

The notation for this family is U(q,r).

6.6.1.2 R Functions

Relevant functions for a uniformly distributed random variable X on (r,s) are:

- **dunif(x,r,s)**, to find $f_X(x)$
- **punif(q,r,s)**, to find $P(X \leq q)$
- **qunif(q,r,s)**, to find c such that $P(X \leq c) = q$
- **runif(n,r,s)**, to generate n independent values of X

As with most such distribution-related functions in R, **x** and **q** can be vectors, so that **punif()** for instance can be used to find the cdf values at multiple points.

6.6.1.3 Example: Modeling of Disk Performance

Uniform distributions are often used to model computer disk requests. Recall that a disk consists of a large number of concentric rings, called **tracks**. When a program issues a request to read or write a file, the **read/write head** must be positioned above the track of the first part of the file. This move, which is called a **seek**, can be a significant factor in disk performance in large systems, e.g. a database for a bank.

If the number of tracks is large, the position of the read/write head, which I'll denote as X, is like a continuous random variable, and often this position is modeled by a uniform distribution. This situation may hold just before a defragmentation operation. After that operation, the files tend to be bunched together in the central tracks of the disk, so as to reduce seek time, and X will not have a uniform distribution anymore.

Each track consists of a certain number of **sectors** of a given size, say 512 bytes each. Once the read/write head reaches the proper track, we must wait for the desired sector to rotate around and pass under the read/write head. It should be clear that a uniform distribution is a good model for this **rotational delay**.

For example, suppose in modeling disk performance, we describe the position X of the read/write head as a number between 0 and 1, representing the innermost and outermost tracks, respectively. Say we assume X has a uniform distribution on (0,1), as discussed above. Consider two consecutive positions (i.e. due to two consecutive seeks), X_1 and X_2, which we'll assume are independent.[1] Let's find $Var(X_1 + X_2)$.

We know from Section 6.6.1.1 that the variance of a U(0,1) distribution is 1/12. Then by independence,

$$Var(X_1 + X_2) = 1/12 + 1/12 = 1/6 \tag{6.38}$$

6.6.1.4 Example: Modeling of Denial-of-Service Attack

In one facet of computer security, it has been found that a uniform distribution is actually a warning of trouble, a possible indication of a **denial-of-service attack**. Here the attacker tries to monopolize, say, a Web server, by inundating it with service requests. According to the research of David Marchette,[2] attackers choose uniformly distributed false IP addresses, a pattern not normally seen at servers.

6.6.2 The Normal (Gaussian) Family of Continuous Distributions

These are the famous "bell-shaped curves," so called because their densities have that shape.[3]

6.6.2.1 Density and Properties

Density and Parameters:

[1] NOT a good assumption for consecutive seeks; think about why.

[2] *Statistical Methods for Network and Computer Security*, David J. Marchette, Naval Surface Warfare Center, `rion.math.iastate.edu/IA/2003/foils/marchette.pdf`.

[3] *All that glitters is not gold"*—Shakespeare

Note that other parametric families, notably the Cauchy, also have bell shapes. The difference lies in the rate at which the tails of the distribution go to 0. However, due to the Central Limit Theorem, to be presented below, the normal family is of prime interest.

The density for a normal distribution is

$$f_W(t) = \frac{1}{\sqrt{2\pi}\sigma} \; e^{-0.5\left(\frac{t-\mu}{\sigma}\right)^2}, -\infty < t < \infty \tag{6.39}$$

Again, this is a two-parameter family, indexed by the parameters μ and σ, which turn out to be the mean[4] and standard deviation μ and σ, The notation for it is $N(\mu, \sigma^2)$ (it is customary to state the variance σ^2 rather than the standard deviation).

The normal family is so important that we have a special chapter on it, Chapter 7.

6.6.3 The Exponential Family of Distributions

Please note: We have been talking here of parametric families of distributions, and in this section will introduce one of the most famous, the family of exponential distributions. This should not be confused, though, with the term *exponential family* that arises in mathematical statistics, which includes exponential distributions but is much broader.

6.6.3.1 Density and Properties

The densities in this family have the form

$$f_W(t) = \lambda e^{-\lambda t}, 0 < t < \infty \tag{6.40}$$

This is a one-parameter family of distributions.

After integration, one finds that $E(W) = \frac{1}{\lambda}$ and $Var(W) = \frac{1}{\lambda^2}$. You might wonder why it is customary to index the family via λ rather than $1/\lambda$ (see (6.40)), since the latter is the mean. But this is actually quite natural, for reasons discussed in Section 8.1.

6.6.3.2 R Functions

Relevant functions for a uniformly distributed random variable X with parameter λ are

- **dexp(x,lambda)**, to find $f_X(x)$
- **pexp(q,lambda)**, to find $P(X \leq q)$

[4] Remember, this is a synonym for expected value.

- **qexp(q,lambda)**, to find c such that $P(X \leq c) = q$
- **rexp(n,lambda)**, to generate n independent values of X

6.6.3.3 Example: Refunds on Failed Components

Suppose a manufacturer of some electronic component finds that its lifetime L is exponentially distributed with mean 10000 hours. They give a refund if the item fails before 500 hours. Let M be the number of items they have sold, up to and including the one on which they make the first refund. Let's find EM and $Var(M)$.

First, notice that M has a geometric distribution! It is the number of independent trials until the first success, where a "trial" is one component, "success" (no value judgment, remember) is giving a refund, and the success probability is

$$P(L < 500) = \int_0^{500} 0.0001 e^{-0.0001t} \; dt = 0.05 \tag{6.41}$$

Then plug p = 0.05 into (4.11) and (4.12).

6.6.3.4 Example: Garage Parking Fees

A certain public parking garage charges parking fees of \$1.50 for the first hour, and \$1 per hour after that. (It is assumed here for simplicity that the time is prorated within each of those defined periods. The reader should consider how the analysis would change if the garage "rounds up" each partial hour.) Suppose parking times T are exponentially distributed with mean 1.5 hours. Let W denote the total fee paid. Let's find E(W) and Var(W).

The key point is that W is a function of T:

$$W = \begin{cases} 1.5T, & \text{if } T \leq 1 \\ 1.5 + 1 \cdot (T-1) = T + 0.5, & \text{if } T > 1 \end{cases} \tag{6.42}$$

That's good, because we know how to find the expected value of a function of a continuous random variable, from (6.28). Defining g() as in (6.42) above, we have

$$EW = \int_0^\infty g(t) \; \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt = \int_0^1 1.5t \; \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt + \int_1^\infty (t+0.5) \; \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt \tag{6.43}$$

The integration is left to the reader.

Now, what about Var(W)? As is often the case, it's easier to use (3.41), so we need to find $E(W^2)$. The above integration becomes

$$E(W^2) = \int_0^\infty g^2(t)\ f_W(t)dt = \int_0^1 (1.5t)^2\ \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt + \int_1^\infty (t+0.5)^2\ \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt \tag{6.44}$$

After evaluating this, we subtract $(EW)^2$, giving us the variance of W.

6.6.3.5 Importance in Modeling

Many distributions in real life have been found to be approximately exponentially distributed. A famous example is the lifetimes of air conditioners on airplanes. Another famous example is interarrival times, such as customers coming into a bank or messages going out onto a computer network. It is used in software reliability studies too.

One of the reasons why this family is used so widely in probabilistic modeling is that it has several remarkable properties, so many that we have a special chapter for this family, Chapter 8.

6.6.4 The Gamma Family of Distributions

6.6.4.1 Density and Properties

Suppose at time 0 we install a light bulb in a lamp, which burns X_1 amount of time. We immediately install a new bulb then, which burns for time X_2, and so on. Assume the X_i are independent random variables having an exponential distribution with parameter λ.

Let

$$T_r = X_1 + ... + X_r, \quad r = 1, 2, 3, ... \tag{6.45}$$

Note that the random variable T_r is the time of the r^{th} light bulb replacement. T_r is the sum of r independent exponentially distributed random variables with parameter λ. The distribution of T_r is called an **Erlang** distribution. Its density can be shown to be

$$f_{T_r}(t) = \frac{1}{(r-1)!} \lambda^r t^{r-1} e^{-\lambda t},\ t > 0 \tag{6.46}$$

This is a two-parameter family.

Again, it's helpful to think in "notebook" terms. Say r = 8. Then we watch the lamp for the durations of eight lightbulbs, recording T_8, the time at which the eighth burns out. We write that time in the first line of our notebook. Then we watch a new batch of eight bulbs, and write the value of T_8 for those bulbs in the second line of our notebook, and so on. Then after recording a very large number of lines in our notebook, we plot a histogram of all the T_8 values. The point is then that that histogram will look like (6.46).

We can generalize this by allowing r to take noninteger values, by defining a generalization of the factorial function:

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} \, dx \tag{6.47}$$

This is called the gamma function, and it gives us the gamma family of distributions, more general than the Erlang:

$$f_W(t) = \frac{1}{\Gamma(r)} \lambda^r t^{r-1} e^{-\lambda t}, \; t > 0 \tag{6.48}$$

(Note that $\Gamma(r)$ is merely serving as the constant that makes the density integrate to 1.0. It doesn't have meaning of its own.)

This is again a two-parameter family, with r and λ as parameters.

A gamma distribution has mean r/λ and variance r/λ^2. In the case of integer r, this follows from (8.1) and the fact that an exponentially distributed random variable has mean and variance $1/\lambda$ and variance $1/\lambda^2$, and it can be derived in general. Note again that the gamma reduces to the exponential when r = 1.

6.6.4.2 Example: Network Buffer

Suppose in a network context (not our ALOHA example), a node does not transmit until it has accumulated five messages in its buffer. Suppose the times between message arrivals are independent and exponentially distributed with mean 100 milliseconds. Let's find the probability that more than 552 ms will pass before a transmission is made, starting with an empty buffer.

Let X_1 be the time until the first message arrives, X_2 the time from then to the arrival of the second message, and so on. Then the time until we accumulate five messages is $Y = X_1 + ... + X_5$. Then from the definition of the gamma family, we see that Y has a gamma distribution with r =

5 and $\lambda = 0.01$. Then

$$P(Y > 552) = \int_{552}^{\infty} \frac{1}{4!} 0.01^5 t^4 e^{-0.01t} \, dt \tag{6.49}$$

This integral could be evaluated via repeated integration by parts, but let's use R instead:

```
> 1 - pgamma(552,5,0.01)
[1] 0.3544101
```

Note that our parameter r is called **shape** in R, and our λ is **rate**.

Again, there are also **dgamma()**, **qgamma()** and **rgamma()**. From the R man page:

```
Usage:

     dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)
     pgamma(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE,
            log.p = FALSE)
     qgamma(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE,
            log.p = FALSE)
     rgamma(n, shape, rate = 1, scale = 1/rate)
```

6.6.4.3 Importance in Modeling

As seen in (8.1), sums of exponentially distributed random variables often arise in applications. Such sums have gamma distributions.

You may ask what the meaning is of a gamma distribution in the case of noninteger r. There is no particular meaning, but when we have a real data set, we often wish to summarize it by fitting a parametric family to it, meaning that we try to find a member of the family that approximates our data well.

In this regard, the gamma family provides us with densities which rise near t = 0, then gradually decrease to 0 as t becomes large, so the family is useful if our data seem to look like this. Graphs of some gamma densities are shown in Figure 6.2.

As you might guess from the network performance analysis example in Section 6.6.4.2, the gamma family does arise often in the network context, and in queuing analysis in general.

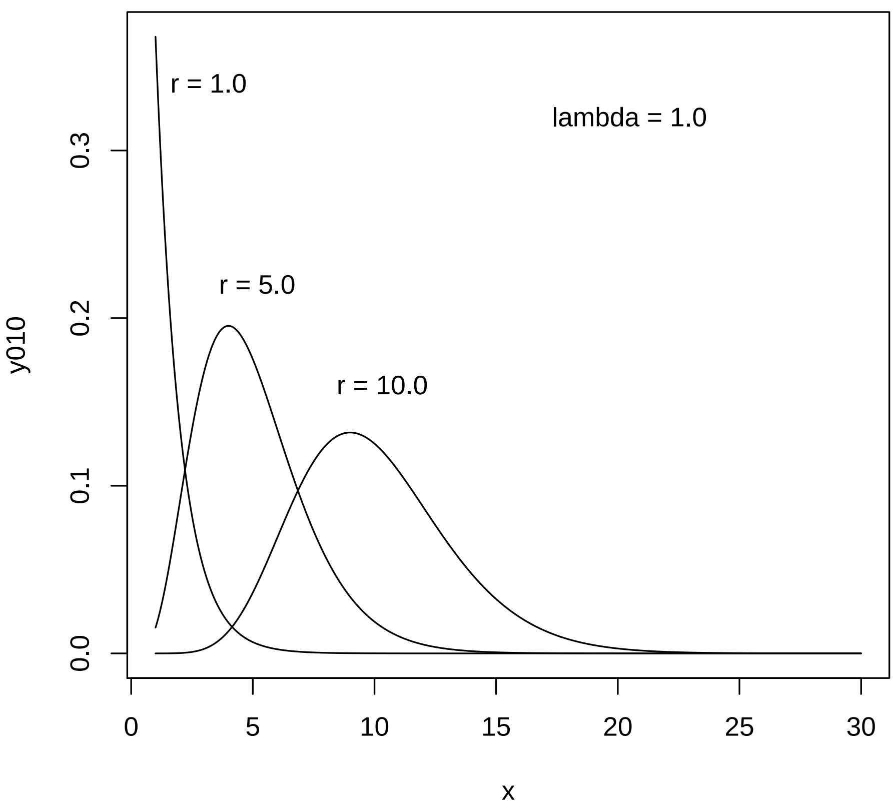


Figure 6.2: Various Gamma Densities

6.6.5 The Beta Family of Distributions

As seen in Figure 6.2, the gamma family is a good choice to consider if our data are nonnegative, with the density having a peak near 0 and then gradually tapering off to the right. What about data in the range (0,1)?

For instance, say trucking company transports many things, including furniture. Let X be the proportion of a truckload that consists of furniture. For instance, if 15% of given truckload is furniture, then X = 0.15. So here we have a distribution with support in (0,1). The beta family provides a very flexible model for this kind of setting, allowing us to model many different concave up or concave down curves.

6.6.5.1 Density Etc.

The densities of the family have the following form:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1} \tag{6.50}$$

There are two parameters, α and β. Figures 6.3 and 6.4 show two possibilities.

The mean and variance are

$$\frac{\alpha}{\alpha+\beta} \tag{6.51}$$

and

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{6.52}$$

Again, there are also **dbeta()**, **qbeta()** and **rbeta()**. From the R man page:

```
Usage:

     dbeta(x, shape1, shape2, ncp = 0, log = FALSE)
     pbeta(q, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
     qbeta(p, shape1, shape2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
     rbeta(n, shape1, shape2, ncp = 0)
```

The graphs mentioned above were generated by running

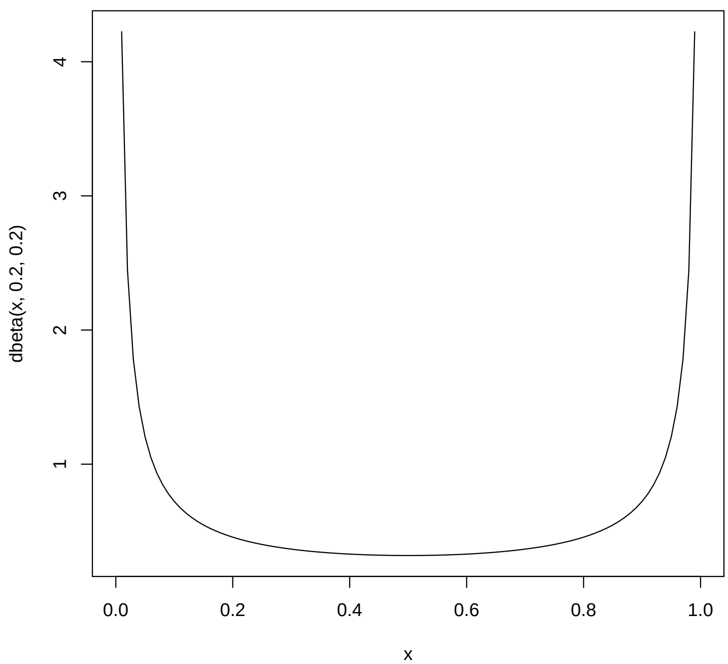


Figure 6.3: Beta Density, $\alpha = 0.2, \beta = 0.2$

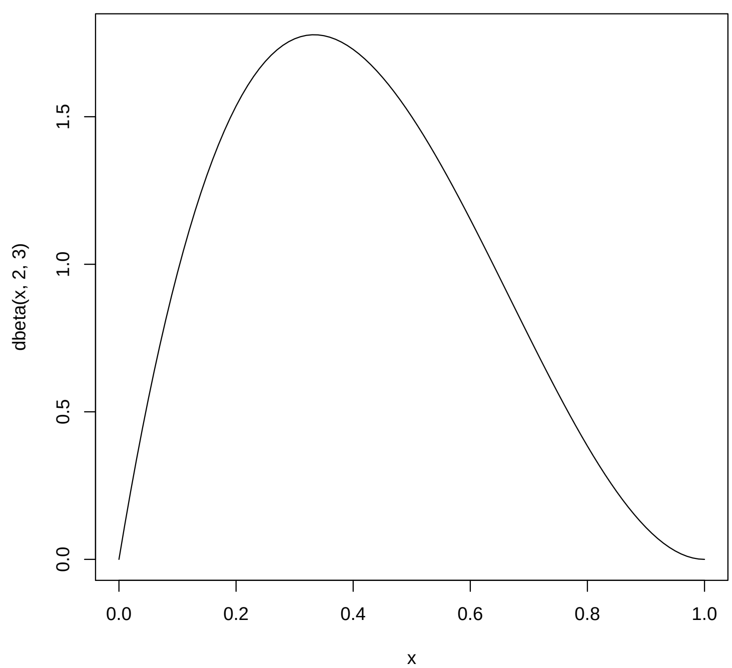


Figure 6.4: Beta Density, $\alpha = 2.0, \beta = 3.0$

```
> curve(dbeta(x,0.2,0.2))
> curve(dbeta(x,2,3)
```

6.6.5.2 Importance in Modeling

As mentioned, the beta family is a natural candidate for modeling a variable having range the interval (0,1).

This family is also popular among **Bayesian** statisticians (Section **??**).

6.7 Choosing a Model

The parametric families presented here are often used in the real world. As indicated previously, this may be done on an empirical basis. We would collect data on a random variable X, and plot the frequencies of its values in a histogram. If for example the plot looks roughly like the curves in Figure 6.2, we could choose this as the family for our model.

Or, our choice may arise from theory. If for instance our knowledge of the setting in which we are working says that our distribution is memoryless, that forces us to use the exponential density family.

In either case, the question as to which member of the family we choose will be settled by using some kind of procedure which finds the member of the family which best fits our data. We will discuss this in detail in our chapters on statistics, especially Chapter **??**.

Note that we may choose not to use a parametric family at all. We may simply find that our data does not fit any of the common parametric families (there are many others than those presented here) very well. Procedures that do not assume any parametric family are termed **nonparametric**.

6.8 Finding the Density of a Function of a Random Variable

Suppose X has, say, a uniform distribution on (1,4). Form a new random variable, $Y = X^2$. How can we find $f_Y()$? Reason as follows. For $1 < t < 16$,

$$
\begin{aligned}
f_Y(t) &= \frac{d}{dt} F_Y(t) \ \text{ (def. of density)} && (6.53)\\
&= \frac{d}{dt} P(Y \leq t) \ \text{ (def. of cdf)} && (6.54)\\
&= \frac{d}{dt} P(X \leq \sqrt{t}) \ \text{ (def. of Y)} && (6.55)\\
&= \frac{d}{dt} F_X[(\sqrt{t}-1)/3] \ \text{ (def. of cdf)} && (6.56)\\
&= \frac{1}{3} f_X(\sqrt{t}) \ \frac{d}{d} \sqrt{t} \ \text{ (Chain Rule)} && (6.57)\\
&= \frac{1}{3} \cdot \frac{1}{2} t^{-0.5} && (6.58)\\
&= \frac{1}{6} t^{-0.5} && (6.59)
\end{aligned}
$$

Other such settings can be handled similarly. Note, though, that the above derivation relied on the fact that $X > 0$. Suppose X has a uniform distribution on (-1,1). Then the above derivation would become, for $0 < t < 1$,

$$
\begin{aligned}
f_Y(t) &= \frac{d}{dt} P(Y \leq t) && (6.60)\\
&= \frac{d}{dt} P(-\sqrt{t} \leq X \leq \sqrt{t}) && (6.61)\\
&= \frac{d}{dt} 2\sqrt{t}/2 && (6.62)\\
&= 0.5 t^{-0.5} && (6.63)
\end{aligned}
$$

6.9 Quantile Functions

First, recall the definition of the **inverse** of a function, say $h()$. The inverse of $h()$, denoted $h^{-1}()$, "brings you back to where you started." If I plug 3 into the squaring function, I get 9, and if I then plug 9 into the square root function, I get back my original 3.[5] So we say that $h(t) = t^2$ and $k(s) = \sqrt{s}$ are inverses of each other. The same relation holds between exp() and ln() and so on.

[5]This assumes, of course, that the domain of my squaring function consists only of the nonnegative numbers. We'll put aside this and similar situations for the time being, but will return.

For a random variable X, its **quantile function** $Q_X(s)$ is defined by

$$Q_X(s) = F_X^{-1}(s) \tag{6.64}$$

This is called the **s quantile** of X.

A well-known example is the **median** of X, the point which half the probability is above and half below. It's the 0.5 quantile of X.

The cdf tells us that cumulative probability for a particular value of X, while the quantile function does the opposite, "bringing us back." Let's make this concrete by considering the random variable Z in Section 6.3.1.

On the one hand, we can ask the question, "What is the probability that Z is at most 1?", with the answer being

$$F_Z(1) = 0.75 \tag{6.65}$$

On the other hand, one can ask conversely, "At what value of X do we have cumulative probability of 0.75?" Here the answer is

$$Q_Z(0.75) = 1 \tag{6.66}$$

It is no coincidence that the word *quantile* has that *-ile* suffix, given your familiarity with the word *percentile*. They are really the same thing.

Suppose for example that 92% of all who take the SAT Math Test have scores of at most 725. In other words, if you look at a randomly chosen test paper, and let X denote its score, then

$$F_X(725) = P(X \leq 725) = 0.92 \tag{6.67}$$

On the other hand, if you are interested in the $92nd$ percentile score for the test, you are saying "Find s such that $P(X \leq s) = 0.92$" you want

$$Q_X(0.92) = 725 \tag{6.68}$$

The reader should note that the R functions we've seen beginning with the letter 'q', such as **qgeom()** and **qunif()**, are quantile functions, hence the name.

However, a problem with discrete random variables is that the quantile values may not be unique. The reader should verify, for instance, in the coin-toss example above, the 0.75 quantile for Z could be not only 1, but also 1.1, 1.288 and so on. So, one should look carefully at the documentation of quantile functions, to see what they do to impose uniqueness. But for continuous random variables there is no such problem.

6.10 Using cdf Functions to Find Probabilities

As we have seen, for many parametric families R has "d/p/q/r" functions, giving the density, cdf, quantile function and random number generator for the given family. How can we use the cdf functions to find probabilities of the form $P(a < X < b)$?

We see the answer in (6.15). We simply evaluate the cdf at b then a, and subtract.

For instance, consider the network buffer example, Section 6.6.4.2. Let's find $P(540 < Y < 562)$:

```
> pgamma(562,5,0.01) - pgamma(540,5,0.01)
[1] 0.03418264
```

Of course, we could also integrate the density,

```
> integrate(function(t) dgamma(t,5,0.01),540,562)
0.03418264 with absolute error < 3.8e-16
```

but R does that for us, there is probably little point in that second approach.

6.11 A General Method for Simulating a Random Variable

Suppose we wish to simulate a random variable X with density f_X for which there is no R function. This can be done via $F_X^{-1}(U)$, where U has a U(0,1) distribution. In other words, we call **runif()** and then plug the result into the inverse of the cdf of X.

For example, say X has the density 2t on (0,1). Then $F_X(t) = t^2$, so $F^{-1}(s) = s^{0.5}$. We can then generate an X as **sqrt(runif(1))**. Here's why:

For brevity, denote F_X^{-1} as G. Our generated random variable is then Y = G(U). Then

$$
\begin{aligned}
F_Y(t) &= P[G(U) \le t] && (6.69)\\
&= P[U \le G^{-1}(t)] && (6.70)\\
&= P[U \le F_X(t)] && (6.71)\\
&= F_X(t) && (6.72)
\end{aligned}
$$

(this last coming from the fact that U has a uniform distribution on (0,1)).

In other words, Y and X have the same cdf, i.e. the same distribution! This is exactly what we want.

Note that this method, though valid, is not necessarily practical, since computing F_X^{-1} may not be easy.

6.12 Example: Writing a Set of R Functions for a Certain Power Family

Consider the family of distributions indexed by positive values of c with densities

$$c\, t^{c-1} \tag{6.73}$$

for t in (0,1) and 0 otherwise..

The cdf is t^c, so let's call this the "tc" family.

Let's find "d", "p", "q" and "r" functions for this family, just like R has for the normal family, the gamma family and so on:

```
# density
dtc <- function(x,c) c * x^(c-1)

# cdf
ptc <- function(x,c) x^c

# quantile function
qtc <- function(q,c) q^(1/c)

# random number generator, n values generated
rtc <- function(n,c) {
   tmp <- runif(n)
   qtc(tmp,c)
}
```

Note that to get **rtc()** we simply plug U(0,1) variates into **qtc()**, according to Section 6.11.

Let's check our work. The mean for the density having c equal to 2 is 2/3 (reader should verify); let's see if a simulation will give us that:

```
> mean(rtc(10000,2))
[1] 0.6696941
```

Sure enough!

6.13 Multivariate Densities

Section 4.14 briefly introduced the notion of multivariate pmfs. Similarly, there are also multivariate densities. Probabilities are then k-fold integrals, where k is the number of random variables.

For instance, a probability involving two variables means taking a double integral of a bivariate density. Since that density can be viewed as a surface in three-dimensional space (just as a univariate density is viewed as a curve in two-dimensional space), a probability is then a volume under that surface (as opposed to area in the univariate case). Conversely, a bivariate density is the mixed partial derivative of the cdf:

$$f_{X,Y}(u,v) = \frac{\partial^2}{\partial u \partial v} F_{X,Y}(u,v) = P(X \leq u,\ Y \leq v) \tag{6.74}$$

In analogy to

$$P(B \mid A) = \frac{P(B \text{ and } A)}{P(A)} \tag{6.75}$$

we can define the conditional density of Y given X:

$$f_{Y|X}(u,v) = \frac{f_{X,Y}(u,v)}{f_X(v)} \tag{6.76}$$

The intuition behind this is that we are conditioning on X being *near* v. Actually,

$$f_{Y|X}(u,v) = \lim_{h \to 0} \ [\text{density of Y} \mid X \in (v-h, v+h)] \tag{6.77}$$

A detailed treatment is presented in Chapter **??**.

6.14 Iterated Expectations

In analogy with (4.68), we have a very useful corresponding formula for the continuous case.

6.14.1 The Theorem

For any random variable W and any continuous random variable V,[6]

$$E(W) = \int_{-\infty}^{\infty} f_V(t)\ E(W \mid V = t)\ dt \tag{6.78}$$

Note that the event $V = t$ has probability 0 for coninuous V. The conditional expectation here is defined in terms of the conditional distribution of W given V; see Section 6.13.

Note too that if we have some event A, we can set W above to the indicator random variable of A (recall (3.9)), yielding

$$P(A) = \int_{-\infty}^{\infty} f_V(t)\ P(A \mid V = t)\ dt \tag{6.79}$$

6.14.2 Example: Another Coin Game

Suppose we have biased coins of various weightings, so that a randomly chosen coin's probability of heads H has density 2t on (0,1). The game has you choose a coin at random, toss it 5 times, and pays you a prize if you get 5 heads. What is your probability of winning?

First, note that the probability of winning, given H = t, is t^5. then (6.79) tells us that

$$P(\text{win}) = \int_0^1 2t\ t^5\ dt = \frac{2}{7} \tag{6.80}$$

6.15 Continuous Random Variables Are "Useful Unicorns"

Recall our random dart example at the outset of this chapter. It must be kept in mind that this is only an idealization. D actually cannot be just any old point in (0,1). To begin with, our measuring instrument has only finite precision. Actually, then, D can only take on a finite number of values. If, say, our precision is four decimal digits, then D can only be 0.0001, 0.0002, ..., 0.9999, making it a discrete random variable after all.[7]

So this modeling of the position of the dart as continuously distributed really is just an idealization. *Indeed, in practice there are NO continuous random variables.* But the coninuous model can be an

[6] The treatment here will be intuitive, rather than being a mathematical definition and proof.

[7] There are also issues such as the nonzero thickness of the dart, and so on, further restricting our measurement.

excellent approximation, and the concept is extremely useful. It's like the assumption of "massless string" in physics analyses; there is no such thing, but it's a good approximation to reality.

As noted, most applications of statistics, and many of probability, are based on continuous distributions. We'll be using them heavily for the remainder of this book.

Chapter 7

The Normal Distributions

Again, these are the famous "bell-shaped curves," so called because their densities have that shape.

7.1 Density and Properties

The density for a normal distribution is

$$f_W(t) = \frac{1}{\sqrt{2\pi}\sigma} \; e^{-0.5\left(\frac{t-\mu}{\sigma}\right)^2}, -\infty < t < \infty \tag{7.1}$$

Again, this is a two-parameter family, indexed by the parameters μ and σ, which turn out to be the mean[1] and standard deviation μ and σ, The notation for it is $N(\mu, \sigma^2)$ (it is customary to state the variance σ^2 rather than the standard deviation).

And we write

$$X \tilde{} N(\mu, \sigma^2) \tag{7.2}$$

to mean that the random variable X has the distribution $N(\mu, \sigma^2)$. (The tilde is read "is distributed as.")

Note: Saying "X has a $N(\mu, \sigma^2)$ distribution" is *more* than simply saying "X has mean μ and variance σ^2." The former statement tells us not only the mean and variance of X, but *also* the fact that X has a "bell-shaped" density in the (7.1) family.

[1]Remember, this is a synonym for expected value.

7.1.1 Closure Under Affine Transformation

The family is closed under affine transformations:

If

$$X \tilde{} N(\mu, \sigma^2) \tag{7.3}$$

and we set

$$Y = cX + d \tag{7.4}$$

then

$$Y \tilde{} N(c\mu + d, c^2\sigma^2) \tag{7.5}$$

For instance, suppose X is the height of a randomly selection UC Davis student, measured in inches. Human heights do have approximate normal distributions; a histogram plot of the student heights would look bell-shaped. Now let Y be the student's height in centimeters. Then we have the situation above, with c = 2.54 and d = 0. The claim about affine transformations of normally distributed random variables would imply that a histogram of Y would again be bell-shaped.

Consider the above statement carefully.

> It is saying much more than simply that Y has mean $c\mu + d$ and variance $c^2\sigma^2$, which would follow from our our "mailing tubes" such as (3.49) *even if X did not have a normal distribution.* The key point is that this new variable Y is also a member of the normal family, i.e. its density is still given by (7.1), now with the new mean and variance.

Let's derive this, using the reasoning of Section 6.8.

For convenience, suppose $c > 0$. Then

$$\begin{aligned}
F_Y(t) &= P(Y \leq t) \quad \text{(definition of } F_Y\text{)} && (7.6)\\
&= P(cX + d \leq t) \quad \text{(definition of Y)} && (7.7)\\
&= P\left(X \leq \frac{t-d}{c}\right) \quad \text{(algebra)} && (7.8)\\
&= F_X\left(\frac{t-d}{c}\right) \quad \text{(definition of } F_X\text{)} && (7.9)
\end{aligned}$$

Therefore

$$
\begin{aligned}
f_Y(t) &= \frac{d}{dt} F_Y(t) \quad \text{(definition of } f_Y) && (7.10)\\
&= \frac{d}{dt} F_X\left(\frac{t-d}{c}\right) \quad \text{(from (7.9))} && (7.11)\\
&= f_X\left(\frac{t-d}{c}\right) \cdot \frac{d}{dt}\frac{t-d}{c} \quad \text{(definition of } f_X \text{ and the Chain Rule)} && (7.12)\\
&= \frac{1}{c} \cdot \frac{1}{\sqrt{2\pi}\sigma}\ e^{-0.5\left(\frac{\frac{t-d}{c}-\mu}{\sigma}\right)^2} \quad \text{(from (7.1)} && (7.13)\\
&= \frac{1}{\sqrt{2\pi}(c\sigma)}\ e^{-0.5\left(\frac{t-(c\mu+d)}{c\sigma}\right)^2} \quad \text{(algebra)} && (7.14)
\end{aligned}
$$

That last expression is the $N(c\mu + d, c^2\sigma^2)$ density, so we are done!

7.1.2 Closure Under Independent Summation

If X and Y are independent random variables, each having a normal distribution, then their sum S = X + Y also is normally distributed.

This is a pretty remarkable phenomenon, not true for most other parametric families. If for instance X and Y each with, say, a U(0,1) distribution, then the density of S turns out to be triangle-shaped, NOT another uniform distribution. (This can be derived using the methods of Section **??**.)

Note that if X and Y are independent and normally distributed, then the two properties above imply that cX + dY will also have a normal distribution, for any constants c and d.

More generally:

> For constants $a_1, ..., a_k$ and *independent* random variables $X_1, ..., X_k$, with
>
> $$X_i \tilde{} N(\mu_i, \sigma_i^2 \qquad (7.15)$$
>
> form the new random variable $Y = a_1 X_1 + ... + a_k X_k$. Then
>
> $$Y \tilde{} N(\sum_{i=1}^k a_i \mu_i, \sum_{i=1}^k a_i^2 \sigma_i^2) \qquad (7.16)$$

Lack of intuition:

The reader should ponder how remarkable this property of the normal family is, because there is no intuitive explanation for it.

Imagine random variables X and Y, each with a normal distribution. Say the mean and variances are 10 and 4 for X, and 18 and 6 for Y. We repeat our experiment 1000 times for our "notebook," i.e. 1000 lines with 2 columns. If we draw a histogram of the X column, we'll get a bell-shaped curve, and the same will be true for the Y column.

But now add a Z column, for $Z = X + Y$. Why in the world should a histogram of the Z column also be bell-shaped?

7.2 R Functions

```
dnorm(x, mean = 0, sd = 1)
pnorm(q, mean = 0, sd = 1)
qnorm(p, mean = 0, sd = 1)
rnorm(n, mean = 0, sd = 1)
```

Here **mean** and **sd** are of course the mean and standard deviation of the distribution. The other arguments are as in our previous examples.

7.3 The Standard Normal Distribution

Definition 11 *If $Z\tilde{}N(0,1)$ we say the random variable Z has a standard normal distribution.*

Note that if $X\tilde{}N(\mu,\sigma^2)$, and if we set

$$Z = \frac{X-\mu}{\sigma} \tag{7.17}$$

then

$$Z\tilde{}N(0,1) \tag{7.18}$$

The above statements follow from the earlier material:

- Define $Z = \frac{X-\mu}{\sigma}$.
- Rewrite it as $Z = \frac{1}{\sigma} \cdot X + (\frac{-\mu}{\sigma})$.

- Since E(cU + d) = c EU + d for any random variable U and constants c and d, we have

$$EZ = \frac{1}{\sigma}EX - \frac{\mu}{\sigma} = 0 \tag{7.19}$$

  and (3.56) and (3.49) imply that Var(X) = 1.

- OK, so we know that Z has mean 0 and variance 1. But does it have a normal distribution? Yes, due to our discussion above titled "Closure Under Affine Transformations."

By the way, the N(0,1) cdf is traditionally denoted by Φ.

7.4 Evaluating Normal cdfs

The function in (7.1) does not have a closed-form indefinite integral. Thus probabilities involving normal random variables must be approximated. Traditionally, this is done with a table for the cdf of N(0,1), which is included as an appendix to almost any statistics textbook; the table gives the cdf values for that distribution.

But this raises a question: There are infinitely many distributions in the normal family. Don't we need a separate table for each? That of course would not be possible, and in fact it turns out that this one table—the one for the N(0,1) distribution— is sufficient for the entire normal family. Though we of course will use R to gt such probabilities, it will be quite instructive to see how these table operations work.

Here's why one table is enough: Say X has an $N(10, 2.5^2)$ distribution. How can we get a probability like, say, $P(X < 12)$ using the N(0,1) table? Write

$$P(X < 12) = P\left(Z < \frac{12-10}{2.5}\right) = P(Z < 0.8) \tag{7.20}$$

Since on the right-hand side Z has a standard normal distribution, we can find that latter probably from the N(0,1) table!

As noted, traditionally it has played a central role, as one could transform any probability involving some normal distribution to an equivalent probability involving N(0,1). One would then use a table of N(0,1) to find the desired probability.

The transformation $Z = (X - \mu)/\sigma$ will play a big role in other contexts in future chapters, but for the sole purpose of simply evaluating normal probabilities, we can be much more direct. Nowadays, probabilities for any normal distribution, not just N(0,1), are easily available by computer. In the R statistical package, the normal cdf for any mean and variance is available via the function **pnorm()**. The call form is

```
pnorm(q,mean=0,sd=1)
```

This returns the value of the cdf evaluated at **q**, for a normal distribution having the specified mean and standard deviation (default values of 0 and 1).

We can use **rnorm()** to simulate normally distributed random variables. The call is

```
rnorm(n,mean=0,sd=1)
```

which returns a vector of **n** random variates from the specified normal distribution.

There are also of course the corresponding density and quantile functions, **dnorm()** and **qnorm()**.

7.5 Example: Network Intrusion

As an example, let's look at a simple version of the network intrusion problem. Suppose we have found that in Jill's remote logins to a certain computer, the number X of disk sectors she reads or writes has an approximate normal distribution with a mean of 500 and a standard deviation of 15.

Before we continue, a comment on modeling: Since the number of sectors is discrete, it could not have an exact normal distribution. But then, no random variable in practice has an exact normal or other continuous distribution, as discussed in Section 6.15, and the distribution can indeed by approximately normal.

Now, say our network intrusion monitor finds that Jill—or someone posing as her—has logged in and has read or written 535 sectors. Should we be suspicious?

To answer this question, let's find $P(X \geq 535)$: Let $Z = (X - 500)/15$. From our discussion above, we know that Z has a N(0,1) distribution, so

$$P(X \geq 535) = P\left(Z \geq \frac{535 - 500}{15}\right) = 1 - \Phi(35/15) = 0.01 \tag{7.21}$$

Again, traditionally we would obtain that 0.01 value from a N(0,1) cdf table in a book. With R, we would just use the function **pnorm()**:

```
> 1 - pnorm(535,500,15)
[1] 0.009815329
```

Anyway, that 0.01 probability makes us suspicious. While it *could* really be Jill, this would be unusual behavior for Jill, so we start to suspect that it isn't her. It's suspicious enough for us to probe more deeply, e.g. by looking at which files she (or the impostor) accessed—were they rare for Jill too?

Now suppose there are two logins to Jill's account, accessing X and Y sectors, with X+Y = 1088. Is this rare for her, i.e. is $P(X + Y > 1088)$? small?

We'll assume X and Y are independent. We'd have to give some thought as to whether this assumption is reasonable, depending on the details of how we observed the logins, etc., but let's move ahead on this basis.

From page 147, we know that the sum S = X+Y is again normally distributed. Due to the properties in Chapter 3, we know S has mean $2 \cdot 500$ and variance $2 \cdot 15^2$. The desired probability is then found via

```
1 - pnorm(1088,1000,sqrt(450))
```

which is about 0.00002. That is indeed a small number, and we should be highly suspicious.

Note again that the normal model (or any other continuous model) can only be approximate, especially in the tails of the distribution, in this case the right-hand tail. But it is clear that S is only rarely larger than 1088, and the matter mandates further investigation.

Of course, this is very crude analysis, and real intrusion detection systems are much more complex, but you can see the main ideas here.

7.6 Example: Class Enrollment Size

After years of experience with a certain course, a university has found that online pre-enrollment in the course is approximately normally distributed, with mean 28.8 and standard deviation 3.1. Suppose that in some particular offering, pre-enrollment was capped at 25, and it hit the cap. Find the probability that the actual demand for the course was at least 30.

Note that this is a conditional probability! Evaulate it as follows. Let N be the actual demand. Then the key point is that we are given that $N \geq 25$, so

$$
\begin{aligned}
P(N \geq 30 | N \geq 25) &= \frac{P(N \geq 30 \text{ and } N \geq 25)}{P(N \geq 25)} \quad ((2.7)) && (7.22)\\
&= \frac{P(N \geq 30)}{P(N \geq 25)} && (7.23)\\
&= \frac{1 - \Phi\left[(30 - 28.8)/3.1\right]}{1 - \Phi\left[(25 - 28.8)/3.1\right]} && (7.24)\\
&= 0.39 && (7.25)
\end{aligned}
$$

Sounds like it may be worth moving the class to a larger room before school starts.

Since we are approximating a discrete random variable by a continuous one, it might be more accurate here to use a **correction for continuity**, described in Section 7.14.

7.7 More on the Jill Example

Continuing the Jill example, suppose there is never an intrusion, i.e. all logins are from Jill herself. Say we've set our network intrusion monitor to notify us every time Jill logs in and accesses 535 or more disk sectors. In what proportion of all such notifications will Jill have accessed at least 545 sectors?

This is $P(X \geq 545—X \geq 535)$. By an analysis similar to that in Section 7.6, this probability is

```
(1 - pnorm(545,500,15)) / (1 - pnorm(535,500,15))
```

7.8 Example: River Levels

Consider a certain river, and L, its level (in feet) relative to its average. There is a flood whenever L > 8, and it is reported that 2.5% of days have flooding. Let's assume that the level L is normally distributed; the above information implies that the mean is 0.

Suppose the standard deviation of L, σ, goes up by 10%. How much will the percentage of flooding days increase?

To solve this, let's first find σ. We have that

$$0.025 = P(L > 8) = P\left(\frac{L-0}{\sigma} > \frac{8-0}{\sigma}\right) \tag{7.26}$$

Since $(L-0)/\sigma$ has a N(0,1) distribution, we can find the 0.975 point in its cdf:

```
> qnorm(0.975,0,1)
[1] 1.959964
```

So,

$$1.96 = \frac{8-0}{\sigma} \tag{7.27}$$

so σ is about 4.

If it increases to 4.4, then we can evaluate $P(L > 8)$ by

```
> 1 - pnorm(8,0,4.4)
[1] 0.03451817
```

So, a 10% increase in σ would lead in this case to about a 40% increase in flood days.

7.9 Example: Upper Tail of a Light Bulb Distribution

Suppose we model light bulb lifetimes as having a normal distribution with mean and standard deviation 500 and 50 hours, respectively. Give a loop-free R expression for finding the value of d such that 30% of all bulbs have lifetime more than d.

You should develop the ability to recognize when we need **p**-series and **q**-series functions. Here we need

```
qnorm(1-0.30,500,50)
```

7.10 The Central Limit Theorem

The Central Limit Theorem (CLT) says, roughly speaking, that a random variable which is a sum of many components will have an approximate normal distribution. So, for instance, human weights are approximately normally distributed, since a person is made of many components. The same is true for SAT test scores,[2] as the total score is the sum of scores on the individual problems.

There are many versions of the CLT. The basic one requires that the summands be independent and identically distributed:[3]

[2]This refers to the raw scores, before scaling by the testing company.

[3]A more mathematically precise statement of the theorem is given in Section **??**.

Theorem 12 *Suppose $X_1, X_2, ...$ are independent random variables, all having the same distribution which has mean m and variance v^2. Form the new random variable $T = X_1 + ... + X_n$. Then for large n, the distribution of T is approximately normal with mean nm and variance nv^2.*

The larger n is, the better the approximation, but typically n = 20 or even n = 10 is enough.

7.11 Example: Cumulative Roundoff Error

Suppose that computer roundoff error in computing the square roots of numbers in a certain range is distributed uniformly on (-0.5,0.5), and that we will be computing the sum of n such square roots. Suppose we compute a sum of 50 square roots. Let's find the approximate probability that the sum is more than 2.0 higher than it should be. (Assume that the error in the summing operation is negligible compared to that of the square root operation.)

Let $U_1, ..., U_{50}$ denote the errors on the individual terms in the sum. Since we are computing a sum, the errors are added too, so our total error is

$$T = U_1 + ... + U_{50} \tag{7.28}$$

By the Central Limit Theorem, since T is a sum, it has an approximately normal distribution, with mean 50 EU and variance 50 Var(U), where U is a random variable having the distribution of the U_i. From Section 6.6.1.1, we know that

$$EU = (-0.5 + 0.5)/2 = 0, \quad Var(U) = \frac{1}{12}[0.5 - (-0.5)]^2 = \frac{1}{12} \tag{7.29}$$

So, the approximate distribution of T is N(0,50/12). We can then use R to find our desired probability:

```
> 1 - pnorm(2,mean=0,sd=sqrt(50/12))
[1] 0.1635934
```

7.12 Example: R Evaluation of a Central Limit Theorem Approximation

Say $W = U_1 + ... + U_{50}$, with the U_i being independent and identically distributed (i.i.d.) with uniform distributions on (0,1). Give an R expression for the approximate value of $P(W < 23.4)$.

W has an approximate normal distribution, with mean 50×0.5 and variance $50 \times (1/12)$. So we need

```
pnorm(23.4,25,sqrt(50/12))
```

7.13 Example: Bug Counts

As an example, suppose the number of bugs per 1,000 lines of code has a Poisson distribution with mean 5.2. Let's find the probability of having more than 106 bugs in 20 sections of code, each 1,000 lines long. We'll assume the different sections act independently in terms of bugs.

Here X_i is the number of bugs in the ith section of code, and T is the total number of bugs. This is another clear candidate for using the CLT.

Since each X_i has a Poisson distribution, $m = v^2 = 5.2$. So, T, being a sum, is approximately distributed normally with mean and variance 20×5.2. So, we can find the approximate probability of having more than 106 bugs:

```
> 1 - pnorm(106,20*5.2,sqrt(20*5.2))
[1] 0.4222596
```

7.14 Example: Coin Tosses

Binomially distributed random variables, though discrete, also are approximately normally distributed. Here's why:

Say T has a binomial distribution with n trials. Then we can write T as a sum of indicator random variables (Section 3.9):

$$T = T_1 + ... + T_n \tag{7.30}$$

where T_i is 1 for a success and 0 for a failure on the ith trial. Since we have a sum of independent, identically distributed terms, the CLT applies. Thus we use the CLT if we have binomial distributions with large n.

For example, let's find the approximate probability of getting more than 12 heads in 20 tosses of a coin. X, the number of heads, has a binomial distribution with n = 20 and p = 0.5 Its mean and variance are then np = 10 and np(1-p) = 5. So, let $Z = (X - 10)/\sqrt{5}$, and write

$$P(X > 12) = P(Z > \frac{12 - 10}{\sqrt{5}}) \approx 1 - \Phi(0.894) = 0.186 \tag{7.31}$$

Or:

```
> 1 - pnorm(12,10,sqrt(5))
[1] 0.1855467
```

The exact answer is 0.132, not too close. Why such a big error? The main reason is n here is rather small. But actually, we can still improve the approximation quite a bit, as follows.

Remember, the reason we did the above normal calculation was that X is approximately normal, from the CLT. This is an approximation of the distribution of a discrete random variable by a continuous one, which introduces additional error.

We can get better accuracy by using the **correction of continuity**, which can be motivated as follows. As an alternative to (7.31), we might write

$$P(X > 12) = P(X \geq 13) = P(Z > \frac{13 - 10}{\sqrt{5}}) \approx 1 - \Phi(1.342) = 0.090 \tag{7.32}$$

That value of 0.090 is considerably smaller than the 0.186 we got from (7.31). We could "split the difference" this way:

$$P(X > 12) = P(X \geq 12.5) = P(Z > \frac{12.5 - 10}{\sqrt{5}}) \approx 1 - \Phi(1.118) = 0.132 \tag{7.33}$$

(Think of the number 13 "owning" the region between 12.5 and 13.5, 14 owning the part between 13.5 and 14.5 and so on.) Since the exact answer to seven decimal places is 0.131588, the strategy has improved accuracy substantially.

The term *correction for continuity* alludes to the fact that we are approximately a discrete distribution by a continuous one.

7.15 Example: Normal Approximation to Gamma Family

Recall from above that the gamma distribution, or at least the Erlang, arises as a sum of independent random variables. Thus the Central Limit Theorem implies that the gamma distribution should be approximately normal for large (integer) values of r. We see in Figure 6.2 that even with r = 10 it is rather close to normal.[4]

7.16 Example: Museum Demonstration

Many science museums have the following visual demonstration of the CLT.

There are many balls in a chute, with a triangular array of r rows of pins beneath the chute. Each ball falls through the rows of pins, bouncing left and right with probability 0.5 each, eventually being collected into one of r bins, numbered 0 to r. A ball will end up in bin i if it bounces rightward in i of the r rows of pins, i = 0,1,...,r. Key point:

> Let X denote the bin number at which a ball ends up. X is the number of rightward bounces ("successes") in r rows ("trials"). Therefore X has a binomial distribution with n = r and p = 0.5

Each bin is wide enough for only one ball, so the balls in a bin will stack up. And since there are many balls, the height of the stack in bin i will be approximately proportional to P(X = i). And since the latter will be approximately given by the CLT, the stacks of balls will roughly look like the famous bell-shaped curve!

There are many online simulations of this museum demonstration, such as `http://www.mathsisfun.com/data/quincunx.html`. By collecting the balls in bins, the apparatus basically simulates a histogram for X, which will then be approximately bell-shaped.

7.17 Importance in Modeling

Needless to say, there are no random variables in the real world that are exactly normally distributed. In addition to our comments at the beginning of this chapter that no real-world random variable has a continuous distribution, there are no practical applications in which a random variable is not bounded on both ends. This contrasts with normal distributions, which extend from $-\infty$ to ∞.

Yet, many things in nature do have approximate normal distributions, so normal distributions play a key role in statistics. Most of the classical statistical procedures assume that one has sampled from a population having an approximate distribution. In addition, it will be seen later than the CLT

[4] It should be mentioned that technically, the CLT, which concerns convergence of cdfs, does not imply convergence of densities. However, under mild mathematical conditions, convergence of densities occurs too.

tells us in many of these cases that the quantities used for statistical estimation are approximately normal, even if the data they are calculated from are not.

Recall from above that the gamma distribution, or at least the Erlang, arises as a sum of independent random variables. Thus the Central Limit Theorem implies that the gamma distribution should be approximately normal for large (integer) values of r. We see in Figure 6.2 that even with r = 10 it is rather close to normal.

Chapter 8

The Exponential Distributions

The family of exponential distributions, Section 6.6.3, has a number of remarkable properties, which contribute to its widespread usage in probabilistic modeling. We'll discuss those here.

8.1 Connection to the Poisson Distribution Family

Suppose the lifetimes of a set of light bulbs are independent and identically distributed **(i.i.d.)**, and consider the following process. At time 0, we install a light bulb, which burns an amount of time X_1. Then we install a second light bulb, with lifetime X_2. Then a third, with lifetime X_3, and so on.

Let

$$T_r = X_1 + ... + X_r \tag{8.1}$$

denote the time of the r^{th} replacement. Also, let N(t) denote the number of replacements up to and including time t. Then it can be shown that if the common distribution of the X_i is exponentially distributed, the N(t) has a Poisson distribution with mean ηt. And the converse is true too: If the X_i are independent and identically distributed and N(t) is Poisson, then the X_i must have exponential distributions. In summary:

Theorem 13 *Suppose $X_1, X_2, ...$ are i.i.d. nonnegative continuous random variables. Define*

$$T_r = X_1 + ... + X_r \tag{8.2}$$

and

$$N(t) = \max\{k : T_k \leq t\} \tag{8.3}$$

Then the distribution of N(t) is Poisson with parameter ηt for all t if and only if the X_i have an exponential distribution with parameter η.

In other words, N(t) will have a Poisson distribution if and only if the lifetimes are exponentially distributed.

Proof

"Only if" part:

The key is to notice that the event $X_1 > t$ is exactly equivalent to $N(t) = 0$. If the first light bulb lasts longer than t, then the count of burnouts at time t is 0, and vice versa. Then

$$\begin{aligned} P(X_1 > t) &= P[N(t) = 0] \quad \text{(see above equiv.)} && (8.4)\\ &= \frac{(\eta t)^0}{0!} \cdot e^{-\eta t} \quad ((4.41) && (8.5)\\ &= e^{-\eta t} && (8.6) \end{aligned}$$

Then

$$f_{X_1}(t) = \frac{d}{dt}(1 - e^{-\eta t}) = \eta e^{-\eta t} \tag{8.7}$$

That shows that X_1 has an exponential distribution, and since the X_i are i.i.d., that implies that all of them have that distribution.

"If" part:

We need to show that if the X_i are exponentially distributed with parameter η, then for u nonnegative and each positive integer k,

$$P[N(u) = k] = \frac{(\eta u)^k e^{-\eta u}}{k!} \tag{8.8}$$

The proof for the case k = 0 just reverses (8.4) above. The general case, not shown here, notes that $N(u) \leq k$ is equivalent to $T_{k+1} > u$. The probability of the latter event can be found by integrating

(6.46) from u to infinity. One needs to perform k-1 integrations by parts, and eventually one arrives at (8.8), summed from 1 to k, as required.

■

The collection of random variables N(t) $t \geq 0$, is called a **Poisson process**.

The relation $E[N(t)] = \eta t$ says that replacements are occurring at an average rate of η per unit time. Thus η is called the **intensity parameter** of the process. It is this "rate" interpretation that makes η a natural indexing parameter in (6.40).

8.2 Memoryless Property of Exponential Distributions

One of the reasons the exponential family of distributions is so famous is that it has a property that makes many practical stochastic models mathematically tractable: The exponential distributions are **memoryless**.

8.2.1 Derivation and Intuition

What the term *memoryless* means for a random variable W is that for all positive t and u

$$P(W > t + u | W > t) = P(W > u) \tag{8.9}$$

Any exponentially distributed random variable has this property. Let's derive this:

$$\begin{aligned}
P(W > t + u | W > t) &= \frac{P(W > t + u \text{ and } W > t)}{P(W > t)} && (8.10) \\
&= \frac{P(W > t + u)}{P(W > t)} && (8.11) \\
&= \frac{\int_{t+u}^{\infty} \lambda e^{-\lambda s} \, ds}{\int_t^{\infty} \lambda e^{-\lambda s} \, ds} && (8.12) \\
&= e^{-\lambda u} && (8.13) \\
&= P(W > u) && (8.14)
\end{aligned}$$

We say that this means that "time starts over" at time t, or that W "doesn't remember" what happened before time t.

It is difficult for the beginning modeler to fully appreciate the memoryless property. Let's make it concrete. Consider the problem of waiting to cross the railroad tracks on Eighth Street in Davis, just west of J Street. One cannot see down the tracks, so we don't know whether the end of the train will come soon or not.

If we are driving, the issue at hand is whether to turn off the car's engine. If we leave it on, and the end of the train does not come for a long time, we will be wasting gasoline; if we turn it off, and the end does come soon, we will have to start the engine again, which also wastes gasoline. (Or, we may be deciding whether to stay there, or go way over to the Covell Rd. railroad overpass.)

Suppose our policy is to turn off the engine if the end of the train won't come for at least s seconds. Suppose also that we arrived at the railroad crossing just when the train first arrived, and we have already waited for r seconds. Will the end of the train come within s more seconds, so that we will keep the engine on? If the length of the train were exponentially distributed (if there are typically many cars, we can model it as continuous even though it is discrete), Equation (8.9) would say that the fact that we have waited r seconds so far is of no value at all in predicting whether the train will end within the next s seconds. The chance of it lasting at least s more seconds right now is no more and no less than the chance it had of lasting at least s seconds when it first arrived.

8.2.2 Uniquely Memoryless

By the way, the exponential distributions are the only continuous distributions which are memoryless. (Note the word *continuous*; in the discrete realm, the family of geometric distributions are also uniquely memoryless.) This too has implications for the theory. A rough proof of this uniqueness is as follows:

Suppose some continuous random variable V has the memoryless property, and let R(t) denote $1 - F_V(t)$. Then from (8.9), we would have

$$R(t+u)]/R(t) = R(u) \tag{8.15}$$

or

$$R(t+u) = R(t)R(u) \tag{8.16}$$

Differentiating both sides with respect to t, we'd have

$$R\prime(t+u) = R\prime(t)R(u) \tag{8.17}$$

Setting t to 0, this would say

$$R\prime(u) = R\prime(0)R(u) \tag{8.18}$$

This is a well-known differential equation, whose solution is

$$R(u) = e^{-cu} \tag{8.19}$$

which is exactly 1 minus the cdf for an exponentially distributed random variable.

8.2.3 Example: "Nonmemoryless" Light Bulbs

Suppose the lifetimes in years of light bulbs have the density 2t/15 on (1,4), 0 elsewhere. Say I've been using bulb A for 2.5 years now in a certain lamp, and am continuing to use it. But at this time I put a new bulb, B, in a second lamp. I am curious as to which bulb is more likely to burn out within the next 1.2 years. Let's find the two probabilities.

For bulb A:

$$P(L > 3.7|L > 2.5) = \frac{P(L > 3.7)}{P(L > 2.5)} = 0.24 \tag{8.20}$$

For bulb B:

$$P(X > 1.2) = \int_{1.2}^{4} 2t/15 \ dt = 0.97 \tag{8.21}$$

So you can see that the bulbs do have "memory." We knew this beforehand, since the exponential distributions are the only continuous ones that have no memory.

8.3 Example: Minima of Independent Exponentially Distributed Random Variables

The memoryless property of the exponential distribution (Section 8.2 leads to other key properties. Here's a famous one:

Theorem 14 *Suppose $W_1, ..., W_k$ are independent random variables, with W_i being exponentially distributed with parameter λ_i. Let $Z = \min(W_1, ..., W_k)$. Then Z too is exponentially distributed with parameter $\lambda_1 + ... + \lambda_k$, and thus has mean equal to the reciprocal of that parameter*

Comments:

- In "notebook" terms, we would have k+1 columns, one each for the W_i and one for Z. For any given line, the value in the Z column will be the smallest of the values in the columns for $W_1, ..., W_k$; Z will be equal to one of them, but not the same one in every line. Then for instance $P(Z = W_3)$ is interpretable in notebook form as the long-run proportion of lines in which the Z column equals the W_3 column.

- It's pretty remarkable that the minimum of independent exponential random variables turns out again to be exponential. Contrast that with Section **??**, where it is found that the minimum of independent uniform random variables does NOT turn out to have a uniform distribution.

- The sum $\lambda_1 + ... + \lambda_n$ in (a) should make good intuitive sense to you, for the following reasons. Recall from Section 8.1 that the parameter λ in an exponential distribution is interpretable as a "light bulb burnout rate."

  Say we have persons 1 and 2. Each has a lamp. Person i uses Brand i light bulbs, i = 1,2. Say Brand i light bulbs have exponential lifetimes with parameter λ_i. Suppose each time person i replaces a bulb, he shouts out, "New bulb!" and each time *anyone* replaces a bulb, I shout out "New bulb!" Persons 1 and 2 are shouting at a rate of λ_1 and λ_2, respectively, so I am shouting at a rate of $\lambda_1 + \lambda_2$.

Proof

$$
\begin{aligned}
F_Z(t) &= P(Z \leq t) \quad \text{(def. of cdf)} && (8.22)\\
&= 1 - P(Z > t) && (8.23)\\
&= 1 - P(W_1 > t \text{ and ... and } W_k > t) \quad (\min > t \text{ iff all } W_i > t) && (8.24)\\
&= 1 - \Pi_i \ P(W_i > t) \quad \text{(indep.)} && (8.25)\\
&= 1 - \Pi_i \ e^{-\lambda_i t} \quad \text{(expon. distr.)} && (8.26)\\
&= 1 - e^{-(\lambda_1 + ... + \lambda_n)t} && (8.27)
\end{aligned}
$$

Taking $\frac{d}{dt}$ of both sides proves the theorem.

■

Also:

Theorem 15 *Under the conditions in Theorem 14,*

$$P(W_i < W_1, ..., W_{i-1}, W_{i+1}, ..., W_k) = \frac{\lambda_i}{\lambda_1 + ... + \lambda_k} \tag{8.28}$$

(There are k terms in the denominator, not k-1.)

Equation (8.28) should be intuitively clear as well from the above "thought experiment" (in which we shouted out "New bulb!"): On average, we have one new Brand 1 bulb every $1/\lambda_1$ time, so in a long time t, we'll have about $t\lambda_1$ shouts for this brand. We'll also have about $t\lambda_2$ shouts for Brand 2. So, a proportion of about

$$\frac{t\lambda_1}{t\lambda_1 + t\lambda_2} \tag{8.29}$$

of the shots are for Brand 1. Also, at any given time, the memoryless property of exponential distributions implies that the time at which I shout next will be the *minimum* of the times at which persons 1 and 2 shout next. This intuitively implies (8.28).

Proof

Again consider the case k = 2, and then use induction.

Let $Z = \min(W_1, W_2)$ as before. Then

$$P(Z = W_1 | W_1 = t) = P(W_2 > t \mid W_1 = t) \tag{8.30}$$

(Note: We are working with continuous random variables here, so quantities like $P(W_1 = t)$ are 0 (though actually $P(Z = W_1)$ is nonzero). So, as mentioned in Section 6.79, quantities like $P(Z = W_1 \mid W_1 = t)$ really mean "the probability that $W_2 > t$ in the conditional distribution of Z given W_1.)

Since W_1 and W_2 are independent,

$$P(W_2 > t \mid W_1 = t) = P(W_2 > t) = e^{-\lambda_2 t} \tag{8.31}$$

Now use (6.79):

$$P(Z = W_1) = \int_0^\infty \lambda_1 e^{-\lambda_1 t} e^{-\lambda_2 t}\, dt = \frac{\lambda_1}{\lambda_1 + \lambda_2} \tag{8.32}$$

as claimed.

■

This property of minima of independent exponentially-distributed random variables developed in this section is key to the structure of continuous-time Markov chains, in Chapter **??**.

8.3.1 Example: Computer Worm

A computer science graduate student at UCD, Senthilkumar Cheetancheri, was working on a worm alert mechanism. A simplified version of the model is that network hosts are divided into groups of size g, say on the basis of sharing the same router. Each infected host tries to infect all the others in the group. When g-1 group members are infected, an alert is sent to the outside world.

The student was studying this model via simulation, and found some surprising behavior. No matter how large he made g, the mean time until an external alert was raised seemed bounded. He asked me for advice.

I modeled the nodes as operating independently, and assumed that if node A is trying to infect node B, it takes an exponentially-distributed amount of time to do so. This is a continuous-time Markov chain. Again, this topic is much more fully developed in Chapter **??**, but all we need here is the result of Section 8.3, that exponential distributions are "memoryless.".

In state i, there are i infected hosts, each trying to infect all of the g-i noninfected hosts. When the process reaches state g-1, the process ends; we call this state an **absorbing state**, i.e. one from which the process never leaves.

Scale time so that for hosts A and B above, the mean time to infection is 1.0. Since in state i there are i(g-i) such pairs, the time to the next state transition is the minimum of i(g-i) exponentially-distributed random variables with mean 1. Theorem 14 tells us that this minimum is also exponentially distributed, with parameter $i(g-i) \cdot 1$. Thus the mean time to go from state i to state i+1 is 1/[i(g-i)].

Then the mean time to go from state 1 to state g-1 is

$$\sum_{i=1}^{g-1} \frac{1}{i(g-i)} \tag{8.33}$$

Using a calculus approximation, we have

$$\int_1^{g-1} \frac{1}{x(g-x)}\ dx = \frac{1}{g} \int_1^{g-1} (\frac{1}{x} + \frac{1}{g-x})\ dx = \frac{2}{g} \ln(g-1) \tag{8.34}$$

The latter quantity goes to zero as $g \to \infty$. This confirms that the behavior seen by the student in simulations holds in general. In other words, (8.33) remains bounded as $g \to \infty$. This is a very interesting result, since it says that the mean time to alert is bounded no matter how big our group size is.

So, even though our model here was quite simple, probably overly so, it did explain why the student was seeing the surprising behavior in his simulations.

8.3.2 Example: Electronic Components

Suppose we have three electronic parts, with independent lifetimes that are exponentially distributed with mean 2.5. They are installed simultaneously. Let's find the mean time until the last failure occurs.

Actually, we can use the same reasoning as for the computer worm example in Section 8.3.1: The mean time is simply

$$1/(3 \cdot 0.4) + 1/(2 \cdot 0.4) + 1/(1 \cdot 0.4) \tag{8.35}$$

8.4 A Cautionary Tale: the Bus Paradox

Suppose you arrive at a bus stop, at which buses arrive according to a Poisson process with intensity parameter 0.1, i.e. 0.1 arrival per minute. Recall that the means that the interarrival times have an exponential distribution with mean 10 minutes. What is the expected value of your waiting time until the next bus?

Well, our first thought might be that since the exponential distribution is memoryless, "time starts over" when we reach the bus stop. Therefore our mean wait should be 10.

On the other hand, we might think that on average we will arrive halfway between two consecutive buses. Since the mean time between buses is 10 minutes, the halfway point is at 5 minutes. Thus it would seem that our mean wait should be 5 minutes.

Which analysis is correct? Actually, the correct answer is 10 minutes. So, what is wrong with the second analysis, which concluded that the mean wait is 5 minutes? The problem is that the second analysis did not take into account the fact that although inter-bus intervals have an exponential distribution with mean 10, *the particular inter-bus interval that we encounter is special.*

8.4.1 Length-Biased Sampling

Imagine a bag full of sticks, of different lengths. We reach into the bag and choose a stick at random. The key point is that not all pieces are equally likely to be chosen; the longer pieces will have a greater chance of being selected.

Say for example there are 50 sticks in the bag, with ID numbers from 1 to 50. Let X denote the length of the stick we obtain if select a stick on an equal-probability basis, i.e. each stick having probability 1/50 of being chosen. (We select a random number I from 1 to 50, and choose the stick with ID number I.) On the other hand, let Y denote the length of the stick we choose by reaching into the bag and pulling out whichever stick we happen to touch first. Intuitively, the distribution of Y should favor the longer sticks, so that for instance $EY > EX$.

Let's look at this from a "notebook" point of view. We pull a stick out of the bag by random ID number, and record its length in the X column of the first line of the notebook. Then we replace the stick, and choose a stick out by the "first touch" method, and record its length in the Y column of the first line. Then we do all this again, recording on the second line, and so on. Again, because the "first touch" method will favor the longer sticks, the long-run average of the Y column will be larger than the one for the X column.

Another example was suggested to me by UCD grad student Shubhabrata Sengupta. Think of a large parking lot on which hundreds of buckets are placed of various diameters. We throw a ball high into the sky, and see what size bucket it lands in. Here the density would be proportional to area of the bucket, i.e. to the square of the diameter.

Similarly, the particular inter-bus interval that we hit is likely to be a longer interval. To see this, suppose we observe the comings and goings of buses for a very long time, and plot their arrivals on a time line on a wall. In some cases two successive marks on the time line are close together, sometimes far apart. If we were to stand far from the wall and throw a dart at it, we would hit the interval between some pair of consecutive marks. Intuitively we are more apt to hit a wider interval than a narrower one.

The formal name for this is **length-biased sampling**.

Once one recognizes this and carefully derives the density of that interval (see below), we discover that that interval does indeed tend to be longer—so much so that the expected value of this interval is 20 minutes! Thus the halfway point comes at 10 minutes, consistent with the analysis which appealed to the memoryless property, thus resolving the "paradox."

In other words, if we throw a dart at the wall, say, 1000 times, the mean of the 1000 intervals we would hit would be about 20. This in contrast to the mean of all of the intervals on the wall, which would be 10.

8.4.2 Probability Mass Functions and Densities in Length-Biased Sampling

Actually, we can intuitively reason out what the density is of the length of the particular inter-bus interval that we hit, as follows.

First consider the bag-of-sticks example, and suppose (somewhat artificially) that stick length X is a discrete random variable. Let Y denote the length of the stick that we pick by randomly touching a stick in the bag.

Again, note carefully that for the reasons we've been discussing here, the distributions of X and Y are different. Say we have a list of all sticks, and we choose a stick at random from the list. Then the length of that stick will be X. But if we choose by touching a stick in the bag, that length will be Y.

Now suppose that, say, stick lengths 2 and 6 each comprise 10% of the sticks in the bag, i.e.

$$p_X(2) = p_X(6) = 0.1 \tag{8.36}$$

Intuitively, one would then reason that

$$p_Y(6) = 3p_Y(2) \tag{8.37}$$

In other words, even though the sticks of length 2 are just as numerous as those of length 6, the latter are three times as long, so they should have triple the chance of being chosen. So, the chance of our choosing a stick of length j depends not only on $p_X(j)$ but also on j itself.

We could write that formally as

$$p_Y(j) \propto jp_X(j) \tag{8.38}$$

where $\propto$ is the "is proportional to" symbol. Thus

$$p_Y(j) = cjp_X(j) \tag{8.39}$$

for some constant of proportionality c.

But a probability mass function must sum to 1. So, summing over all possible values of j (whatever they are), we have

$$1 = \sum_j p_Y(j) = \sum_j cjp_X(j) \tag{8.40}$$

That last term is c E(X)! So, c = 1/EX, and

$$p_Y(j) = \frac{1}{EX} \cdot jp_X(j) \tag{8.41}$$

The continuous analog of (8.41) is

$$f_Y(t) = \frac{1}{EX} \cdot tf_X(t) \tag{8.42}$$

So, for our bus example, in which $f_X(t) = 0.1e^{-0.1t}, \ t > 0$ and EX = 10,

$$f_Y(t) = 0.01te^{-0.1t} \tag{8.43}$$

You may recognize this as an Erlang density with r = 2 and $\lambda = 0.1$. That distribution does indeed have mean 20, consistent with the discussion at the end of Section 8.4.1.

Chapter 9

Introduction to Confidence Intervals

The idea of a confidence interval is central to statistical inference. But actually, you already know about it—from the term *margin of error* in news reports about opinion polls.

9.1 The "Margin of Error" and Confidence Intervals

To explain the idea of margin of error, let's begin with a problem that has gone unanswered so far:

In our simulations in previous units, it was never quite clear how long the simulation should be run, i.e. what value to set for **nreps** in Section 2.14.7. Now we will finally address this issue.

As our example, consider the Bus Paradox, which presented in Section 8.4: Buses arrive at a certain bus stop at random times, with interarrival times being independent exponentially distributed random variables with mean 10 minutes. You arrive at the bus stop every day at a certain time, say four hours (240 minutes) after the buses start their morning run. What is your mean wait μ for the next bus?

We found mathematically that, due to the memoryless property of the exponential distribution, our wait is again exponentially distributed with mean 10. But suppose we didn't know that, and we wished to find the answer via simulation. (Note to reader: Keep in mind throughout this example that we will be pretending that we don't know the mean wait is actually 10. Reminders of this will be brought up occasionally.)

We could write a program to do this:

```
doexpt <- function(opt) {
   lastarrival <- 0.0
   while (lastarrival < opt)
```

```
      lastarrival <- lastarrival + rexp(1,0.1)
   return(lastarrival-opt)
}

observationpt <- 240
nreps <- 1000
waits <- vector(length=nreps)
for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
cat("approx. mean wait = ",mean(waits),"\n")
```

Running the program yields

```
approx. mean wait = 9.653743
```

Note that μ is a population mean, where our "population" here is the set of all possible bus wait times (some more frequent than others). Our simulation, then, drew a sample of size 1000 from that population. The expression **mean(waits)** was our sample mean.

Now, was 1000 iterations enough? How close is this value 9.653743 to the true expected value of waiting time?[1]

What we would like to do is something like what the pollsters do during presidential elections, when they say "Ms. X is supported by 62% of the voters, with a margin of error of 4%." In other words, we want to be able to attach a margin of error to that figure of 9.653743 above. We do this in the next section.

9.2 Confidence Intervals for Means

We are now set to make use of the infrastructure that we've built up in the preceding sections of this chapter. Everything will hinge on understanding that the sample mean is a random variable, with a known approximate distribution.

The goal of this section (and several that follow) is to develop a notion of margin of error, just as you see in the election campaign polls. This raises two questions:

(a) What do we mean by "margin of error"?

(b) How can we calculate it?

[1] Of course, continue to ignore the fact that we know that this value is 10.0. What we're trying to do here is figure out how to answer "how close is it" questions in general, when we don't know the true mean.

9.2.1 Basic Formulation

So, suppose we have a random sample $W_1, ..., W_n$ from some population with mean μ and variance σ^2.

Recall that (**??**) has an approximate N(0,1) distribution. We will be interested in the central 95% of the distribution N(0,1). Due to symmetry, that distribution has 2.5% of its area in the left tail and 2.5% in the right one. Through the R call **qnorm(0.025)**, or by consulting a N(0,1) cdf table in a book, we find that the cutoff points are at -1.96 and 1.96. In other words, if some random variable T has a N(0,1) distribution, then $P(-1.96 < T < 1.96) = 0.95$.

Thus

$$0.95 \approx P\left(-1.96 < \frac{\overline{W} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \tag{9.1}$$

(Note the approximation sign.) Doing a bit of algebra on the inequalities yields

$$0.95 \approx P\left(\overline{W} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{W} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{9.2}$$

Now remember, not only do we not know μ, we also don't know σ. But we can estimate it, as we saw, via (**??**). One can show (the details will be given in Section **??**) that (9.2) is still valid if we substitute s for σ, i.e.

$$0.95 \approx P\left(\overline{W} - 1.96\frac{s}{\sqrt{n}} < \mu < \overline{W} + 1.96\frac{s}{\sqrt{n}}\right) \tag{9.3}$$

In other words, we are about 95% sure that the interval

$$(\overline{W} - 1.96\frac{s}{\sqrt{n}}, \overline{W} + 1.96\frac{s}{\sqrt{n}}) \tag{9.4}$$

contains μ. This is called a 95% **confidence interval** for μ. The quantity $1.96\frac{s}{\sqrt{n}}$ is the margin of error.

9.2.2 Example: Simulation Output

We could add this feature to our program in Section 9.1:

```
doexpt <- function(opt) {
   lastarrival <- 0.0
   while (lastarrival < opt)
      lastarrival <- lastarrival + rexp(1,0.1)
   return(lastarrival-opt)
}

observationpt <- 240
nreps <- 10000
waits <- vector(length=nreps)
for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
wbar <- mean(waits)
cat("approx. mean wait =",wbar,"\n")
s2 <- mean(waits^2) - wbar^2
s <- sqrt(s2)
radius <- 1.96*s/sqrt(nreps)
cat("approx. CI for EW =",wbar-radius,"to",wbar+radius,"\n")
```

When I ran this, I got 10.02565 for the estimate of EW, and got an interval of (9.382715, 10.66859). Note that the margin of error is the radius of that interval, about 1.29/2. We would then say, "We are about 95% confident that the true mean wait time is between 9.38 and 10.67."

What does this really mean? This question is of the utmost importance. We will devote an entire section to it, Section 9.3.

Note that our analysis here is approximate, based on the Central Limit Theorem, which was applicable because $\overline{W}$ involves a sum. We are making no assumption about the density of the population from which the W_i are drawn. However, if that population density itself is normal, then an exact confidence interval can be constructed. This will be discussed in Section **??**.

9.3 Meaning of Confidence Intervals

9.3.1 A Weight Survey in Davis

Consider the question of estimating the mean weight, denoted by μ, of all adults in the city of Davis. Say we sample 1000 people at random, and record their weights, with W_i being the weight of the i^{th} person in our sample.[2]

Now remember, we don't know the true value of that population mean, μ—again, that's why we are collecting the sample data, to estimate μ! Our estimate will be our sample mean, $\overline{\mathbf{W}}$. But we don't know how accurate that estimate might be. That's

[2]Do you like our statistical pun here? Typically an example like this would concern people's heights, not weights. But it would be nice to use the same letter for random variables as in Section 9.2, i.e. the letter W, so we'll have our example involve people's weights instead of heights. It works out neatly, because the word *weight* has the same sound as *wait*.

the reason we form the confidence interval, as a gauge of the accuracy of $\overline{\mathbf{W}}$ as an estimate of μ.

Say our interval (9.4) turns out to be (142.6,158.8). We say that we are about 95% confident that the mean weight μ of all adults in Davis is contained in this interval. **What does this mean?**

Say we were to perform this experiment many, many times, recording the results in a notebook: We'd sample 1000 people at random, then record our interval $(\overline{W} - 1.96\frac{s}{\sqrt{n}}, \overline{W} + 1.96\frac{s}{\sqrt{n}})$ on the first line of the notebook. Then we'd sample another 1000 people at random, and record what interval we got that time on the second line of the notebook. This would be a different set of 1000 people (though possibly with some overlap), so we would get a different value of $\overline{W}$ and so, thus a different interval; it would have a different center and a different radius. Then we'd do this a third time, a fourth, a fifth and so on.

Again, each line of the notebook would contain the information for a different random sample of 1000 people. There would be two columns for the interval, one each for the lower and upper bounds. And though it's not immediately important here, note that there would also be columns for W_1 through W_{1000}, the weights of our 1000 people, and columns for $\overline{W}$ and s.

Now here is the point: Approximately 95% of all those intervals would contain μ, the mean weight in the entire adult population of Davis. The value of μ would be unknown to us—once again, that's why we'd be sampling 1000 people in the first place—but it does exist, and it would be contained in approximately 95% of the intervals.

As a variation on the notebook idea, think of what would happen if you and 99 friends each do this experiment. Each of you would sample 1000 people and form a confidence interval. Since each of you would get a different sample of people, you would each get a different confidence interval. What we mean when we say the confidence level is 95% is that of the 100 intervals formed—by you and 99 friends—about 95 of them will contain the true population mean weight. Of course, you hope you yourself will be one of the 95 lucky ones! But remember, you'll never know whose intervals are correct and whose aren't.

Now remember, in practice we only take *one* sample of 1000 people. Our notebook idea here is merely for the purpose of understanding what we mean when we say that we are about 95% confident that one interval we form does contain the true value of μ.

9.3.2 More About Interpretation

Some statistics instructors give students the odd warning, "You can't say that the probability is 95% that μ is IN the interval; you can only say that the probability is 95% confident that the interval CONTAINS μ." This of course is nonsense. As any fool can see, the following two statements are equivalent:

- "μ is in the interval"
- "the interval contains μ"

So it is ridiculous to say that the first is incorrect. Yet many instructors of statistics say so.

Where did this craziness come from? Well, way back in the early days of statistics, some instructor was afraid that a statement like "The probability is 95% that μ is in the interval" would make it sound like μ is a random variable. Granted, that was a legitimate fear, because μ is not a random variable, and without proper warning, some learners of statistics might think incorrectly. The random entity is the interval (both its center and radius), not μ; $\overline{W}$ and s in (9.4) vary from sample to sample, so the interval is indeed the random object here, not μ.

So, it was reasonable for teachers to warn students not to think μ is a random variable. But later on, some misguided instructor must have then decided that it is incorrect to say "μ is in the interval," and others then followed suit. They continue to this day, sadly.

A variant on that silliness involves saying that one can't say "The probability is 95% that μ is in the interval," because μ is either in the interval or not, so that "probability" is either 1 or 0! That is equally mushy thinking.

Suppose, for example, that I go into the next room and toss a coin, letting it land on the floor. I return to you, and tell you the coin is lying on the floor in the next room. I know the outcome but you don't. What is the probability that the coin came up heads? To me that is 1 or 0, yes, but to you it is 50%, in any practical sense.

It is also true in the "notebook" sense. If I do this experiment many times—go to the next room, toss the coin, come back to you, go to the next room, toss the coin, come back to you, etc., one line of the notebook per toss—then in the long run 50% of the lines of the notebook have Heads in the Outcome column.

The same is true for confidence intervals. Say we conduct many, many samplings, one per line of the notebook, with a column labeled Interval Contains Mu. Unfortunately, we ourselves don't get to see that column, but it exists, and in the long run 95% of the entries in the column will be Yes.

Finally, there are those who make a distinction between saying "There is a 95% probability that..." and "We are 95% confident that..." That's silly too. What else could "95% confident" mean if not 95% probability?

Consider the experiment of tossing two fair dice. The probability is 34/36, or about 94%, that we get a total that is different from 2 or 12. As we toss the dice, what possible distinction could be made between saying, "The probability is 94% that we will get a total between 3 and 11" and saying, "We are 94% confident that we will get a total between 3 and 11"? The notebook interpretation supports both phrasings, really. The words *probability* and *confident* should not be given much weight here; remember the quote at the beginning of our Chapter 1:

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

9.4 Confidence Intervals for Proportions

So we know how to find confidence intervals for means. How about proportions?

9.4.1 Derivation

For example, in an election opinion poll, we might be interested in the proportion p of people in the entire population who plan to vote for candidate A.

We will estimate p by taking a random sample of n voters, and finding the *sample* proportion of voters who plan to vote for A. The latter is usually denoted $\widehat{p}$, pronounced "p-hat." (The symbol $\widehat{}$ is often used in statistics to mean "estimate of.")

Assign to each voter in the population a value of Y, 1 if he/she plans to vote for A, 0 otherwise. Let Y_i be the value of Y for the ith person in our sample. Then

$$\widehat{p} = \overline{Y} \tag{9.5}$$

where $\overline{Y}$ is the sample mean among the Y_i, and p is the population mean of Y.

So we are really working with means after all, and thus in order to get a confidence interval for p from $\widehat{p}$, we can use (9.4)! We have that an approximate 95% confidence interval for p is

$$\left(\widehat{p} - 1.96 s/\sqrt{n}, \widehat{p} + 1.96 s/\sqrt{n}\right) \tag{9.6}$$

where as before s^2 is the sample variance among the Y_i, defined in **??**.

But there's more, because we can exploit the fact that in this special case, each Y_i is either 1 or 0, in order to save ourselves a bit of computation, as follows:

Recalling the convenient form of s^2, (**??**), we have

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2 &(9.7)\\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i - \overline{Y}^2 &(9.8)\\
&= \overline{Y} - \overline{Y}^2 &(9.9)\\
&= \widehat{p} - \widehat{p}^2 &(9.10)
\end{aligned}
$$

Then (9.6) simplifies to

$$
\left(\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}, \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\right) \qquad (9.11)
$$

9.4.2 That n vs. n-1 Thing Again

Recall Section **??**, in which it was noted that this book's definition of the sample variance, (**??**), is a little at odds with the the way most books define it, (**??**). The above derivation sheds a bit more light on this topic.

In the way I've defined things here, I was consistent: I divided by n both in (**??**) and in (9.7). Yet most books divide by n-1 in the former case but by n in the latter case! Their version of (9.11) is exactly the same as mine, yet they use a different s in (9.4)—even though they too observe that the proportions case is just a special case of estimating means (as in (9.5)). So, another reason to divide by n in (**??**) is to be consistent.

Again, the difference is usually minuscule anyway, but conceptually it's important to understand. As noted earlier, the n-1 divisor is really just a historical accident.

9.4.3 Simulation Example Again

In our bus example above, suppose we also want our simulation to print out the (estimated) probability that one must wait longer than 6.4 minutes. As before, we'd also like a margin of error for the output.

We incorporate (9.11) into our program:

```
doexpt <- function(opt) {
   lastarrival <- 0.0
   while (lastarrival < opt)
```

```
      lastarrival <- lastarrival + rexp(1,0.1)
   return(lastarrival-opt)
}

observationpt <- 240
nreps <- 1000
waits <- vector(length=nreps)
for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
wbar <- mean(waits)
cat("approx. mean wait =",wbar,"\n")
s2 <- (mean(waits^2) - mean(wbar)^2)
s <- sqrt(s2)
radius <- 1.96*s/sqrt(nreps)
cat("approx. CI for EW =",wbar-radius,"to",wbar+radius,"\n")
prop <- length(waits[waits > 6.4]) / nreps
s2 <- prop*(1-prop)
s <- sqrt(s2)
radius <- 1.96*s/sqrt(nreps)
cat("approx. P(W > 6.4) =",prop,", with a margin of error of",radius,"\n")
```

When I ran this, the value printed out for $\widehat{p}$ was 0.54, with a margin of error of 0.03, thus an interval of (0.51,0.57). We would say, "We don't know the exact value of $P(W > 6.4)$, so we ran a simulation. The latter estimates this probability to be 0.54, with a 95% margin of error of 0.03."

9.4.4 Example: Davis Weights

Note again that this uses the same principles as our Davis weights example. Suppose we were interested in estimating the proportion of adults in Davis who weigh more than 150 pounds. Suppose that proportion is 0.45 in our sample of 1000 people. This would be our estimate $\widehat{p}$ for the population proportion p, and an approximate 95% confidence interval (9.11) for the population proportion would be (0.42,0.48). We would then say, "We are 95% confident that the true population proportion p of people who weigh over 150 pounds is between 0.42 and 0.48."

Note also that although we've used the word *proportion* in the Davis weights example instead of *probability*, they are the same. If I choose an adult at random from the population, the probability that his/her weight is more than 150 is equal to the proportion of adults in the population who have weights of more than 150.

And the same principles are used in opinion polls during presidential elections. Here p is the population proportion of people who plan to vote for the given candidate. This is an unknown quantity, which is exactly the point of polling a sample of people—to estimate that unknown quantity p. Our estimate is $\widehat{p}$, the proportion of people in our sample who plan to vote for the given candidate, and n is the number of people that we poll. We again use (9.11).

9.4.5 Interpretation

The same interpretation holds as before. Consider the examples in the last section:

- If each of you and 99 friends were to run the R program at the beginning of Section 9.4.4, you 100 people would get 100 confidence intervals for $P(W > 6.4)$. About 95 of you would have intervals that do contain that number.
- If each of you and 99 friends were to sample 1000 people in Davis and come up with confidence intervals for the true population proportion of people who weight more than 150 pounds, about 95 of you would have intervals that do contain that true population proportion.
- If each of you and 99 friends were to sample 1200 people in an election campaign, to estimate the true population proportion of people who will vote for candidate X, about 95 of you will have intervals that do contain this population proportion.

Of course, this is just a "thought experiment," whose goal is to understand what the term "95% confident" really means. In practice, we have just one sample and thus compute just one interval. But we say that the interval we compute has a 95% chance of containing the population value, since 95% of all intervals will contain it.

9.4.6 (Non-)Effect of the Population Size

Note that in both the Davis and election examples, it doesn't matter what the size of the population is. The approximate distribution of $\widehat{p}$ is N(p,p(1-p)/n), so the accuracy of $\widehat{p}$, depends only on p and n. So when people ask, "How a presidential election poll can get by with sampling only 1200 people, when there are more than 100,000,000 voters in the U.S.?" now you know the answer. (We'll discuss the question "Why 1200?" below.)

Another way to see this is to think of a situation in which we wish to estimate the probability p of heads for a certain coin. We toss the coin n times, and use $\widehat{p}$ as our estimate of p. Here our "population"—the population of all coin tosses—is infinite, yet it is still the case that 1200 tosses would be enough to get a good estimate of p.

9.4.7 Inferring the Number Polled

A news report tells us that in a poll, 54% of those polled supported Candidate A, with a 2.2% margin of error. Assuming that the methods here were used, with a 95% level of confidence, let's find the approximate number polled.

$$0.022 = 1.96 \times \sqrt{0.54 \cdot 0.46/n} \tag{9.12}$$

Solving, we find that n is approximately 1972.

9.4.8 Planning Ahead

Now, why do the pollsters often sample 1200 people?

First, note that the maximum possible value of $\widehat{p}(1-\widehat{p})$ is 0.25.[3] Then the pollsters know that their margin of error with n = 1200 will be at most $1.96 \times 0.5/\sqrt{1200}$, or about 3%, even before they poll anyone. They consider 3% to be sufficiently accurate for their purposes, so 1200 is the n they choose.

9.5 General Formation of Confidence Intervals from Approximately Normal Estimators

We would now like to move on to constructing confidence intervals for other settings than the case handled so far, estimation of a single population mean or proportion.

9.5.1 The Notion of a Standard Error

Suppose we are estimating some population quantity θ based on sample data $Y_1, ..., Y_n$. So far, our only examples have had θ as a population mean μ, a population proportion p. But we'll see other examples as things unfold in this and subsequent chapters.

Consider an estimate for θ, $\widehat{\theta}$, and suppose that the estimator is composed of some sum for which the Central Limit Theorem applies,[4], so that the approximate distribution of $\widehat{\theta}$ is normal with mean θ and some variance.

Ponder this sequence of points:

- $\widehat{\theta}$ is a random variable.
- Thus $\widehat{\theta}$ has a variance.

[3]Use calculus to find the maximum value of f(x) = x(1-x).

[4]Using more advanced tools (Section **??**), one can show approximate normality even in many nonlinear cases.

- Thus $\widehat{\theta}$ has a standard deviation η.
- Unfortunately, η is an unknown population quantity.
- But we may be able to estimate η from our sample data. Call that estimate $\widehat{\eta}$.
- We refer to $\widehat{\eta}$ as the *standard error* of $\widehat{\theta}$, or s.e.$\widehat{\theta}$.

Back in Section 9.2.1, we found a standard error for $\overline{W}$, the sample mean, using the following train of thought:

- $Var(\overline{W}) = \frac{\sigma^2}{n}$
- $\widehat{Var}(\overline{W}) = \frac{s^2}{n}$
- s.e.$(\overline{W}) = \frac{s}{\sqrt{n}}$

In many cases, deriving a standard error is more involved than the above, But the point is this:

> Suppose $\widehat{\theta}$ is a sample-based estimator of a population quantity θ, and that, due to being composed of sums or some other reason, $\widehat{\theta}$ is approximately normally distributed with mean θ, and some (possibly unknown) variance. Then the quantity
>
> $$\frac{\widehat{\theta} - \theta}{\text{s.e.}(\widehat{\theta})} \tag{9.13}$$
>
> has an approximate N(0,1) distribution.

9.5.2 Forming General Confidence Intervals

That means we can mimic the derivation that led to (9.4). As with (9.1), write

$$0.95 \approx P\left(-1.96 < \frac{\widehat{\theta} - \theta}{\text{s.e.}(\widehat{\theta})} < 1.96\right) \tag{9.14}$$

After going through steps analogous to those following (9.1), we find that an approximate 95% confidence interval for θ is

$$\widehat{\theta} \pm 1.96 \cdot \text{s.e.}(\widehat{\theta}) \tag{9.15}$$

In other words, the margin of error is 1.96 s.e.$(\widehat{\theta})$.

The standard error of the estimate is one of the most commonly-used quantities in statistical applications. You will encounter it frequently in the output of R, for instance, and in the subsequent portions of this book. Make sure you understand what it means and how it is used.

And note again that $\sqrt{\widehat{p}(1-\widehat{p})/n}$ is the standard error of $\widehat{p}$.

9.5.3 Standard Errors of Combined Estimators

Here is further chance to exercise your skills in the mailing tubes regarding variance.

Suppose we have two population values to estimate, ω and γ, and that we are also interested in the quantity $\omega + 2\gamma$. We'll estimate the latter with $\hat{\omega} + 2\hat{\gamma}$. Suppose the standard errors of $\hat{\omega}$ and $\hat{\gamma}$ turn out to be 3.2 and 8.8, respectively, and that the two estimators are independent and approximately normal.[5] Let's find the standard error of $\hat{\omega} + 2\hat{\gamma}$.

We know from the material surrounding (7.16) that $\hat{\omega}+2\hat{\gamma}$ has an approximately normal distribution with variance

$$Var(\hat{\omega}) + 2^2 Var(\hat{\gamma}) \tag{9.16}$$

Thus the standard error of $\hat{\omega} + 2\hat{\gamma}$ is

$$\sqrt{\cdot 3.2^2 + 2^2 \cdot 8.8^2} \tag{9.17}$$

Now that we know the standard error of $\hat{\omega} + 2\hat{\gamma}$, we can use it in (9.15). We add and subtract 1.96 times (9.17) to $\hat{\omega} + 2\hat{\gamma}$, and that is our interval.

In general, for constants a and b, an approximate 95% confidence interval for the population quantity $a\omega + b\gamma$ is

$$a\hat{\omega} + b\hat{\gamma} \pm 1.96 \sqrt{a^2 s.e.^2(\hat{\omega}) + b^2 s.e.^2(\hat{\gamma})} \tag{9.18}$$

We can go even further. If $\hat{\omega}$ and $\hat{\gamma}$ are not independent but have known covariance, we can use the methods of Chapter **??** to obtain a standard error for any linear combination of these two estimators.

[5] Technically, the term *standard error* is only used for approximately normal estimators anyway.

9.6 Confidence Intervals for Differences of Means or Proportions

9.6.1 Independent Samples

Suppose in our sampling of people in Davis we are mainly interested in the difference in weights between men and women. Let $\overline{X}$ and n_1 denote the sample mean and sample size for men, and let $\overline{Y}$ and n_2 for the women. Denote the population means and variances by μ_i and σ_i^2, i = 1,2. We wish to find a confidence interval for $\mu_1 - \mu_2$. The natural estimator for that quantity is $\overline{X} - \overline{Y}$.

So, how can we form a confidence interval for $\mu_1 - \mu_2$ using $\overline{X} - \overline{Y}$? Since the latter quantity is composed of sums, we can use (9.15) and (9.18). Here:

- $a = 1, \ b = -1$
- $\omega = \mu_1, \ \gamma = \mu_2$
- $\hat{\omega} = \overline{X}, \ \hat{\gamma} = \overline{Y}$

But we know from before that $s.e.(\overline{X} = s_1/\sqrt{n}$, where s_1^2 is the sample variance for the men,

$$s_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \overline{X})^2 \tag{9.19}$$

and similarly for $\overline{Y}$ and the women. So, we have

$$\text{s.e.}(\overline{X} - \overline{Y}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{9.20}$$

Thus (9.15) tells us that an approximate 95% confidence interval for $\mu_1 - \mu_2$ is

$$\left(\overline{X} - \overline{Y} - 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \overline{X} - \overline{Y} + 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) \tag{9.21}$$

What about confidence intervals for the difference in two population proportions $p_1 - p_2$? Recalling that in Section 9.4 we noted that proportions are special cases of means, we see that finding a confidence interval for the difference in two proportions is covered by (9.21). Here

- $\overline{X}$ reduces to $\widehat{p}_1$

- $\overline{Y}$ reduces to $\widehat{p}_2$
- s_1^2 reduces to $\widehat{p}_1(1-\widehat{p}_1)$
- s_2^2 reduces to $\widehat{p}_2(1-\widehat{p}_2)$

So, (9.21) reduces to

$$\widehat{p_1} - \widehat{p_2} \pm R \tag{9.22}$$

where the radius R is

$$1.96 \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \tag{9.23}$$

9.6.2 Example: Network Security Application

In a network security application, C. Mano *et al*[6] compare round-trip travel time for packets involved in the same application in certain wired and wireless networks. The data was as follows:

sample	sample mean	sample s.d.	sample size
wired	2.000	6.299	436
wireless	11.520	9.939	344

We had observed quite a difference, 11.52 versus 2.00, but could it be due to sampling variation? Maybe we have unusual samples? This calls for a confidence interval!

Then a 95% confidence interval for the difference between wireless and wired networks is

$$11.520 - 2.000 \pm 1.96\sqrt{\frac{9.939^2}{344} + \frac{6.299^2}{436}} = 9.52 \pm 1.22 \tag{9.24}$$

So you can see that there is a big difference between the two networks, even after allowing for sampling variation.

[6] RIPPS: Rogue Identifying Packet Payload Slicer Detecting Unauthorized Wireless Hosts Through Network Traffic Conditioning, C. Mano and a ton of other authors, ACM TRANSACTIONS ON INFORMATION SYSTEMS AND SECURITY, May 2007.

9.6.3 Dependent Samples

Note carefully, though, that a key point above was the independence of the two samples. By contrast, suppose we wish, for instance, to find a confidence interval for $\nu_1 - \nu_2$, the difference in mean heights in Davis of 15-year-old and 10-year-old children, and suppose our data consist of pairs of height measurements at the two ages on *the same children*. In other words, we have a sample of n children, and for the i^{th} child we have his/her height U_i at age 15 and V_i at age 10. Let $\overline{U}$ and $\overline{V}$ denote the sample means.

The problem is that the two sample means are not independent. If a child is taller than his/her peers at age 15, he/she was probably taller than them when they were all age 10. In other words, for each i, V_i and U_i are positively correlated, and thus the same is true for $\overline{V}$ and $\overline{U}$. Thus we cannot use (9.21).

As always, it is instructive to consider this in "notebook" terms. Suppose on one particular sample at age 10—one line of the notebook—we just happen to have a lot of big kids. Then $\overline{V}$ is large. Well, if we look at the same kids later at age 15, they're liable to be bigger than the average 15-year-old too. In other words, among the notebook lines in which $\overline{V}$ is large, many of them will have $\overline{U}$ large too.

Since $\overline{U}$ is approximately normally distributed with mean ν_1, about half of the notebook lines will have $\overline{U} > \nu_1$. Similarly, about half of the notebook lines will have $\overline{V} > \nu_2$. But the nonindependence will be reflected in MORE than one-fourth of the lines having both $\overline{U} > \nu_1$ and $\overline{V} > \nu_2$. (If the two sample means were 100% correlated, that fraction would be 1.0.)

Contrast that with a sample scheme in which we sample some 10-year-olds and some 15-year-olds, say at the same time. Now *there are different kids in each of the two samples.* So, if by happenstance we get some big kids in the first sample, that has no impact on which kids we get in the second sample. In other words, $\overline{V}$ and $\overline{U}$ will be independent. In this case, one-fourth of the lines will have both $\overline{U} > \nu_1$ and $\overline{V} > \nu_2$.

So, we cannot get a confidence interval for $\nu_1 - \nu_2$ from (9.21), since the latter assumes that the two sample means are independent. What to do?

The key to the resolution of this problem is that the random variables $T_i = V_i - U_i$, i = 1,2,...,n are still independent. Thus we can use (9.4) on these values, so that our approximate 95% confidence interval is

$$(\overline{T} - 1.96\frac{s}{\sqrt{n}}, \overline{T} + 1.96\frac{s}{\sqrt{n}}) \tag{9.25}$$

where $\overline{T}$ and s^2 are the sample mean and sample variance of the T_i.

A common situation in which we have dependent samples is that in which we are comparing two dependent proportions. Suppose for example that there are three candidates running for a political office, A, B and C. We poll 1,000 voters and ask whom they plan to vote for. Let p_A, p_B and p_C be the three population proportions of people planning to vote for the various candidates, and let $\widehat{p}_A$, $\widehat{p}_B$ and $\widehat{p}_C$ be the corresponding sample proportions.

Suppose we wish to form a confidence interval for $p_A - p_B$. Clearly, the two sample proportions are not independent random variables, since for instance if $\widehat{p}_A = 1$ then we know for sure that $\widehat{p}_B$ is 0.

Or to put it another way, define the indicator variables U_i and V_i as above, with for example U_i being 1 or 0, according to whether the i^{th} person in our sample plans to vote for A or not, with V_i being defined similarly for B. Since U_i and V_i are "measurements" on *the same person*, they are not independent, and thus $\widehat{p}_A$ and $\widehat{p}_B$ are not independent either.

Note by the way that while the two sample means in our kids' height example above were positively correlated, in this voter poll example, the two sample proportions are negatively correlated.

So, we cannot form a confidence interval for $p_A - p_B$ by using (9.22). What can we do instead?

We'll use the fact that the vector $(N_A, N_B, N_C)^T$ has a multinomial distribution, where N_A, N_B and N_C denote the numbers of people in our sample who state they will vote for the various candidates (so that for instance $\widehat{p}_A = N_A/1000$).

Now to compute $Var(\widehat{p}_A - \widehat{p}_B)$, we make use of (**??**):

$$Var(\widehat{p}_A - \widehat{p}_B) = Var(\widehat{p}_A) + Var(\widehat{p}_B) - 2Cov(\widehat{p}_A, \widehat{p}_B) \tag{9.26}$$

Or, we could have taken a matrix approach, using (**??**) with A equal to the row vector (1,-1,0).

So, using (**??**), the standard error of $\widehat{p}_A - \widehat{p}_B$ is

$$\sqrt{0.001\widehat{p}_A(1-\widehat{p}_A) + 0.001\widehat{p}_B(1-\widehat{p}_B) + 0.002\widehat{p}_A\widehat{p}_B} \tag{9.27}$$

9.6.4 Example: Machine Classification of Forest Covers

Remote sensing is machine classification of type from variables observed aerially, typically by satellite. The application we'll consider here involves forest cover type for a given location; there are seven different types. (See Blackard, Jock A. and Denis J. Dean, 2000, "Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables," *Computers and Electronics in Agriculture*, 24(3):131-151.) Direct observation of the cover type is either too expensive or may suffer from land access permission issues. So, we wish to guess cover type from other variables that we can more easily obtain.

One of the variables was the amount of hillside shade at noon, which we'll call HS12. *Here's our goal:* Let μ_1 and μ_2 be the population mean HS12 among sites having cover types 1 and 2, respectively. If $\mu_1 - \mu_2$ is large, then HS12 would be a good predictor of whether the cover type is 1 or 2.

So, we wish to estimate $\mu_1 - \mu_2$ from our data, in which we do know cover type. There were over 50,000 observations, but for simplicity we'll just use the first 1,000 here. Let's find an approximate 95% confidence interval for $\mu_1 - \mu_2$. The two sample means were 223.8 and 226.3, with s values of 15.3 and 14.3, and the sample sizes were 226 and 585.

Using (9.21), we have that the interval is

$$223.8 - 226.3 \pm 1.96\sqrt{\frac{15.3^2}{226} + \frac{14.3^2}{585}} = -2.5 \pm 2.3 = (-4.8, -0.3) \tag{9.28}$$

Given that HS12 values are in the 200 range (see the sample means), this difference between them actually is not very large. This is a great illustration of an important principle, it will turn out in Section 10.11.

As another illustration of confidence intervals, let's find one for the difference in population proportions of sites that have cover types 1 and 2. Our sample estimate is

$$\widehat{p}_1 - \widehat{p}_2 = 0.226 - 0.585 = -0.359 \tag{9.29}$$

The standard error of this quantity, from (9.27), is

$$\sqrt{0.001 \cdot 0.226 \cdot 0.7740.001 \cdot 0.585 \cdot 0.415 + 002 \cdot 0.226 \cdot 0.585} = 0.019 \tag{9.30}$$

That gives us a confidence interval of

$$-0.359 \pm 1.96 \cdot 0.019 = (-0.397, -0.321) \tag{9.31}$$

Chapter 10

Introduction to Significance Tests

Suppose (just for fun, but with the same pattern as in more serious examples) you have a coin that will be flipped at the Super Bowl to see who gets the first kickoff. (We'll assume slightly different rules here. The coin is not "called." Instead, it is agreed beforehand that if the coin comes up heads, Team A will get the kickoff, and otherwise it will be Team B.) You want to assess for "fairness." Let p be the probability of heads for the coin.

You could toss the coin, say, 100 times, and then form a confidence interval for p using (9.11). The width of the interval would tell you the margin of error, i.e. it tells you whether 100 tosses were enough for the accuracy you want, and the location of the interval would tell you whether the coin is "fair" enough.

For instance, if your interval were (0.49,0.54), you might feel satisfied that this coin is reasonably fair. In fact, **note carefully that even if the interval were, say, (0.502,0.506), you would still consider the coin to be reasonably fair**; the fact that the interval did not contain 0.5 is irrelevant, as the entire interval would be reasonably near 0.5.

However, this process would not be the way it's traditionally done. Most users of statistics would use the toss data to test the **null hypothesis**

$$H_0 : p = 0.5 \tag{10.1}$$

against the **alternate hypothesis**

$$H_A : p \neq 0.5 \tag{10.2}$$

For reasons that will be explained below, this procedure is called **significance testing**. It forms the very core of statistical inference as practiced today. This, however, is unfortunate, as there are some serious problems that have been recognized with this procedure. We will first discuss the mechanics of the procedure, and then look closely at the problems with it in Section 10.11.

10.1 The Basics

Here's how significance testing works.

The approach is to consider H_0 "innocent until proven guilty," meaning that we assume H_0 is true unless the data give strong evidence to the contrary. **KEEP THIS IN MIND!**—we are continually asking, "What if...?"

The basic plan of attack is this:

> We will toss the coin n times. Then we will believe that the coin is fair unless the number of heads is "suspiciously" extreme, i.e. much less than n/2 or much more than n/2.

Let p denote the true probability of heads for our coin. As in Section 9.4.1, let $\widehat{p}$ denote the proportion of heads in our sample of n tosses. We observed in that section that $\widehat{p}$ is a special case of a sample mean (it's a mean of 1s and 0s). We also found that the standard deviation of $\widehat{p}$ is $\sqrt{p(1-p)/n}$, so that the standard error (Section 9.5.1) is $\sqrt{\widehat{p}(1-\widehat{p})/n}$.

In other words, the material surrounding (9.13) tells us that

$$\frac{\widehat{p}-p}{\sqrt{\frac{1}{n}\cdot\widehat{p}(1-\widehat{p})}} \tag{10.3}$$

has an approximate N(0,1) distribution.

But remember, we are going to assume H_0 for now, until and unless we find strong evidence to the contrary. So, in the numerator of (10.3), p = 0.5. And in the denominator, why use the standard error when we "know" (under our provisional assumption) that p = 0.5? The exact standard deviation of $\widehat{p}$ is $\sqrt{p(1-p)/n}$. So, replace (10.3) by

$$Z = \frac{\widehat{p}-0.5}{\sqrt{\frac{1}{n}\cdot 0.5(1-0.5)}} \tag{10.4}$$

then Z has an approximate N(0,1) distribution under the assumption that H_0 is true.

Now recall from the derivation of (9.4) that -1.96 and 1.96 are the lower- and upper-2.5% points of the N(0,1) distribution. Thus,

$$P(Z < -1.96 \text{ or } Z > 1.96) \approx 0.05 \tag{10.5}$$

Now here is the point: After we collect our data, in this case by tossing the coin n times, we compute $\widehat{p}$ from that data, and then compute Z from (10.4). If Z is smaller than -1.96 or larger than 1.96, we reason as follows:

> Hmmm, Z would stray that far from 0 only 5% of the time. So, either I have to believe that a rare event has occurred, or I must abandon my assumption that H_0 is true.

For instance, say n = 100 and we get 62 heads in our sample. That gives us Z = 2.4, in that "rare" range. We then **reject** H_0, and announce to the world that this is an unfair coin. We say, "The value of p is significantly different from 0.5."

The 5% "suspicion criterion" used above is called the **significance level**, typically denoted α. One common statement is "We rejected H_0 at the 5% level."

On the other hand, suppose we get 47 heads in our sample. Then Z = -0.60. Again, taking 5% as our significance level, this value of Z would not be deemed suspicious, as it occurs frequently. We would then say "We accept H_0 at the 5% level," or "We find that p is not significantly different from 0.5."

The word *significant* is misleading. It should NOT be confused with *important*. It simply is saying we don't believe the observed value of Z is a rare event, which it would be under H_0; we have instead decided to abandon our believe that H_0 is true.

Note by the way that Z values of -1.96 and 1.96 correspond getting $50 - 1.96 \cdot 0.5 \cdot \sqrt{100}$ or $50 + 1.96 \cdot 0.5 \cdot \sqrt{100}$ heads, i.e. roughly 40 or 60. In other words, we can describe our rejection rule to be "Reject if we get fewer than 40 or more than 60 heads, out of our 100 tosses."

10.2 General Testing Based on Normally Distributed Estimators

In Section 9.5, we developed a method of constructing confidence intervals for general approximately normally distributed estimators. Now we do the same for significance testing.

Suppose $\widehat{\theta}$ is an approximately normally distributed estimator of some population value θ. Then to test $H_0 : \theta = c$, form the test statistic

$$Z = \frac{\widehat{\theta} - c}{s.e.(\widehat{\theta})} \tag{10.6}$$

where $s.e.(\widehat{\theta})$ is the standard error of $\widehat{\theta}$,[1] and proceed as before:

> Reject $H_0 : \theta = c$ at the significance level of $\alpha = 0.05$ if $|Z| \geq 1.96$.

10.3 Example: Network Security

Let's look at the network security example in Section 9.6.2 again. Here $\widehat{\theta} = \overline{X} - \overline{Y}$, and c is presumably 0 (depending on the goals of Mano *et al*). From 9.20, the standard error works out to 0.61. So, our test statistic (10.6) is

$$Z = \frac{\overline{X} - \overline{Y} - 0}{0.61} = \frac{11.52 - 2.00}{0.61} = 15.61 \tag{10.7}$$

This is definitely larger in absolute value than 1.96, so we reject H_0, and conclude that the population mean round-trip times are different in the wired and wireless cases.

10.4 The Notion of "p-Values"

Recall the coin example in Section 10.1, in which we got 62 heads, i.e. Z = 2.4. Since 2.4 is considerably larger than 1.96, our cutoff for rejection, we might say that in some sense we not only rejected H_0, we actually strongly rejected it.

To quantify that notion, we compute something called the **observed significance level**, more often called the **p-value**.

We ask,

> We rejected H_0 at the 5% level. Clearly, we would have rejected it even at some small—thus more stringent—levels. What is the smallest such level? Call this the p-value of the test.

[1] See Section 9.5. Or, if we know the exact standard deviation of $\widehat{\theta}$ under H_0, which was the case in our coin example above, we could use that, for a better normal approximation.

By checking a table of the N(0,1) distribution, or by calling **pnorm(2.40)** in R, we would find that the N(0,1) distribution has area 0.008 to the right of 2.40, and of course by symmetry there is an equal area to the left of -2.40. That's a total area of 0.016. In other words, we would have been able to reject H_0 even at the much more stringent significance level of 0.016 (the 1.6% level) instead of 0.05. So, Z = 2.40 would be considered even more significant than Z = 1.96. In the research community it is customary to say, "The p-value was 0.016." [2] The smaller the p-value, the more significant the results are considered.

In our network security example above in which Z was 15.61, the value is literally "off the chart"; **pnorm(15.61)** returns a value of 1. Of course, it's a tiny bit less than 1, but it is so far out in the right tail of the N(0,1) distribution that the area to the right is essentially 0. So the p-value would be essentially 0, and the result would be treated as very, very highly significant.

In computer output or research reports, we often see small p-values being denoted by asterisks. There is generally one asterisk for p under 0.05, two for p less than 0.01, three for 0.001, etc. The more asterisks, the more significant the data is supposed to be. See for instance the R regression output on page 214.

10.5 Example: Bank Data

Consider again the bank marketing data in Section **??**. Our comparison was between marketing campaign success rates for married and unmarried customers. The p-value was quite tiny, 2.2×10^{-16}, but be careful interpreting this.

First, don't take that p-value as exact by any means. Though our sample sizes are certainly large enough for the Central Limit Theorem to work well, that is in the heart of the distribution, not the far tails. So, just take the p-value as "tiny," and leave it at that.

Second, although the standard description for a test with such a small p-value is "very highly significant," keep in mind that the difference between the two groups was not that large. The confidence interval we are 95% confident that the population success rate is between 3.3% and 4.6% less for married people. That is an interesting difference and possibly of some use to the marketing people, but it is NOT large.

[2]The 'p' in "p-value" of course stands for "probability," meaning the probably that a N(0,1) random variable would stray as far, or further, from 0 as our observed Z here. By the way, be careful not to confuse this with the quantity p in our coin example, the probability of heads.

10.6 One-Sided H_A

Suppose that—somehow—we are sure that our coin in the example above is either fair or it is more heavily weighted towards heads. Then we would take our alternate hypothesis to be

$$H_A : p > 0.5 \tag{10.8}$$

A "rare event" which could make us abandon our belief in H_0 would now be if Z in (10.4) is very large in the positive direction. So, with $\alpha = 0.05$, we call **qnorm(0.95)**, and find that our rule would now be to reject H_0 if $Z > 1.65$.

One-sided tests are not common, as their assumptions are often difficult to justify.

10.7 Exact Tests

Remember, the tests we've seen so far are all approximate. In (10.4), for instance, $\widehat{p}$ had an approximate normal distribution, so that the distribution of Z was approximately N(0,1). Thus the significance level α was approximate, as were the p-values and so on.[3]

But the only reason our tests were approximate is that we only had the *approximate* distribution of our test statistic Z, or equivalently, we only had the approximate distribution of our estimator, e.g. $\widehat{p}$. If we have an *exact* distribution to work with, then we can perform an exact test.

10.7.1 Example: Test for Biased Coin

Let's consider the coin example again, with the one-sided alternative (10.8) . To keep things simple, let's suppose we toss the coin 10 times. We will make our decision based on X, the number of heads out of 10 tosses. Suppose we set our threshhold for "strong evidence" again H_0 to be 8 heads, i.e. we will reject H_0 if $X \geq 8$. What will α be?

$$\alpha = \sum_{i=8}^{10} P(X = i) = \sum_{i=8}^{10} \binom{10}{i} \left(\frac{1}{2}\right)^{10} = 0.055 \tag{10.9}$$

That's not the usual 0.05. Clearly we cannot get an exact significance level of 0.05,[4] but our α is

[3] Another class of probabilities which would be approximate would be the **power** values. These are the probabilities of rejecting H_0 if the latter is not true. We would speak, for instance, of the power of our test at p = 0.55, meaning the chances that we would reject the null hypothesis if the true population value of p were 0.55.

[4] Actually, it could be done by introducing some randomization to our test.

exactly 0.055, so this is an exact test.

So, we will believe that this coin is perfectly balanced, unless we get eight or more heads in our 10 tosses. The latter event would be very unlikely (probability only 5.5%) if H_0 were true, so we decide not to believe that H_0 is true.

10.7.2 Example: Improved Light Bulbs

Suppose lifetimes of lightbulbs are exponentially distributed with mean μ. In the past, $\mu = 1000$, but there is a claim that the new light bulbs are improved and $\mu > 1000$. To test that claim, we will sample 10 lightbulbs, getting lifetimes $X_1, ..., X_{10}$, and compute the sample mean $\overline{X}$. We will then perform a significance test of

$$H_0 : \mu = 1000 \tag{10.10}$$

vs.

$$H_A : \mu > 1000 \tag{10.11}$$

It is natural to have our test take the form in which we reject H_0 if

$$\overline{X} > w \tag{10.12}$$

for some constant w chosen so that

$$P(\overline{X} > w) = 0.05 \tag{10.13}$$

under H_0. Suppose we want an exact test, not one based on a normal approximation.

Remember, we are making our calculations under the assumption that H_0 is true. Now recall (Section 6.6.4.1) that $10\overline{X}$, the sum of the X_i, has a gamma distribution, with r = 10 and $\lambda = 0.001$. So, we can find the w for which $P(\overline{X} > w) = 0.05$ by using R's **qgamma()**:

```
> qgamma(0.95,10,0.001)
[1] 15705.22
```

So, we reject H_0 if our sample mean is larger than 1570.5.

Now suppose it turns out that $\overline{X} = 1624.2$. Under H_0 there was only a 0.05 chance that $\overline{X}$ would exceed 1570.5, so we would reject H_0 with $\alpha = 0.05$. But what if we had set w to 1624.2? We didn't do so, of course, but what if? The computation

```
> 1 - pgamma(16242,10,0.001)
[1] 0.03840629
```

shows that we would have rejected H_0 even if we had originally set α to the more stringent criterion of 0.038 instead of 0.05. So we report that the p-value was 0.038.

The idea of a p-value is to indicate in our report "how strongly" we rejected H_0. Arguably there is a bit of game-playing in p-values, as there is with significance testing in general. This will be pursued in Section 10.11.

10.7.3 Example: Test Based on Range Data

Suppose lifetimes of some electronic component formerly had an exponential distribution with mean 100.0. However, it's claimed that now the mean has increased. (Suppose we are somehow sure it has not decreased.) Someone has tested 50 of these new components, and has recorded their lifetimes, $X_1, ..., X_{50}$. Unfortunately, they only reported to us the range of the data, $R = \max_i X_i - \min_i X_i$, not the individual X_i. We will need to do a significance test with this limited data, at the 0.05 level.

Recall that the variance of an exponential random variable is the square of its mean. Intutively, then, the larger this population mean of X, the larger the mean of the range R. In other words, the form of the test should be to reject H_0 if R is greater than some cutoff value c. So, we need to find the value of c to make α equal to 0.05.

Unfortunately, we can't do this analytically, i.e. mathematically, as the distribution of R is far too complex. This we'll have to resort to simulation.[5] Here is code to do that:

```
# code to determine the cutoff point for significance
# at 0.05 level

nreps <- 200000
n <- 50

rvec <- vector(length=nreps)
for (i in 1:nreps) {
   x <- rexp(n,0.01)
```

[5] I am still referring to the following as an exact test, as we are not using any statistical approximation, such as the Central Limit Theorm.

```
   rng <- range(x)
   rvec[i] <- rng[2] - rng[1]
}

rvec <- sort(rvec)
cutoff <- rvec[ceiling(0.95*nreps)]
cat("reject H0 if R >",rvec[cutoff],"\n")
```

Here we general **nreps** samples of size 50 from an exponential distribution having mean 100. Note that since we are setting α, a probability defined in the setting in which H_0 is true, we assume the mean is 100. For each of the **nreps** samples we find the value of R, recording it in **rvec**. We then take the 95^{th} percentile of those values, which is the c for which P(R > c) = 0.05.[6]

The value of c output by the code was 220.4991. A second run yielded, 220.9304, and a third 220.7099. The fact that these values varied little among themselves indicates that our value of **nreps**, 200000, was sufficiently large.

10.7.4 Exact Tests under a Normal Distribution Assumption

If you are willing to assume that you are sampling from a normally-distributed population, then the Student-t test is nominally exact. The R function **t.test()** performs this operation, with the argument **alternative** set to be either **"less"** or **"greater"**.

10.8 Don't Speak of "the Probability That H_0 Is True"

It is very important to understand that throughout this chapter, we cannot speak of "the probability that H_0 is true," because we have no probablistic structure on H_0.

Consider the fair-coin problem at the beginning of this chapter. Suppose we hope to make a statement like, say, "Given that we got 62 heads out of 100 tosses, we find that the probability that this is a fair coin is 0.04." What kind of derivation would need to go into this? It would go along the following lines:

[6] Of course, this is approximate. The greater the value of **nreps**, the better the approximation.

$$
\begin{aligned}
P(H_0 \text{ is true} \mid \text{our data}) &= \frac{P(H_0 \text{ is true and our data})}{P(\text{our data})} \\
&= \frac{P(H_0 \text{ is true and our data})}{P(H_0 \text{ is true and our data}) + P(H_0 \text{ is false and our data})} \\
&= \frac{P(H_0 \text{ true})\ P(\text{our data} \mid H_0 \text{ true})}{P(H_0 \text{ true})\ P(\text{our data} \mid H_0 \text{ true}) + P(H_0 \text{ false})\ P(\text{our data} \mid H_0 \text{ false})}
\end{aligned}
\tag{10.14}
$$

Through our modeling process, e.g. the discussion surrounding (10.4), we can calculate $P(\text{our data} \mid H_0 \text{ is true})$. (The false case would be more complicated, since there are many different kinds of false cases here, for different values of p, but could be handled similarly.) But what we don't have is $P(H_0 \text{ is true})$.

We could certainly try to model that latter quantity, say by taking a sample of all possible pennies (if our coin is a penny), doing very extensive testing of them;[7] the proportion found to be fair would then be $P(H_0 \text{ is true})$. But lacking that, we have no probablistic structure for $P(H_0 \text{ is true})$, and thus cannot use language like "the probability that H_0 is true,"

10.9 R Computation

The R function **t.test()**, discussed in Section **??**, does both confidence intervals and tests, including p-values in the latter case.

10.10 The Power of a Test

In addition to the significance level of a test, we may also be interested in its **power** (or its many power values, as will be seen).

10.10.1 Example: Coin Fairness

For example, consider our first example in this chapter, in which we were testing a coin for fairness (Section 10.1). Our rule for a test at a 0.05 significance level turned out to be that we reject H_0 if we get fewer than 40 or more than 60 heads out of our 100 tosses. We might ask the question, say:

> Suppose the true heads probability is 0.65. We don't know, of course, but what if that were the case. That's a pretty substantial departure from H_0, so hopefully we would reject. Well, what is the probability that we would indeed reject?

[7] Say, 100,000 tosses per coin.

We could calculate this. Let N denote the number of heads. Then the desired probability is $P(N < 40 \text{ or } N > 60) = P(N < 40) + P(N > 60)$. Let's find the latter.[8]

Once again, since N has a binomial distribution, it is approximately normal, in this case with mean $np = 100 \times 0.65 = 65$ and variance $np(1-p) = 100 \times 0.65 \times 0.35 = 22.75$. Then $P(N > 60)$ is about

```
1 - pnorm(60,65,sqrt(22.75))
```

or about 0.85. So we would be quite likely to decide this is an unfair coin if (unknown to us) the true probability of heads is 0.65.

We say that the power of this test *at* p = 0.65 is 0.85. There is a different power for each p.

10.10.2 Example: Improved Light Bulbs

Let's find the power of the test in Section 10.7.2, at $\mu = 1250$. Recall that we reject H_0 if $\overline{X} > 1570.522$. Thus our power is

```
1 - pgamma(15705.22,10,1/1250)
```

This turns out to be about 0.197. So, if (remember, this is just a "what if?") the true new mean were 1250, we'd only have about a 20% chance of discovering that the new bulbs are improved.

10.11 What's Wrong with Significance Testing—and What to Do Instead

The first principle is that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists.—Richard Feynman, Nobel laureate in physics

"Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path"—Paul Meehl, professor of psychology and the philosophy of science

Significance testing is a time-honored approach, used by tens of thousands of people every day. But it is "wrong." I use the quotation marks here because, although significance testing is mathematically correct, it is at best noninformative and at worst seriously misleading.

[8] The former would be found similarly, but would come out quite small.

10.11.1 History of Significance Testing, and Where We Are Today

We'll see why significance testing has serious problems shortly, but first a bit of history.

When the concept of significance testing, especially the 5% value for α, was developed in the 1920s by Sir Ronald Fisher, many prominent statisticians opposed the idea—for good reason, as we'll see below. But Fisher was so influential that he prevailed, and thus significance testing became the core operation of statistics.

So, significance testing became entrenched in the field, in spite of being widely recognized as faulty, to this day. Most modern statisticians understand this, even if many continue to engage in the practice.[9] Here are a few places you can read criticism of testing:

- There is an entire book on the subject, *The Cult of Statistical Significance*, by S. Ziliak and D. McCloskey. Interestingly, on page 2, they note the prominent people who have criticized testing. Their list is a virtual "who's who" of statistics, as well as physics Nobel laureate Richard Feynman and economics Nobelists Kenneth Arrow and Milton Friedman.

- See `http://www.indiana.edu/~stigtsts/quotsagn.html` for a nice collection of quotes from famous statisticians on this point.

- There is an entire chapter devoted to this issue in one of the best-selling elementary statistics textbooks in the nation.[10]

- The Federal Judicial Center, which is the educational and research arm of the federal court system, commissioned two prominent academics, one a statistics professor and the other a law professor, to write a guide to statistics for judges: *Reference Guide on Statistics.* David H. Kaye. David A. Freedman, at

  `http://www.fjc.gov/public/pdf.nsf/lookup/sciman02.pdf/$file/sciman02.pdf`

  There is quite a bit here on the problems of significance testing, and especially p.129.

10.11.2 The Basic Fallacy

To begin with, **it's absurd to test H_0 in the first place**, because we know *a priori* that H_0 is false.

[9] Many are forced to do so, e.g. to comply with government standards in pharmaceutical testing. My own approach in such situations is to quote the test results but then point out the problems, and present confidence intervals as well.

[10] *Statistics*, third edition, by David Freedman, Robert Pisani, Roger Purves, pub. by W.W. Norton, 1997.

Consider the coin example, for instance. No coin is absolutely perfectly balanced, and yet that is the question that significance testing is asking:

$$H_0 : p = 0.5000000000000000000000000000... \tag{10.15}$$

We know before even collecting any data that the hypothesis we are testing is false, and thus it's nonsense to test it.

But much worse is this word "significant." Say our coin actually has p = 0.502. From anyone's point of view, that's a fair coin! But look what happens in (10.4) as the sample size n grows. If we have a large enough sample, eventually the denominator in (10.4) will be small enough, and $\widehat{p}$ will be close enough to 0.502, that Z will be larger than 1.96 and we will declare that p is "significantly" different from 0.5. But it isn't! Yes, 0.502 is different from 0.5, but NOT in any significant sense in terms of our deciding whether to use this coin in the Super Bowl.

The same is true for government testing of new pharmaceuticals. We might be comparing a new drug to an old drug. Suppose the new drug works only, say, 0.4% (i.e. 0.004) better than the old one. Do we want to say that the new one is "signficantly" better? This wouldn't be right, especially if the new drug has much worse side effects and costs a lot more (a given, for a new drug).

Note that in our analysis above, in which we considered what would happen in (10.4) as the sample size increases, we found that eventually *everything* becomes "signficiant"—even if there is no practical difference. This is especially a problem in computer science applications of statistics, because they often use very large data sets. A data mining application, for instance, may consist of hundreds of thousands of retail purchases. The same is true for data on visits to a Web site, network traffic data and so on. In all of these, the standard use of significance testing can result in our pouncing on very small differences that are quite insignificant to us, yet will be declared "significant" by the test.

Conversely, if our sample is too small, we can miss a difference that actually *is* significant—i.e. important to us—and we would declare that p is NOT significantly different from 0.5. In the example of the new drug, this would mean that it would be declared as "not significantly better" than the old drug, even if the new one is much better but our sample size wasn't large enough to show it.

In summary, the basic problems with significance testing are

- H_0 is improperly specified. What we are really interested in here is whether p is *near* 0.5, not whether it is *exactly* 0.5 (which we know is not the case anyway).
- Use of the word *significant* is grossly improper (or, if you wish, grossly misinterpreted).

Significance testing forms the very core usage of statistics, yet you can now see that it is, as I said above, "at best noninformative and at worst seriously misleading." This is widely recognized by thinking statisticians and prominent scientists, as noted above. But the practice of significance testing is too deeply entrenched for things to have any prospect of changing.

10.11.3 You Be the Judge!

This book has been written from the point of view that every educated person should understand statistics. It impacts many vital aspects of our daily lives, and many people with technical degrees find a need for it at some point in their careers.

In other words, statistics is something to be *used*, not just learned for a course. You should think about it critically, especially this material here on the problems of significance testing. You yourself should decide whether the latter's widespread usage is justified.

10.11.4 What to Do Instead

Note carefully that I am not saying that we should not make a decision. We *do* have to decide, e.g. decide whether a new hypertension drug is safe or in this case decide whether this coin is "fair" enough for practical purposes, say for determining which team gets the kickoff in the Super Bowl. But it should be an informed decision, and even testing the modified H_0 above would be much less informative than a confidence interval.

In fact, the real problem with significance tests is that they **take the decision out of our hands**. They make our decision mechanically for us, not allowing us to interject issues of importance to us, such possible side effects in the drug case.

So, what can we do instead?

In the coin example, we could set limits of fairness, say require that p be no more than 0.01 from 0.5 in order to consider it fair. We could then test the hypothesis

$$H_0 : 0.49 \leq p \leq 0.51 \tag{10.16}$$

Such an approach is almost never used in practice, as it is somewhat difficult to use and explain. But even more importantly, what if the true value of p were, say, 0.51001? Would we still really want to reject the coin in such a scenario?

Forming a confidence interval is the far superior approach. The width of the interval shows us whether n is large enough for $\widehat{p}$ to be reasonably accurate, and the location of the interval tells us whether the coin is fair enough for our purposes.

Note that in making such a decision, we do NOT simply check whether 0.5 is in the interval. That would make the confidence interval reduce to a significance test, which is what we are trying to avoid. If for example the interval is (0.502,0.505), we would probably be quite satisfied that the coin is fair enough for our purposes, even though 0.5 is not in the interval.

On the other hand, say the interval comparing the new drug to the old one is quite wide and more or less equal positive and negative territory. Then the interval is telling us that the sample size just isn't large enough to say much at all.

Significance testing is also used for model building, such as for predictor variable selection in regression analysis (a method to be covered in Chapter 11). The problem is even worse there, because there is no reason to use $\alpha = 0.05$ as the cutoff point for selecting a variable. In fact, even if one uses significance testing for this purpose—again, very questionable—some studies have found that the best values of α for this kind of application are in the range 0.25 to 0.40, far outside the range people use in testing.

In model building, we still can and should use confidence intervals. However, it does take more work to do so. We will return to this point in our unit on modeling, Chapter **??**.

10.11.5 Decide on the Basis of "the Preponderance of Evidence"

I was in search of a one-armed economist, so that the guy could never make a statement and then say: "on the other hand"—President Harry S Truman

If all economists were laid end to end, they would not reach a conclusion—Irish writer George Bernard Shaw

In the movies, you see stories of murder trials in which the accused must be "proven guilty beyond the shadow of a doubt." But in most noncriminal trials, the standard of proof is considerably lighter, **preponderance of evidence**. This is the standard you must use when making decisions based on statistical data. Such data cannot "prove" anything in a mathematical sense. Instead, it should be taken merely as evidence. The width of the confidence interval tells us the likely accuracy of that evidence. We must then weigh that evidence against other information we have about the subject being studied, and then ultimately make a decision on the basis of the preponderance of all the evidence.

Yes, juries must make a decision. But they don't base their verdict on some formula. Similarly, you the data analyst should not base your decision on the blind application of a method that is usually of little relevance to the problem at hand—significance testing.

10.11.6 Example: the Forest Cover Data

In Section 9.6.4, we found that an approximate 95% confidence interval for $\mu_1 - \mu_2$ was

$$223.8 - 226.3 \pm 2.3 = (-4.8, -0.3) \tag{10.17}$$

Clearly, the difference in HS12 between cover types 1 and 2 is tiny when compared to the general size of HS12, in the 200s. Thus HS12 is not going to help us guess which cover type exists at a given location. Yet with the same data, we would reject the hypothesis

$$H_0 : \mu_1 = \mu_2 \tag{10.18}$$

and say that the two means are "significantly" different, which sounds like there is an important difference—which there is not.

10.11.7 Example: Assessing Your Candidate's Chances for Election

Imagine an election between Ms. Smith and Mr. Jones, with you serving as campaign manager for Smith. You've just gotten the results of a very small voter poll, and the confidence interval for p, the fraction of voters who say they'll vote for Smith, is (0.45,0.85). Most of the points in this interval are greater than 0.5, so you would be highly encouraged! You are certainly not sure of the final election result, as a small part of the interval is below 0.5, and anyway voters might change their minds between now and the election. But the results would be highly encouraging.

Yet a significance test would say "There is no significant difference between the two candidates. It's a dead heat." Clearly that is not telling the whole story. The point, once again, is that **the confidence interval is giving you much more information than is the significance test.**

Chapter 11

Linear Regression

In many senses, this chapter and several of the following ones form the real core of statistics, especially from a computer science point of view.

In this chapter and the next, we are interested in relations between variables, in two main senses:

- In **regression analysis**, we are interested in the relation of one variable with one or more others.
- In other kinds of analyses, such as **principal components analysis**, we are interested in relations among several variables, symmetrically, i.e. not having one variable play a special role.

Note carefully that *many types of methods that go by another name are actually regression methods.* Examples are the **classification problem**, **discriminant analysis**, **pattern recognition**, **machine learning** and so on. We'll return to this point in Chapter **??**.

11.1 The Goals: Prediction and Description

Prediction is difficult, especially when it's about the future.—Yogi Berra[1]

Before beginning, it is important to understand the typical goals in regression analysis.

[1]Yogi Berra (1925-2015) is a former baseball player and manager, famous for his malapropisms, such as "When you reach a fork in the road, take it"; "That restaurant is so crowded that no one goes there anymore"; and "I never said half the things I really said."

- **Prediction:** Here we are trying to predict one variable from one or more others.
- **Description:** Here we wish to determine which of several variables have a greater effect on (or relation to) a given variable. An important special case is that in which we are interested in determining the effect of one predictor variable, **after the effects of the other predictors are removed.**

Denote the **predictor variables** by, $X^{(1)}, ..., X^{(r)}$, alluding to the Prediction goal. They are also called **independent variables** or **explanatory variables** (the latter term highlighting the Description goal) The variable to be predicted, Y, is often called the **response variable**, or the **dependent variable**. Note that one or more of the variables—whether the predictors or the response variable—may be indicator variables (Section 3.9). Another name for response variables of that type is **dummy variables**.

Methodology for this kind of setting is called **regression analysis**. If the response variable Y is an indicator variable, the values 1 and 0 to indicate class membership, we call this the **classification problem**. (If we have more than two classes, we need several Ys.)

In the above context, we are interested in the relation of a single variable Y with other variables $X^{(i)}$. But in some applications, we are interested in the more symmetric problem of relations *among* variables $X^{(i)}$ (with there being no Y). A typical tool for the case of continuous random variables is **principal components analysis**, and a popular one for the discrete case is **log-linear model**; both will be discussed later in this chapter.

11.2 Example Applications: Software Engineering, Networks, Text Mining

Example: As an aid in deciding which applicants to admit to a graduate program in computer science, we might try to predict Y, a faculty rating of a student after completion of his/her first year in the program, from $X^{(1)}$ = the student's CS GRE score, $X^{(2)}$ = the student's undergraduate GPA and various other variables. Here our goal would be Prediction, but educational researchers might do the same thing with the goal of Description. For an example of the latter, see Predicting Academic Performance in the School of Computing & Information Technology (SCIT), *35th ASEE/IEEE Frontiers in Education Conference*, by Paul Golding and Sophia McNamarah, 2005.

Example: In a paper, Estimation of Network Distances Using Off-line Measurements, *Computer Communications*, by Prasun Sinha, Danny Raz and Nidhan Choudhuri, 2006, the authors wanted to predict Y, the round-trip time (RTT) for packets in a network, using the predictor variables $X^{(1)}$ = geographical distance between the two nodes, $X^{(2)}$ = number of router-to-router hops, and other offline variables. The goal here was primarily Prediction.

Example: In a paper, Productivity Analysis of Object-Oriented Software Developed in a Commercial Environment, *Software—Practice and Experience*, by Thomas E. Potok, Mladen Vouk and Andy Rindos, 1999, the authors mainly had an Description goal: What impact, positive or negative, does the use of object-oriented programming have on programmer productivity? Here they predicted Y = number of person-months needed to complete the project, from $X^{(1)}$ = size of the project as measured in lines of code, $X^{(2)} = 1$ or 0 depending on whether an object-oriented or procedural approach was used, and other variables.

Example: Most **text mining** applications are classification problems. For example, the paper Untangling Text Data Mining, *Proceedings of ACL'99*, by Marti Hearst, 1999 cites, *inter alia*, an application in which the analysts wished to know what proportion of patents come from publicly funded research. They were using a patent database, which of course is far too huge to feasibly search by hand. That meant that they needed to be able to (reasonably reliably) predict Y = 1 or 0, according to whether the patent was publicly funded from a number of $X^{(i)}$, each of which was an indicator variable for a given key word, such as "NSF." They would then treat the predicted Y values as the real ones, and estimate their proportion from them.

Example: A major health insurance company wanted to have a tool to predict which of its members would be likely to need hospitalization in the next year. Here Y = 1 or 0, according to whether the patient turns out to be hospitalized, and the predictor variables were the members' demographics, previous medical history and so on. (Interestingly, rather hiring its own data scientist to do the analysis, the company put the problem on Kaggle, a site that holds predictive analytics competitions, `www.kaggle.com`.)

11.3 Adjusting for Covariates

The first statistical consulting engagement I ever worked involved something called *adjusting for covariates*. I was retained by the Kaiser hospital chain to investigate how heart attack patients fared at the various hospitals—did patients have a better chance to survive in some hospitals than in others? There were four hospitals of particular interest.

I could have simply computed raw survival rates, say the proportion of patients who survive for a month following a heart attack, and then used the methods of Section 9.4, for instance. This could have been misleading, though, because one of the four hospitals served a largely elderly population. A straight comparison of survival rates might then unfairly paint that particular hospital as giving lower quality of care than the others.

So, we want to somehow adjust for the effects of age. I did this by setting Y to 1 or 0, for survival, $X^{(1)}$ to age, and $X^{(2+i)}$ to be an indicator random variable for whether the patient was at hospital i, i = 1,2,3.[2]

[2]Note that there is no i = 4 case, since if the first three hospital variables are all 0, that already tells us that this

11.4 What Does "Relationship" Really Mean?

Consider the Davis city population example again. In addition to the random variable W for weight, let H denote the person's height. Suppose we are interested in exploring the relationship between height and weight.

As usual, we must first ask, **what does that really mean**? What do we mean by "relationship"? Clearly, there is no exact relationship; for instance, a person's weight is not an exact function of his/her height.

Effective use of the methods to be presented here requires an understanding of what exactly is meant by the term *relationship* in this context.

11.4.1 Precise Definition

Intuitively, we would guess that mean weight increases with height. To state this precisely, the key word in the previous sentence is *mean*.

Take Y to be the weight W and $X^{(1)}$ to be the height H, and define

$$m_{W;H}(t) = E(W|H = t) \tag{11.1}$$

This looks abstract, but it is just common-sense stuff. For example, $m_{W;H}(68)$ would be the mean weight of all people in the population of height 68 inches. The value of $m_{W;H}(t)$ varies with t, and we would expect that a graph of it would show an increasing trend with t, reflecting that taller people tend to be heavier.

We call $m_{W;H}$ the **regression function of W on H**. In general, $m_{Y;X}(t)$ means the mean of Y among all units in the population for which $X = t$.[3]

Note the word *population* in that last sentence. The function m() is a population function.

So we have:

> **Major Point 1:** When we talk about the *relationship* of one variable to one or more others, we are referring to the regression function, which expresses the mean of the first variable as a function of the others. The key word here is *mean*!

patient was at the fourth hospital.

[3]The word "regression" is an allusion to the famous comment of Sir Francis Galton in the late 1800s regarding "regression toward the mean." This referred to the fact that tall parents tend to have children who are less tall—closer to the mean—with a similar statement for short parents. The predictor variable here might be, say, the father's height F, with the response variable being, say, the son's height S. Galton was saying that E(S | F) < F.

i ↓, j →	0	1	2	3
0	0.0079	0.0952	0.1429	0.0317
1	0.0635	0.2857	0.1905	0.1587
2	0.0476	0.0952	0.0238	0.000

Table 11.1: Bivariate pmf for the Marble Problem

11.4.2 (Rather Artificial) Example: Marble Problem

Recall the marble selection example in Section **??**: Suppose we have a bag containing two yellow marbles, three blue ones and four green ones. We choose four marbles from the bag at random, without replacement. Let Y and B denote the number of yellow and blue marbles that we get. Let's find $m_{Y;B}(2)$.

For convenience, Table 11.1 shows what we found before for P(Y = i and B = j).

Now keep in mind that since $m_{Y;B}(t)$ is the conditional mean of Y given B, we need to use conditional probabilities to compute it. For our example here of $m_{Y;B}(2)$, we need the probabilities $P(Y = k|B = 2)$. For instance,

$$
\begin{aligned}
P(Y = 1|B = 2) &= \frac{p_{Y,B}(1,2)}{p_B(2)} && (11.2)\\
&= \frac{0.1905}{0.1492 + 0.1905 + 0.0238} && (11.3)\\
&= 0.5333 && (11.4)
\end{aligned}
$$

The other conditional $P(Y = k|B = 2)$ are then found to be 0.1429/0.3572 = 0.4001 for k = 0 and 0.0238/0.3572 = 0.0667 for k = 2.

$$m_{Y;B}(2) = 0.4001 \cdot 0 + 0.5333 \cdot 1 + 0.0667 \cdot 2 = 0.667 \tag{11.5}$$

11.5 Estimating That Relationship from Sample Data

The marble example in the last section was rather artificial, in that the exact distribution of the variables was known (Table 11.1). In real applications, we don't know this distribution, and must estimate it from sample data.

As noted, $m_{W;H}(t)$ is a population function, dependent on population distributions. How can we estimate this function from sample data?

Toward that end, let's again suppose we have a random sample of 1000 people from Davis, with

$$(H_1, W_1), ..., (H_{1000}, W_{1000}) \tag{11.6}$$

being their heights and weights. We again wish to use this data to estimate population values, meaning the population regression function of W on H, $m_{W;H}(t)$. But the difference here is that we are estimating a whole function now, the whole curve $m_{W;H}(t)$. That means we are estimating infinitely many values, with one $m_{W;H}(t)$ value for each t.[4] How do we do this?

One approach would be as follows. Say we wish to find $\widehat{m}_{W;H}(t)$ (note the hat, for "estimate of"!) at t = 70.2. In other words, we wish to estimate the mean weight—in the population—among all people of height 70.2. What we could do is look at all the people in our sample who are within, say, 1.0 inch of 70.2, and calculate the average of all their weights. This would then be our $\widehat{m}_{W;H}(t)$.

11.5.1 Parametric Models for the Regression Function m()

There are many methods like the above (Chapter **??**), but the traditional method is to choose a parametric model for the regression function. That way we estimate only a finite number of quantities instead of an infinite number. This would be good in light of Section **??**.

Typically the parametric model chosen is linear, i.e. we assume that $m_{W;H}(t)$ is a linear function of t:

$$m_{W;H}(t) = ct + d \tag{11.7}$$

for some constants c and d. If this assumption is reasonable—meaning that though it may not be exactly true it is reasonably close—then it is a huge gain for us over a nonparametric model. Do you see why? Again, the answer is that instead of having to estimate an infinite number of quantities, we now must estimate only two quantities—the parameters c and d.

[4]Of course, the population of Davis is finite, but there is the conceptual population of all people who *could* live in Davis.

Equation (11.7) is thus called a **parametric** model of $m_{W;H}()$. The set of straight lines indexed by c and d is a two-parameter family, analogous to parametric families of distributions, such as the two-parametric gamma family; the difference, of course, is that in the gamma case we were modeling a density function, and here we are modeling a regression function.

Note that c and d are indeed population parameters in the same sense that, for instance, r and λ are parameters in the gamma distribution family. We must estimate c and d from our sample data.

So we have:

> **Major Point 2:** The function $m_{W;H}(t)$ is a population entity, so we must estimate it from our sample data. To do this, we have a choice of either assuming that $m_{W;H}(t)$ takes on some parametric form, or making no such assumption.
>
> If we opt for a parametric approach, the most common model is linear, i.e. (11.7). Again, the quantities c and d in (11.7) are population values, and as such, we must estimate them from the data.

11.5.2 Estimation in Parametric Regression Models

So, how can we estimate these population values c and d? We'll go into details in Section 11.10, but here is a preview:

Using the result on page 57, together with the principle of iterated expectation, (4.68) and (6.78), we can show that the minimum value of the quantity

$$E\left[(W-g(H))^2\right] \tag{11.8}$$

overall all possible functions $g(H)$, is attained by setting

$$g(H) = m_{W;H}(H) \tag{11.9}$$

In other words, $m_{W;H}(H)$ is the optimal predictor of W among all possible functions of H, in the sense of minimizing mean squared prediction error.[5]

Since we are assuming the model (11.7), this in turn means that:

[5] But if we wish to minimize the mean absolute prediction error, $E(|W-g(H)|)$, the best function turns out to be is $g(H) = \text{median}(W|H)$.

The quantity

$$E\left[(W - (uH + v))^2\right] \tag{11.10}$$

is minimized by setting u = c and v = d.

This then gives us a clue as to how to estimate c and d from our data, as follows.

If you recall, in earlier chapters we've often chosen estimators by using sample analogs, e.g. s^2 as an estimator of σ^2. Well, the sample analog of (11.10) is

$$\frac{1}{n}\sum_{i=1}^{n}[W_i - (uH_i + v)]^2 \tag{11.11}$$

Here (11.10) is the mean squared prediction error using u and v in the population, and (11.11) is the mean squared prediction error using u and v in our sample. Since u = c and v = d minimize (11.10), it is natural to estimate c and d by the u and v that minimize (11.11).

Using the "hat" notation common for estimators, we'll denote the u and v that minimize (11.11) by $\hat{c}$ and $\hat{d}$, respectively. These numbers are then the classical **least-squares estimators** of the population values c and d.

> **Major Point 3:** In statistical regression analysis, one uses a linear model as in (11.7), estimating the coefficients by minimizing (11.11).

We will elaborate on this in Section 11.10.

11.5.3 More on Parametric vs. Nonparametric Models

Suppose we're interested in the distribution of battery lifetimes, and we have a sample of them, say $B_1, ..., B_{100}$. We wish to estimate the density of lifetimes in the population of all batteries of this kind, $f_B(t)$.

We have two choices:

(a) We can simply plot a histogram of our data, which we found in Chapter **??** is actually a density estimator. We are estimating infinitely many population quantities, namely the heights of the curve $f_B(t)$ at infinitely many values of t.

(b) We could postulate a model for the distribution of battery lifetime, say using the gamma family (Section 6.6.4). Then we would estimate just two parameters, λ and r.

What are the pros and cons of (a) versus (b)? The approach (a) is nice, because we don't have to make any assumptions about the form of the curve $f_B(t)$; we just estimate it directly, with the histogram or other method from Chapter **??**. But we are, in essence, using a finite amount of data to estimate an infinite values.

As to (b), it requires us to estimate only two parameters, which is nice. Also, having a nice, compact parametric form for our estimate is appealing. But we have the problem of having to make an assumption about the form of the model. We then have to see how well the model fits the data, say using the methods in Chapter **??**. If it turns out not to fit well, we may try other models (e.g. from the Weibull family, not presented in this book).

The above situation is exactly parallel to what we are studying in the present chapter. The analogy here of estimating a density function is estimating a regression function. The analog of the histogram in (a) is the "average the people near a given height" method. The analog here of using a parametric family of densities, such as the gamma, is using a parametric family of straight lines. And the analog of comparing several candidate parametric density models is to compare several regression models, e.g. adding quadratic or cubic terms (t^2, t^3) for height in (11.7). (See Section **??** for reading on model assessment methods.)

Most statistical analysts prefer parameteric models, but nonparametric approaches are becoming increasingly popular.

11.6 Example: Baseball Data

Let's do a regression analysis of weight against height in the baseball player data introduced in Section **??**.

11.6.1 R Code

I ran R's **lm()** ("linear model") function to perform the regression analysis:

```
> summary(lm(players$Weight ~ players$Height))

Call:
lm(formula = players$Weight ~ players$Height)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-51.988 -13.147   1.218  11.694  70.012

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -155.092     17.699  -8.763   <2e-16 ***
players$Height    4.841      0.240  20.168   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.78 on 1031 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.2829,Adjusted R-squared: 0.2822
F-statistic: 406.7 on 1 and 1031 DF,  p-value: < 2.2e-16
```

This illustrates the **polymorphic** nature of R: The same function can be applied to different classes of objects. Here **summary()** is such a function; another common one is **plot()**. So, we can call **summary()** on an object of any class, at least, any one for which a **summary()** function has been written. In the above R output, we called **summary()** on an object of type **"lm"**; the R interpreter checked the class of our object, and then accordingly called **summary.lm()**. But it's convenient for us, since we ignore all that and simply call **summary()** no matter what our object is.

The call `lm(players$Weight ~players$Height)` specified that my response and predictor variables were the Weight and Height columns in the **players** data frame.

Note: The variables here are specified in an R data frame. One can also specify via a matrix, which gives more flexibility. For example,

```
lm(y ~ x[,c(2,3,7)])
```

to predict **y** from columns 2, 3 and 7 of **x**.

11.6.2 A Look through the Output

Next, note that **lm()** returns a lot of information (even more than shown above), all packed into an object of type **"lm"**.[6] By calling **summary()** on that object, I got some of the information. It gave me more than we'll cover for now, but the key is that it told me that the sample estimates of

[6]R class names are quoted.

c and d are

$$\hat{d} = -155.092 \tag{11.12}$$

$$\hat{c} = 4.841 \tag{11.13}$$

In other words, our estimate for the function giving mean weight in terms of height is

```
mean weight = -155.092 + 4.841 height
```

Do keep in mind that this is just an estimate, based on the sample data; it is not the population mean-weight-versus-height function. So for example, our *sample estimate* is that an extra inch in height corresponds on average to about 4.8 more pounds in weight.

We can form a confidence interval to make that point clear, and get an idea of how accurate our estimate is. The R output tells us that the standard error of $\hat{d}$ is 0.240. Making use of Section 9.5, we add and subtract 1.96 times this number to $\hat{d}$ to get our interval: (4.351,5.331). So, we are about 95% confident that the true slope, c, is in that interval.

Note the column of output labled "t value." This is again a Student-t test, with the p-value given in the last column, labeled "$Pr(>|t|)$." Let's discuss this. In the row of the summary above regarding the Height variable, for example, we are testing

$$H_0 : c = 0 \tag{11.14}$$

R is using a Student-t distribution for this, while we have been using the the N(0,1) distribution, based on the Central Limit Theorem approximation. For all but the smallest samples, the difference is negligible. Consider:

Using (10.6), we would test (11.14) by forming the quotient

$$\frac{4.841 - 0}{0.240} = 20.17 \tag{11.15}$$

This is essentially the same as the 20.168 we see in the above summary. In other words, don't worry that R uses the Student-t distribution while we use (10.6).

At any rate, 20.17 is way larger than 1.96, thus resulting in rejection of H_0. The p-value is then the area to the left of -20.17 and to the right of 20.17, which we could compute using **pnorm()**. But R has already done this for us, reporting that the p-value is 2×10^{-16}.

What about the **residuals**? Here we go back to the original (H_i, W_i) data with our slope and intercept estimates, and "predict" each W_i from the corresponding H_i. The residuals are the resulting prediction errors. In other words, the ith residual is

$$W_i - (\hat{d} + \hat{c} H_i) \tag{11.16}$$

You might wonder why we would try to predict the data that we already know! But the reason for doing this is to try to assess how well we can predict future cases, in which we know height but not weight. If we can "predict" well in our known data, maybe we'll do well later with unknown data. This will turn out to be somewhat overoptimistic, we'll see, but again, the residuals should be of at least *some* value in assessing the predictive ability of our model. So, the R output reports to us what the smallest and largest residual values were.

The R^2 values will be explained in Section **??**.

Finally, the F-test is a significance test that c = d = 0. Since this book does not regard testing as very useful, this aspect will not be pursued here.

11.7 Multiple Regression: More Than One Predictor Variable

Note that X and t could be vector-valued. For instance, we could have Y be weight and have X be the pair

$$X = \left(X^{(1)}, X^{(2)}\right) = (H, A) = (\text{height, age}) \tag{11.17}$$

so as to study the relationship of weight with height and age. If we used a linear model, we would write for $t = (t_1, t_2)$,

$$m_{W;H,A}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 \tag{11.18}$$

In other words

$$\text{mean weight} = \beta_0 + \beta_1 \text{ height} + \beta_2 \text{ age} \tag{11.19}$$

Once again, keep in mind that (11.18) and (11.19) are models for the population. We assume that (11.18), (11.19) or whichever model we use is an exact representation of the relation in the population. And of course, our derivations below assume our model is correct.

(It is traditional to use the Greek letter β to name the coefficients in a linear regression model.)

So for instance $m_{W;H,A}(68, 37.2)$ would be the mean weight in the population of all people having height 68 and age 37.2.

In analogy with (11.11), we would estimate the β_i by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} [W_i - (u + vH_i + wA_i)]^2 \tag{11.20}$$

with respect to u, v and w. The minimizing values would be denoted $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

We might consider adding a third predictor, gender:

$$\text{mean weight} = \beta_0 + \beta_1 \text{ height} + \beta_2 \text{ age} + \beta_3 \text{ gender} \tag{11.21}$$

where **gender** is an indicator variable, 1 for male, 0 for female. Note that we would not have two gender variables, since knowledge of the value of one such variable would tell us for sure what the other one is. (It would also make a certain matrix noninvertible, as we'll discuss later.)

11.8 Example: Baseball Data (cont'd.)

So, let's regress weight against height and age:

```
> summary(lm(players$Weight ~ players$Height + players$Age))

Call:
lm(formula = players$Weight ~ players$Height + players$Age)

Residuals:
    Min      1Q  Median      3Q     Max
-50.794 -12.141  -0.304  10.737  74.206

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -192.6564    17.8905 -10.769  < 2e-16 ***
players$Height    4.9746     0.2341  21.247  < 2e-16 ***
players$Age       0.9647     0.1249   7.722  2.7e-14 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.3 on 1030 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.3221,Adjusted R-squared: 0.3208
F-statistic: 244.8 on 2 and 1030 DF,  p-value: < 2.2e-16
```

So, our regression function coefficient estimates are $\hat{\beta}_0 = -192.6564$, $\hat{\beta}_1 = 4.9746$ and $\hat{\beta}_2 = 0.9647$. For instance, we estimate from our sample data that 10 years' extra age results, on average, of a weight gain about about 9.6 pounds—for people of a given height. This last condition is very important.

11.9 Interaction Terms

Equation (11.18) implicitly says that, for instance, the effect of age on weight is the same at all height levels. In other words, the difference in mean weight between 30-year-olds and 40-year-olds is the same regardless of whether we are looking at tall people or short people. To see that, just plug 40 and 30 for age in (11.18), with the same number for height in both, and subtract; you get $10\beta_2$, an expression that has no height term.

That assumption is not a good one, since the weight gain in aging tends to be larger for tall people than for short ones. If we don't like this assumption, we can add an **interaction term** to (11.18), consisting of the product of the two original predictors. Our new predictor variable $X^{(3)}$ is equal to $X^{(1)}X^{(2)}$, and thus our regression function is

$$m_{W;H}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_1 t_2 \tag{11.22}$$

If you perform the same subtraction described above, you'll see that this more complex model does not assume, as the old did, that the difference in mean weight between 30-year-olds and 40-year-olds is the same regardless of we are looking at tall people or short people.

Recall the study of object-oriented programming in Section 11.1. The authors there set $X^{(3)} = X^{(1)}X^{(2)}$. The reader should make sure to understand that without this term, we are basically saying that the effect (whether positive or negative) of using object-oriented programming is the same for any code size.

Though the idea of adding interaction terms to a regression model is tempting, it can easily get out of hand. If we have k basic predictor variables, then there are $\binom{k}{2}$ potential two-way interaction terms, $\begin{pmatrix} k \\ 3 \end{pmatrix}$ three-way terms and so on. Unless we have a very large amount of data, we run a big risk of overfitting (Section **??**). And with so many interaction terms, the model would be difficult to interpret.

We can add even more interaction terms by introducing powers of variables, say the square of height in addition to height. Then (11.22) would become

$$m_{W;H}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_1 t_2 + \beta_4 t_1^2 \tag{11.23}$$

This square is essentially the "interaction" of height with itself. If we believe the relation between weight and height is quadratic, this might be worthwhile, but again, this means more and more predictors.

So, we may have a decision to make here, as to whether to introduce interaction terms. For that matter, it may be the case that age is actually not that important, so we even might consider dropping that variable altogether. These questions will be pursued in Section **??**.

11.10 Parametric Estimation of Linear Regression Functions

So, how did R compute those estimated regression coefficients? Let's take a look.

11.10.1 Meaning of "Linear"

Here we model $m_{Y;X}$ as a linear function of $X^{(1)}, ..., X^{(r)}$:

$$m_{Y;X}(t) = \beta_0 + \beta_1 t^{(1)} + ... + \beta_r t^{(r)} \tag{11.24}$$

Note that the term **linear regression** does NOT necessarily mean that the graph of the regression function is a straight line or a plane. We could, for instance, have one predictor variable set equal to the square of another, as in (11.23).

Instead, the word *linear* refers to the regression function being linear in the parameters. So, for instance, (11.23) is a linear model; if for example we multiple β_0, β_1 and β_2 by 8, then $m_{A;b}(s)$ is multiplied by 8.

A more literal look at the meaning of "linear" comes from the matrix formulation (**??**) below.

11.10.2 Random-X and Fixed-X Regression

Consider our earlier example of estimating the regression function of weight on height. To make things, simple, say we sample only 5 people, so our data is $(H_1, W_1), ..., (H_5, W_5)$. and we measure height to the nearest inch.

In our "notebook" view, each line of our notebook would have 5 heights and 5 weights. Since we would have a different set of 5 people on each line, in the H_1 column will generally have different values from line to line, though occasionally two consecutive lines will have the same value. H_1 is a random variable. We can regression analysis in this setting **random-X** regression.

We could, on the other hand, set up our sampling plan so that we sample one person each of heights 65, 67, 69, 71 and 73. These values would then stay the same from line to line. The H_1 column, for instance, would consist entirely of 65s. This is called **fixed-X regression**.

So, the probabilistic structure of the two settings is different. However, it turns out not to matter much, for the following reason.

Recall that the definition of the regression function, concerns the *conditional* distribution of W given H. So, our analysis below will revolve around that conditional distribution, in which case H becomes nonrandom anyway.

11.10.3 Point Estimates and Matrix Formulation

So, how do we estimate the β_i? Keep in mind that the β_i are population values, which we need to estimate them from our data. How do we do that? For instance, how did R compute the $\widehat{\beta}_i$ in Section 11.6? As previewed in Section 11.5, the usual method is least-squares. Here we will go into the details.

For concreteness, think of the baseball data, and let H_i, A_i and W_i denote the height, age and weight of the ith player in our sample, i = 1,2,...,1033. As in (11.11), the estimation methodology involves finding the values of u_i which minimize the sum of squared differences between the actual W values and their predicted values using the u_i:

$$\sum_{i=1}^{1033} [W_i - (u_0 + u_1 H_i + u_2 A_i)]^2 \tag{11.25}$$

When we find the minimizing u_i, we will set our estimates for the population regression coefficients β_i in (11.24):

$$\widehat{\beta}_0 = u_0 \tag{11.26}$$

$$\widehat{\beta}_1 = u_1 \tag{11.27}$$

$$\widehat{\beta}_2 = u_2 \tag{11.28}$$

Obviously, this is a calculus problem. We set the partial derivatives of (11.25) with respect to the u_i to 0, giving use three linear equations in three unknowns, and then solve.

In linear algebra terms, define

$$V = \begin{pmatrix} W_1 \\ W_2 \\ ... \\ W_{1033} \end{pmatrix}, \tag{11.29}$$

$$u = \begin{pmatrix} u_0 \\ u_1 \\ u_2 \end{pmatrix} \tag{11.30}$$

and

$$Q = \begin{pmatrix} 1 & H_1 & A_1 \\ 1 & H_2 & A_2 \\ ... \\ 1 & H_{1033} & A_{1033} \end{pmatrix} \tag{11.31}$$

Then

$$E(V \mid Q) = Q\beta \tag{11.32}$$

To see this, look at the first player, of height 74 and age 22.99 (Section **??**). We are modeling the mean weight in the population for all players of that height and weight as

$$\text{mean weight} = \beta_0 + \beta_1 \ 74 + \beta_2 \ 22.99 \tag{11.33}$$

The top row of Q will be (1,74,22.99), so the top row of $Q\beta$ will be $\beta_0 + \beta_1 \ 74 + \beta_2 \ 22.99$ — which exactly matches (11.33). Note the need for the 1s column in Q.

we can write (11.25) as

$$(V - Qu)'(V - Qu) \tag{11.34}$$

Whatever vector u minimizes (11.34), we set our estimated β vector $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)'$ to that u.

Then it can be shown that, after all the partial derivatives are taken and set to 0, the solution is

$$\hat{\beta} = (Q'Q)^{-1}Q'V \tag{11.35}$$

For the general case (11.24) with n observations (n = 1033 in the baseball data), the matrix Q has n rows and r+1 columns. Column i+1 has the sample data on predictor variable i.

Note that we are conditioning on Q in (11.32). This is the standard approach, especially since there is the case of nonrandom X. Thus we will later get conditional confidence intervals, which is fine. To avoid clutter, I will sometimes not show the conditioning explicitly, and thus for instance will write, for example, Cov(V) instead of Cov(V|Q).

It turns out that $\hat{\beta}$ is an unbiased estimate of β:[7]

$$\begin{aligned}
E\hat{\beta} &= E[(Q'Q)^{-1}Q'V] \quad (11.35) && (11.36)\\
&= (Q'Q)^{-1}Q'EV \quad \text{(linearity of E())} && (11.37)\\
&= (Q'Q)^{-1}Q' \cdot Q\beta \quad (\textbf{??}) && (11.38)\\
&= \beta && (11.39)
\end{aligned}$$

In some applications, we assume there is no constant term β_0 in (11.24). This means that our Q matrix no longer has the column of 1s on the left end, but everything else above is valid.

11.10.4 Approximate Confidence Intervals

As noted, R gives you standard errors for the estimated coefficients. Where do they come from?

As usual, we should not be satisfied with just point estimates, in this case the $\widehat{\beta}_i$. We need an indication of how accurate they are, so we need confidence intervals. In other words, we need to use the $\widehat{\beta}_i$ to form confidence intervals for the β_i.

[7] Note that here we are taking the expected value of a vector, as in Chapter **??**.

For instance, recall the study on object-oriented programming in Section 11.1. The goal there was primarily Description, specifically assessing the impact of OOP. That impact is measured by β_2. Thus, we want to find a confidence interval for β_2.

Equation (11.35) shows that the $\widehat{\beta}_i$ are sums of the components of V, i.e. the W_j. So, the Central Limit Theorem implies that the $\widehat{\beta}_i$ are approximately normally distributed. That in turn means that, in order to form confidence intervals, we need standard errors for the β_i. How will we get them?

Note carefully that so far we have made NO assumptions other than (11.24). Now, though, we need to add an assumption:[8]

$$Var(Y|X=t) = \sigma^2 \tag{11.40}$$

for all t. Note that this and the independence of the sample observations (e.g. the various people sampled in the Davis height/weight example are independent of each other) implies that

$$Cov(V|Q) = \sigma^2 I \tag{11.41}$$

where I is the usual identiy matrix (1s on the diagonal, 0s off diagonal).

Be sure you understand what this means. In the Davis weights example, for instance, it means that the variance of weight among 72-inch tall people is the same as that for 65-inch-tall people. That is not quite true—the taller group has larger variance—but research into this has found that as long as the discrepancy is not too bad, violations of this assumption won't affect things much.

We can derive the covariance matrix of $\hat{\beta}$ as follows. Again to avoid clutter, let $B = (Q'Q)^{-1}$. A theorem from linear algebra says that Q'Q is symmetric and thus B is too. Another theorem says that for any conformable matrices U and V, then (UV)' = V'U'. Armed with that knowledge, here we go:

$$\begin{aligned}
Cov(\hat{\beta}) &= Cov(BQ'V) \text{ ((11.35))} && (11.42)\\
&= BQ'Cov(V)(BQ')' \text{ (\textbf{??})} && (11.43)\\
&= BQ'\sigma^2 I(BQ')' \text{ (11.41)} && (11.44)\\
&= \sigma^2 BQ'QB \text{ (lin. alg.)} && (11.45)\\
&= \sigma^2 (Q'Q)^{-1} \text{ (def. of B)} && (11.46)
\end{aligned}$$

[8]Actually, we could derive some usable, though messy,standard errors without this assumption.

Whew! That's a lot of work for you, if your linear algebra is rusty. But it's worth it, because (11.46) now gives us what we need for confidence intervals. Here's how:

First, we need to estimate σ^2. Recall first that for any random variable U, $Var(U) = E[(U - EU)^2]$, we have

$$\begin{aligned}
\sigma^2 &= Var(Y|X = t) && (11.47)\\
&= Var(Y|X^{(1)} = t_1, ..., X^{(r)} = t_r) && (11.48)\\
&= E\left[\{Y - m_{Y;X}(t)\}^2\right] && (11.49)\\
&= E\left[(Y - \beta_0 - \beta_1 t_1 - ... - \beta_r t_r)^2\right] && (11.50)
\end{aligned}$$

Thus, a natural estimate for σ^2 would be the sample analog, where we replace E() by averaging over our sample, and replace population quantities by sample estimates:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i^{(1)} - ... - \hat{\beta}_r X_i^{(r)})^2 \tag{11.51}$$

As in Chapter **??**, this estimate of σ^2 is biased, and classicly one divides by n-(r+1) instead of n. But again, it's not an issue unless r+1 is a substantial fraction of n, in which case you are overfitting and shouldn't be using a model with so large a value of r.

So, the estimated covariance matrix for $\hat{\beta}$ is

$$\widehat{Cov}(\hat{\beta}) = s^2(Q'Q)^{-1} \tag{11.52}$$

The diagonal elements here are the squared standard errors (recall that the standard error of an estimator is its estimated standard deviation) of the β_i. (And the off-diagonal elements are the estimated covariances between the β_i.) Since the first standard errors you ever saw, in Section 9.5, included factors like $1/\sqrt{n}$, you might wonder why you don't see such a factor in (11.52).

The answer is that such a factor is essentially there, in the following sense. Q'Q consists of various sums of products of the X values, and the larger n is, then the larger the elements of Q'Q are. So, $(Q'Q)^{-1}$ already has something like a "1/n" factor in it.

R's **vcov()** function, applied to the output of **lm()** will give you (11.52) (subject to a bias correction factor that we'll discuss in Section **??**, but that we'll dismiss as unimportant.

11.11 Example: Baseball Data (cont'd.)

Let us use **vcov()** to obtain the estimated covariance matrix of the vector $\widehat{\beta}$ for our baseball data.

```
> lmout <- lm(players$Weight ~ players$Height + players$Age)
> vcov(lmout)
               (Intercept) players$Height  players$Age
(Intercept)    320.0706223   -4.102047105 -0.607718793
players$Height  -4.1020471    0.054817211  0.002160128
players$Age     -0.6077188    0.002160128  0.015607390
```

The first command saved the output of **lm()** in a variable that we chose to name **lmout**; we then called **vcov()** on that object.

For instance, the estimated variance of $\widehat{\beta}_1$ is 0.054817211. Actually, we already knew this, because the standard error of $\widehat{\beta}_1$ was reported earlier to be 0.2341, and $0.2341^2 = 0.054817211$.

But now we can find more. Say we wish to compute a confidence interval for the population mean weight of players who are 72 inches tall and age 30. That quantity is equal to

$$\beta_0 + 72\beta_1 + 30\beta_2 = (1, 72, 30)\beta \tag{11.53}$$

which we will estimate by

$$(1, 72, 30)\widehat{\beta} \tag{11.54}$$

Thus, using (**??**), we have

$$\widehat{Var}(\widehat{\beta}_0 + 72\widehat{\beta}_1 + 30\widehat{\beta}_2) = (1, 72, 30)A \begin{pmatrix} 1 \\ 72 \\ 30 \end{pmatrix} \tag{11.55}$$

where A is the matrix in the R output above.

The square root of this quantity is the standard error of $\widehat{\beta}_0 + 72\widehat{\beta}_1 + 30\widehat{\beta}_2$. We add and subtract 1.96 times that square root to $\widehat{\beta}_0 + 72\widehat{\beta}_1 + 30\widehat{\beta}_2$, and then have an approximate 95% confidence interval for the population mean weight of players who are 72 inches tall and age 30.

11.12 Dummy Variables

Recall our example in Section 11.2 concerning a study of software engineer productivity. To review, the authors of the study predicted Y = number of person-months needed to complete the project, from $X^{(1)}$ = size of the project as measured in lines of code, $X^{(2)} = 1$ or 0 depending on whether an object-oriented or procedural approach was used, and other variables.

As mentioned at the time, $X^{(2)}$ is an indicator variable, often called a "dummy" variable in the regression context.

Let's generalize that a bit. Suppose we are comparing two different object-oriented languages, C++ and Java, as well as the procedural language C. Then we could change the definition of $X^{(2)}$ to have the value 1 for C++ and 0 for non-C++, and we could add another variable, $X^{(3)}$, which has the value 1 for Java and 0 for non-Java. Use of the C language would be implied by the situation $X^{(2)} = X^{(3)} = 0$.

Note that we do NOT want to represent Language by a single value having the values 0, 1 and 2, which would imply that C has, for instance, double the impact of Java.

11.13 Example: Baseball Data (cont'd.)

Let's now bring the Position variable into play. First, what is recorded for that variable?

```
> levels(players$Position)
[1] "Catcher"           "Designated_Hitter" "First_Baseman"
[4] "Outfielder"        "Relief_Pitcher"    "Second_Baseman"
[7] "Shortstop"         "Starting_Pitcher"  "Third_Baseman"
```

So, all the outfield positions have been simply labeled "Outfielder," though pitchers have been separated into starters and relievers.

Technically, this variable, **players$Position**, is an R **factor**. This is a fancy name for an integer vector with labels, such that the labels are normally displayed rather than the codes. So actually catchers are coded 1, designated hitters 2, first basemen 3 and so on, but in displaying the data frame, the labels are shown rather than the codes.

The designated hitters are rather problematic, as they only exist in the American League, not the National League. Let's restrict our analysis to the other players:

```
> nondh <- players[players$Position != "Designated_Hitter",]
> nrow(players)
[1] 1034
```

```
> nrow(nondh)
[1] 1016
```

This requires some deconstruction. The expression `players$Position !="Designated_Hitter"` gives us a vector of True and False values. Then `players[players$Position !="Designated_Hitter",]` consists of all rows of **players** corresponding to a True value. Result: We've deleted the designated hitters, assigning the result to **nondh**. A comparison of numbers of rows show that there were only 18 designated hitters in the data set anyway.

Let's consolidate into four kinds of positions: infielders, outfielders, catchers and pitchers. First, switch to numeric codes, in a vector we'll name **poscodes**:

```
> poscodes <- as.integer(nondh$Position)
> head(poscodes)
[1] 1 1 1 3 3 6
> head(nondh$Position)
[1] Catcher       Catcher       Catcher       First_Baseman  First_Baseman
[6] Second_Baseman
9 Levels: Catcher Designated_Hitter First_Baseman ... Third_Baseman
```

Now consolidate into three dummy variables:

```
> infld <- as.integer(poscodes==3 | poscodes==6 | poscodes==7 | poscodes==9)
> outfld <- as.integer(poscodes==4)
> pitcher <- as.integer(poscodes==5 | poscodes==8)
```

Again, remember that catchers are designated via the other three dummies being 0.

So, let's run the regression:

```
> summary(lm(nondh$Weight ~ nondh$Height + nondh$Age + infld + outfld + pitcher))

Call:
lm(formula = nondh$Weight ~ nondh$Height + nondh$Age + infld +
    outfld + pitcher)

Residuals:
    Min      1Q  Median      3Q     Max
-49.669 -12.083  -0.386  10.410  75.081

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -193.2557    19.0127 -10.165  < 2e-16 ***
```

```
nondh$Height    5.1075     0.2520  20.270  < 2e-16 ***
nondh$Age       0.8844     0.1251   7.068 2.93e-12 ***
infld          -7.7727     2.2917  -3.392 0.000722 ***
outfld         -6.1398     2.3169  -2.650 0.008175 **
pitcher        -8.3017     2.1481  -3.865 0.000118 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.1 on 1009 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.3286,Adjusted R-squared: 0.3253
F-statistic: 98.76 on 5 and 1009 DF,  p-value: < 2.2e-16
```

The estimated coefficients for the position variables are all negative. For example, for a given height and age, pitchers are on average about 8.3 pounds lighter than catchers, while outfielders are about 6.1 pounds lighter than catchers.

What if we want to compare infielders and outfielders, say form a confidence interval for $\beta_3 - \beta_4$? Then we'd do a computation like (11.55), with a vector (0,0,0,1,-1,0) instead of (1,72,30).

11.14 What Does It All Mean?—Effects of Adding Predictors

Keep in mind the twin goals of regression analysis, Prediction and Description. In applications in which Description is the goal, we are keenly interested in the signs and magnitudes of the β_i,[9] especially their signs. We do need to be careful, just as we saw in Section 10.11; the sign of a coefficient usually won't be of much interest if the magnitude is near 0. Subject to that caution, discussion of regression results often centers on the sign of a coefficient: Is there a positive relationship between the response variable and a predictor, holding the other predictors constant?

That latter phrase, *holding the other predictors constant*, is key. Recall for example our example at the start of this chapter on a study of the effects of using the object-oriented programming paradigm. Does OOP help or hurt productivity? Since longer programs often take longer to write, the researchers wanted to correct for program length, so they used that as a predictor, in addition to a dummy variable for OOP. In other words, they wanted to know the impact of OOP on productivity, holding program length constant.

So, in studying a predictor variable, it may matter greatly which other predictors one is using. Let's examine the baseball data in this regard.

[9]As estimated from the $\widehat{\beta}_i$.

In Section 11.8, we added the age variable as our second predictor, height being the first. This resulted in the coefficient of height increasing from 4.84 to 4.97. This is not a large change, but what does it tell us? It suggests that older players tend to be shorter. No, this doesn't mean the players shrink with age—shrinkage does occur among the elderly, but likely not here—but rather that other phenomena are at work. It could be, for instance, that shorter players tend to have longer careers. This in turn might be due to a situation in which certain positions whose players tend to be tall have shorter careers. All of this could be explored, say starting with calculating the correlation between height and age.[10]

To develop some intuition on this, consider the following artificial population of eight people:

gender	height	weight
male	66	150
male	70	165
male	70	175
male	70	185
female	66	120
female	66	130
female	66	140
female	70	155

Here is the weight-height relationship for men, i.e. the mean weight for each height group:

men:

height	mean weight
66	150
70	175

$$\beta_{height} = (175 - 150)/4 = 6.25 \tag{11.56}$$

women:

height	mean weight
66	130
70	155

$$\beta_{height} = (155 - 130)/4 = 6.25 \tag{11.57}$$

The coefficient of height is the same for both gender subpopulations.

[10]The R function **cor()** computes the correlation between its first two arguments if they are vectors.

But look what happens when we remove gender from the analysis:

all:

height	mean weight
66	135
70	170

$$\beta_{height} = (170 - 135)/4 = 8.75 \tag{11.58}$$

In other words, the beta coefficient for height is 8.75 if gender is not in the equation, but is only 6.25 if we add in gender. For a given height, men in this population tend to be heavier, and since the men tend to be taller, that inflated the height coefficient in the genderless analysis.

Returning to the baseball example, recall that in in Section 11.13, we added the position variables to height and age as predictors. The coefficient for height, which had increased when we added in the age variable, now increased further, while the coefficient for age decreased, compared to the results in Section 11.8. Those heavy catchers weren't separated out from the other players in our previous analysis, and now that we are separating them from the rest, the relationship of weight versus height and age is now clarified.

Such thinking was central to another baseball example, in *Mere Mortals: Retract This Article*, Gregory Matthews blog,

`http://statsinthewild.wordpress.com/2012/08/23/mere-mortals-retract-this-article/`.

There the author took exception to someone else's analysis that purported to show that professional baseball players have a higher mortality rate than do pro football players. This was counterintuitive, since football is more of a contact sport. It turned out that the original analysis had been misleading, as it did not use age as a predictor.

Clearly, the above considerations are absolutely crucial to effective use of regression analysis for the Description goal. This insight is key—don't do regression without it! And for the same reasons, whenever you read someone else's study, do so with a skeptical eye.

Chapter 12

Relations Among Variables

It is a very sad thing that nowadays there is so little useless information—Oscar Wilde, famous 19th century writer

Unlike the case of regression analysis, where the response/dependent variable plays a central role, we are now interested in symmetric relations among several variables. Often our goal is **dimension reduction**, meaning compressing our data into just a few important variables.

Dimension reduction ties in to the Oscar Wilde quote above, which is a complaint that there is too *much* information of the use*ful* variety. We are concerned here with reducing the complexity of that information to a more manageable, simple set of variables.

Here we cover two of the most widely-used methods, **principal components analysis** for continuous variables, and the **log-linear model** for the discrete case. We also introduce **clustering**.

12.1 Principal Components Analysis (PCA)

Consider a random vector $X = (X_1, X_2)'$. Suppose the two components of X are highly correlated with each other. Then for some constants c and d,

$$X_2 \approx c + dX_1 \tag{12.1}$$

Then in a sense there is really just one random variable here, as the second is nearly equal to some linear combination of the first. The second provides us with almost no new information, once we have the first.

In other words, even though the vector X roams in two-dimensional space, it usually sticks close to a one-dimensional object, namely the line (12.1). We saw a graph illustrating this in our chapter on multivariate distributions, page **??**.

In general, consider a k-component random vector

$$X = (X_1, ..., X_k)' \tag{12.2}$$

We again wish to investigate whether just a few, say w, of the X_i tell almost the whole story, i.e. whether most X_j can be expressed approximately as linear combinations of these few X_i. In other words, even though X is k-dimensional, it tends to stick close to some w-dimensional subspace.

Note that although (12.1) is phrased in prediction terms, we are not (or more accurately, not necessarily) interested in prediction here. We have not designated one of the $X^{(i)}$ to be a response variable and the rest to be predictors.

Once again, the Principle of Parsimony is key. If we have, say, 20 or 30 variables, it would be nice if we could reduce that to, for example, three or four. This may be easier to understand and work with, albeit with the complication that our new variables would be linear combinations of the old ones.

12.1.1 How to Calculate Them

Here's how it works. Let Σ denote the covariance matrix of X. The theory of linear algebra says that since Σ is a symmetric matrix, it is diagonalizable, i.e. there is a real matrix Q for which

$$Q'\Sigma Q = D \tag{12.3}$$

where D is a diagonal matrix. (A related approach is **singular value decomposition**.) The columns C_i of Q are the eigenvectors of Σ, and it turns out that they are orthogonal to each other, i.e. their dot product is 0.

Let

$$W_i = C_i' X, \ \ i = 1, ..., k \tag{12.4}$$

so that the W_i are scalar random variables, and set

$$W = (W_1, ..., W_k)' \tag{12.5}$$

Then

$$W = Q'X \tag{12.6}$$

Now, use the material on covariance matrices from our chapter on random vectors, page **??**,

$$Cov(W) = Cov(Q'X) = Q'Cov(X)Q = D \quad \text{(from (12.3))} \tag{12.7}$$

Note too that if X has a multivariate normal distribution (which we are not assuming), then W does too.

Let's recap:

- We have created new random variables W_i as linear combinations of our original X_j.
- The W_i are uncorrelated. Thus if in addition X has a multivariate normal distribution, so that W does too, then the W_i will be independent.
- The variance of W_i is given by the ith diagonal element of D.

The W_i are called the **principal components** of the distribution of X.

It is customary to relabel the W_i so that W_1 has the largest variance, W_2 has the second-largest, and so on. We then choose those W_i that have the larger variances, and discard the others, because the latter, having small variances, are close to constant and thus carry no information.

Note that an alternate definition of the first principal component is a value of u that maximizes $u'X$ subject to u having length 1. The second principal component maximizes $u'X$ subject to u having length 1 and subject to the second component being orthogonal to the first, and so on.

All this will become clearer in the example below.

12.1.2 Example: Forest Cover Data

Let's try using principal component analysis on the forest cover data set we've looked at before. There are 10 continuous variables.[1]

In my R run, the data set (not restricted to just two forest cover types, but consisting only of the first 1000 observations) was in the object **f**. Here are the call and the results:

[1] There are also many discrete ones.

```
> prc <- prcomp(f[,1:10])
> summary(prc)
Importance of components:
                            PC1      PC2      PC3      PC4      PC5 PC6
Standard deviation     1812.394 1613.287 1.89e+02 1.10e+02 96.93455 30.16789
Proportion of Variance    0.552    0.438 6.01e-03 2.04e-03  0.00158 0.00015
Cumulative Proportion     0.552    0.990 9.96e-01 9.98e-01  0.99968 0.99984
                            PC7      PC8 PC9  PC10
Standard deviation     25.95478 16.78595 4.2 0.783
Proportion of Variance  0.00011  0.00005 0.0 0.000
Cumulative Proportion   0.99995  1.00000 1.0 1.000
```

You can see from the variance values here that R has scaled the W_i so that their variances sum to 1.0. (It has not done so for the standard deviations, which are for the nonscaled variables.) This is fine, as we are only interested in the variances relative to each other, i.e. saving the principal components with the larger variances.

What we see here is that eight of the 10 principal components have very small variances, i.e. are close to constant. In other words, though we have 10 variables $X_1, ..., X_{10}$, there is really only two variables' worth of information carried in them.

So for example if we wish to predict forest cover type from these 10 variables, we should only use two of them. We could use W_1 and W_2, but for the sake of interpretability we stick to the original X vector. We can use any two of the X_i, though typically it would be two that have large coefficients in the top two principal components..

The coefficients of the linear combinations which produce W from X, i.e. the Q matrix, are available via **prc$rotation**.

12.1.3 Scaling

If your original variables range quite a bit in variance, you should have **prcomp()** scale them first, so they all have standard deviation 1.[2] The argument name is **scale**, of course.

Without scaling, the proportion-of-total-variance type of analysis discussed above may be misleading, as large-variance variables may dominate.

12.1.4 Scope of Application

PCA makes no assumptions about the data. It is strictly an exploratory/descriptive tool.

However, it should be noted that the motivation we presented for PCA at the beginning of this chapter involved correlations among our original variables. This is further highlighted by the fact

[2] And mean 0, though this is irrelevant, as Σ is all that matter.

that the PCs are calculated based on the covariance matrix of the data, which except for scale is the same as the correlation matrix.

This in turn implies that each variable is at least *ordinal* in nature, i.e. that it makes sense to speak of the impact of larger or smaller values of a variable.

Note, though, that an indicator random variable is inherently ordinal! So, if you have a *categorical* variable, i.e. one that simply codes what category an individual falls into (such as Democratic, Republican, independent), then you can convert it to a set of indicator variables, and potentially get some insight into the relation between this variable and others.

This can be especially valuable if, as is often the case, your data consists of a mixture of ordinal and categorical variables.

12.1.5 Example: Turkish Teaching Evaluation Data

This data, again from the UCI Machine Learning Repository, was introduced in Section **??**. Let's try PCA on it:

```
> tpca <- prcomp(turk,scale=T)
> summary(tpca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
PC7
Standard deviation     4.8008 1.1296 0.98827 0.62725 0.59837 0.53828 0.50587
Proportion of Variance 0.7947 0.0440 0.03368 0.01357 0.01235 0.00999 0.00882
Cumulative Proportion  0.7947 0.8387 0.87242 0.88598 0.89833 0.90832 0.91714
                           PC8     PC9    PC10    PC11    PC12    PC13
PC14
Standard deviation     0.45182 0.42784 0.41517 0.37736 0.37161 0.36957 0.3450
Proportion of Variance 0.00704 0.00631 0.00594 0.00491 0.00476 0.00471 0.0041
Cumulative Proportion  0.92418 0.93050 0.93644 0.94135 0.94611 0.95082 0.9549
                          PC15    PC16    PC17    PC18    PC19    PC20
PC21
Standard deviation     0.34114 0.33792 0.33110 0.32507 0.31687 0.30867 0.3046
Proportion of Variance 0.00401 0.00394 0.00378 0.00364 0.00346 0.00329 0.0032
Cumulative Proportion  0.95894 0.96288 0.96666 0.97030 0.97376 0.97705 0.9802
                          PC22    PC23    PC24    PC25    PC26    PC27
PC28
Standard deviation     0.29083 0.29035 0.28363 0.27815 0.26602 0.26023 0.23621
Proportion of Variance 0.00292 0.00291 0.00277 0.00267 0.00244 0.00234 0.00192
Cumulative Proportion  0.98316 0.98607 0.98884 0.99151 0.99395 0.99629 0.99821
```

```
                           PC29
Standard deviation     0.22773
Proportion of Variance 0.00179
Cumulative Proportion  1.00000
```

This is remarkable—the first PC accounts for 79% of the variance of the set of 29 variables. In other words, in spite of the survey asking supposedly 29 different aspects of the course, they can be summarized largely in just one variable. Let's see what that variable is:

```
> tpca$rotation[,1]
         Q1          Q2          Q3          Q4          Q5
Q6
-0.16974120 -0.18551431 -0.18553930 -0.18283025 -0.18973563 -0.18635256
         Q7          Q8          Q9         Q10         Q11         Q12
-0.18730028 -0.18559928 -0.18344211 -0.19241585 -0.18388873 -0.18184118
        Q13         Q14         Q15         Q16         Q17         Q18
-0.19430111 -0.19462822 -0.19401115 -0.19457451 -0.18249389 -0.19320936
        Q19         Q20         Q21         Q22         Q23         Q24
-0.19412781 -0.19335127 -0.19232101 -0.19232914 -0.19554282 -0.19328500
        Q25         Q26         Q27         Q28  difficulty
-0.19203359 -0.19186433 -0.18751777 -0.18855570 -0.01712709
```

This is even more remarkable. Except for the "difficulty" variable, all the Qi have about the same coefficients (the same **loadings**). In other words, just one question would have been enough, and it wouldn't matter mcuh which one were used.

The second PC, though only accounting for 4% of the total variation, is still worth a look:

```
> tpca$rotation[,2]
         Q1          Q2          Q3          Q4          Q5
Q6
 0.32009850  0.22046468  0.11432028  0.23340347  0.20236372  0.19890471
         Q7          Q8          Q9         Q10         Q11         Q12
 0.24025046  0.24477543  0.13198060  0.19239207  0.11064523  0.20881773
        Q13         Q14         Q15         Q16         Q17         Q18
-0.09943140 -0.15193169 -0.15089563 -0.03494282 -0.26163096 -0.11646066
        Q19         Q20         Q21         Q22         Q23         Q24
-0.14424468 -0.18729978 -0.21208705 -0.21650494 -0.09349599 -0.05372049
        Q25         Q26         Q27         Q28  difficulty
-0.20342350 -0.10790888 -0.05928032 -0.20370705 -0.27672177
```

Here the "difficulty" variable now shows up, and some of the Qi become unimportant.