From Algorithms to Z-Scores:

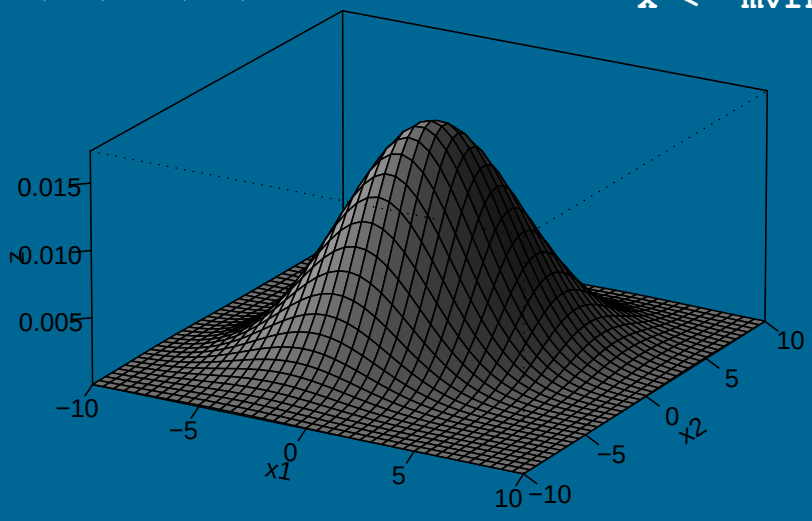
Probabilistic and Statistical Modeling in Computer Science

Norm Matloff

University of California, Davis

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

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



See Creative Commons license at

http://heather.cs.ucdavis.edu/ matloff/probstatbook.html

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 mathematics and 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 parallel processing, analysis of social networks, 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 UNESCO. 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 the author of two published textbooks, and of a number of widely-used Web tutorials on computer topics, such as the Linux operating system and the Python programming language. He and Dr. Peter Salzman are authors of *The Art of Debugging with GDB, DDD, and Eclipse*. Prof. Matloff's book on the R programming language, *The Art of R Programming*, is due to be published in 2010. He is also the author of several open-source textbooks, including *From Algorithms to Z-Scores: Probabilistic and Statistical Modeling in Computer Science* (`http://heather.cs.ucdavis.edu/probstatbook`), and *Programming on Parallel Machines* (`http://heather.cs.ucdavis.edu/~matloff/ParProcBook.pdf`).

Contents

Preface

Why is this book different from all other books on probability and statistics?

First, the book stresses computer science applications. Though other books of this nature have been published, notably the outstanding text by K.S. Trivedi, this book has much more coverage of statistics, including a full chapter titled Statistical Relations Between Variables. This should prove especially valuable, as maching learning and data mining now play a significant role in computer science.

Second, 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 intuition involving 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.

Due to the emphasis on intution, there is lesser treatment of mathematical theory. This book does not define probability spaces in the "mini-measure theory" taken by most texts. However, all models and so on are described precisely in terms of random variables and distributions. And the material is somewhat more mathematical than most at this level in the sense that it makes extensive usage of linear algebra.

Finally, the R statistical/data manipulation 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. It is recommended that my online tutorial on R programming, *R for Programmers* (`http://heather.cs.ucdavis.edu/~matloff/R/RProg.pdf`), be used as a supplement.

As prerequisites, the student must know calculus, basic matrix algebra, and have 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.

A couple of points regarding computer usage:

- In the mathematical exercises, the instructor is urged to require that the students not only do the mathematical derivations but also check their results by writing R simulation code. This gives the students better intuition, and has the huge practical benefit that its gives partial confirmation that the student's answer is correct.
- In the chapters on statistics, it is crucial that students apply the concepts in thought-provoking exercises on real data. Nowadays there are many good sources for real data sets available. Here are a few to get you started:
  - UC Irvine Machine Learning Repository, `http://archive.ics.uci.edu/ml/datasets.html`
  - UCLA Statistics Dept. data sets, `http://www.stat.ucla.edu/data/`
  - Dr. B's Wide World of Web Data, `http://research.ed.asu.edu/multimedia/DrB/Default.htm`
  - StatSci.org, at `http://www.statsci.org/datasets.html`
  - University of Edinburgh School of Informatics, `http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html`

  Note that R has the capability of reading files on the Web, e.g.

```
> z <- read.table("http://heather.cs.ucdavis.edu/~matloff/z")
```

This work is licensed under a Creative Commons Attribution-No Derivative Works 3.0 United States License. The details may be viewed at `http://creativecommons.org/licenses/by-nd/3.0/us/`, but in essence it states that you are free to use, copy and distribute the work, but you must attribute the work to me and not "alter, transform, or build upon" it. If you are using the book, either in teaching a class or for your own learning, I would appreciate your informing me. I retain copyright in all non-U.S. jurisdictions, but permission to use these materials in teaching is still granted, provided the licensing information here is displayed.

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 mechanical 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

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 at 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.

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 9.)

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 $(1-p)^2$

Thus

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

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

2.2 The Crucial Notion of a Repeatable Experiment

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

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

Unfortunately, the notion of sample space becomes mathematically tricky when developed for more complex probability models. Indeed, it requires graduate-level math. And much worse, one loses all the intuition. 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 3.19.

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,

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

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]
- 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.**

[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.

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

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

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

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

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

Here are some more very important definitions and properties:

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

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

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

- If A and B are not disjoint, then

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{2.3}$$

  In the disjoint case, that subtracted term is 0, so (2.3) reduces to (2.2).

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

$$P(A \text{ and } B) = P(A) \cdot P(B) \tag{2.4}$$

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

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

- 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.

- If A and B are not independent, the equation (2.4) generalizes to

$$P(A \text{ and } B) = P(A)P(B|A) \tag{2.5}$$

  Note that if A and B actually are independent, then $P(B|A) = P(B)$, and (2.5) reduces to (2.4).

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. (2.4) and (2.5), 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, (2.4) and 2.5 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 (2.4) and (2.5)."

Be ready for this!

This mailing tube metaphor will be mentioned often, such as in Section 3.4.3 .

2.5 Basic Probability Computations: ALOHA Network Example

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

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

- We have n network nodes, sharing a common communications channel.

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

Now, in Equation (2.1) we found that

$$P(X_1 = 2) = p^2 + (1-p)^2 = 0.52 \tag{2.6}$$

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

$$\begin{aligned} P(X_1 = 2) &= P(\underbrace{C_1 \text{ and } C_2}\text{ or }\underbrace{\text{not } C_1 \text{ and not } C_2}) && (2.7)\\ &= P(C_1 \text{ and } C_2) + P(\text{ not } C_1 \text{ and not } C_2) \text{ (from (2.2)} && (2.8)\\ &= P(C_1)P(C_2) + P(\text{ not } C_1)P(\text{ not } C_2) \text{ (from (2.4)} && (2.9)\\ &= p^2 + (1-p)^2 && (2.10) \end{aligned}$$

(The underbraces in (2.7) 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:

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

(2.8): 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.

(2.9): The two nodes act physically independently of each other. Thus the events C_1 and C_2 are stochastically independent, so we applied (2.4). Then we did the same for D_1 and D_2.

Note carefully that in Equation (2.7), our first step was to **"break big events down into small events,"** in this case breaking the event $\{X_1 = 2\}$ down into the events C_1 and C_2 and D_1 and D_2. This is a central part of most probability computations. In calculating a probability, ask yourself, **"How can it happen?"**

Note that the ideas, "break big events down into small events" and asking "'How can it happen?" should be considered "mailing tubes" too.

Good tip: When you solve problems like this, write out the *and* and *or* conjunctions like I've done above. This helps!

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:

$$\begin{aligned} P(X_2 = 2) &= P(X_1 = 0 \text{ and } X_2 = 2 \text{ or } X_1 = 1 \text{ and } X_2 = 2 \text{ or } X_1 = 2 \text{ and } X_2 = 2) \\ &= P(X_1 = 0 \text{ and } X_2 = 2) \\ &+ P(X_1 = 1 \text{ and } X_2 = 2) \\ &+ P(X_1 = 2 \text{ and } X_2 = 2) \end{aligned} \tag{2.11}$$

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 (2.5). 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

$$P(X_1 = 1 \text{ and } X_2 = 2) = P(X_1 = 1)P(X_2 = 2|X_1 = 1) \tag{2.12}$$

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

$$P(X_1 = 1) = 2p(1 - p) = 0.48 \tag{2.13}$$

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

$$P(X_2 = 2|X_1 = 1) = q[p^2 + (1-p)^2] = 0.41 \tag{2.14}$$

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 one more; 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 (2.5), we know that

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

We computed both numerator and denominator here before, in Equations (2.12) and (2.11), 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 (2.15), 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.

Note by the way that events involving X_2 are NOT independent of those involving X_1. For instance, we found in (2.15) that

$$P(X_1 = 1|X_2 = 2) = 0.43 \tag{2.16}$$

yet from (2.13) we have

$$P(X_1 = 1) = 0.48. \tag{2.17}$$

2.6 Bayes' Rule

Following (2.15) above, we noted that the ingredients had already been computed, in (2.12) and (2.11). If we go back to the derivations in those two equations and substitute in (2.15), we have

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

$$= \frac{P(X_1 = 1 \text{ and } X_2 = 2)}{P(X_1 = 1 \text{ and } X_2 = 2) + P(X_1 = 2 \text{ and } X_2 = 2)} \tag{2.19}$$

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

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

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)} \tag{2.21}$$

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

2.7 ALOHA in the Notebook Context

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

- Run the network for two epochs, starting with both nodes active, the first time, and write the outcome on the first line of the notebook.
- Run the network for two epochs, starting with both nodes active, the second time, and write the outcome on the second line of the notebook.
- Run the network for two epochs, starting with both nodes active, the third time, and write the outcome on the third line of the notebook.
- Run the network for two epochs, starting with both nodes active, the fourth time, and write the outcome on the fourth line of the notebook.
- 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 $X_2 = 2$	$X_2 = 2 \mid X_1 = 2$
1	Yes	No	No	No
2	No	No	No	NA
3	Yes	Yes	Yes	Yes
4	Yes	No	No	No
5	Yes	Yes	Yes	Yes
6	No	No	No	NA
7	No	Yes	No	NA

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

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

- Among those first seven lines in the notebook, 4/7 of them have $X_1 = 2$. After many, many lines, this fraction will be approximately 0.52.
- Among those first seven lines in the notebook, 3/7 of them have $X_2 = 2$. After many, many lines, this fraction will be approximately 0.47.[5]
- Among those first seven lines in the notebook, 3/7 of them have $X_1 = 2 \text{ and } X_2 = 2$. 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 Example: Divisibility of Random Integers

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

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

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

$$
\begin{aligned}
P(N=2) &= P(\text{fail in step 5 and succeed in step 4}) \quad (\text{“How can it happen?”}) && (2.22)\\
&= P(\text{fail in step 5}) \quad \text{P(succeed in step 4 | fail in step 5)} \quad ((2.5)) && (2.23)\\
&= [1-q(5)]q(4) && (2.24)\\
&= \frac{4}{5} \cdot \frac{1}{4} && (2.25)\\
&= \frac{1}{5} && (2.26)
\end{aligned}
$$

But there's more.

First, note that q(i) is either equal or approximately equal to 1/i. Then following the derivation in (2.22), you'll find that

$$P(N=j) \approx \frac{1}{5} \tag{2.27}$$

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

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

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

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

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

2.9 Example: A Simple Board Game

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

A token advances according to the roll of a single die. If a player lands on square 3, he/she gets a bonus turn. Let's find the probability that a player has yet to make a complete circuit of the board after the first

turn (including the bonus, if any). Let R denote his first roll, and let B be his bonus if there is one, with B being set to 0 if there is no bonus. Then

$$\begin{aligned}
P(\text{no complete circuit}) &= P(R+B \leq 7) && (2.28)\\
&= P(R \leq 6, R \neq 3, B = 0 \text{ or } R = 3, B \leq 4) && (2.29)\\
&= P(R \leq 6, R \neq 3, B = 0) + P(R = 3, B \leq 4) && (2.30)\\
&= P(R \leq 6, R \neq 3) + P(R = 3, B \leq 4) && (2.31)\\
&= P(R \leq 6, R \neq 3) + P(R = 3)\ P(B \leq 4) && (2.32)\\
&= \frac{5}{6} + \frac{1}{6} \cdot \frac{4}{6} && (2.33)\\
&= \frac{17}{18} && (2.34)
\end{aligned}$$

According to a telephone report of the game, you hear that on A'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.

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

$$\begin{aligned}
P(B > 0 | R + B = 4) &= \frac{P(R+B=4, B>0)}{P(R+B=4)} && (2.35)\\
&= \frac{P(R+B=4, B>0)}{P(R+B=4, B>0) + P(R+B=4, B=0)} && (2.36)\\
&= \frac{P(R=3, B=1)}{P(R=3, B=1) + P(R=4)} && (2.37)\\
&= \frac{\frac{1}{6} \cdot \frac{1}{6}}{\frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6}} && (2.38)\\
&= \frac{1}{7} && (2.39)
\end{aligned}$$

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.

2.10 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.
- Assume the bus is so large that it never becomes full, so the new passengers can always get on.
- Suppose the bus is empty when it arrives at its first stop.

Let L_i denote the number of passengers on the bus as it *leaves* its ith stop, i = 1,2,3,... Let's find some probabilities, say $P(L_2 = 0)$.

For convenience, let B_i denote the number of new passengers who board the bus at the ith stop. Then

$$\begin{aligned} P(L_2 = 0) &= P(B_1 = 0 \text{ and } L_2 = 0 \text{ or } B_1 = 1 \text{ and } L_2 = 0 \text{ or } B_1 = 2 \text{ and } L_2 = 0) && (2.40)\\ &= \sum_{i=0}^{2} P(B_1 = i \text{ and } L_2 = 0) && (2.41)\\ &= \sum_{i=0}^{2} P(B_1 = i) P(L_2 = 0 | B_1 = i) && (2.42)\\ &= 0.5^2 + (0.4)(0.2)(0.5) + (0.1)(0.2^2)(0.5) && (2.43)\\ &= 0.292 && (2.44) \end{aligned}$$

2.11 Solution Strategies

The example in Section 2.9 shows typical stategies in exploring solutions to probability problems, such as:

- Name what seem to be the important variables, in this case R and B.
- Write the given probability in terms of those named variables, e.g.

$$P(\text{no complete circuit}) = P(R + B \leq 7) \tag{2.45}$$

above.

- Ask the famous question, "How can it happen?" In the above case:

$$P(R \leq 6, R \neq 3, B = 0 \text{ or } R = 3, B \leq 4) \tag{2.46}$$

- 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. 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"?)

- 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 Section 2.5. After you gain more experience, you can start skipping steps, but not in the initial learning period.

- Solving probably problmes—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. Just try some cases until you find one that works, in the sense that you can evaluate both factors.

2.12 Simulation

Note to readers: The R simulation examples in this book provide a valuable supplement to your developing insight into this material.

To learn about the syntax (e.g. <- as the assignment operator), see the first chapter of my book on R programming, at http://heather.cs.ucdavis.edu/~matloff/R/NMRIntro.pdf.

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).

2.12.1 Simulation of the ALOHA Example

Following is a computation via simulation of the *approximate* value of $P(X_1 = 2)$, $P(X_2 = 2)$ and $P(X_2 = 2|X_1 = 1)$, using the R statistical language, the language of choice of professional statisticans. It is open source, it's statistically correct (not all statistical packages are so), has dazzling graphics capabilities, etc.

```
# 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 (i 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
```

```
18        numactive <- X1
19        if (X1 == 1 && runif(1) < q) numactive <- numactive + 1
20        # send?
21        if (numactive == 1)
22           if (runif(1) < p) X2 <- 0
23           else X2 <- 1
24        else {  # numactive = 2
25           numsend <- 0
26           for (i in 1:2)
27              if (runif(1) < p) numsend <- numsend + 1
28           if (numsend == 1) X2 <- 1
29           else X2 <- 2
30        }
31        if (X2 == 2) countx2eq2 <- countx2eq2 + 1
32        if (X1 == 1) {  # do tally for the cond. prob.
33           countx1eq1 <- countx1eq1 + 1
34           if (X2 == 2) countx2eq2givx1eq1 <- countx2eq2givx1eq1 + 1
35        }
36     }
37     # print results
38     cat("P(X1 = 2):",countx1eq2/nreps,"\n")
39     cat("P(X2 = 2):",countx2eq2/nreps,"\n")
40     cat("P(X2 = 2 | X1 = 1):",countx2eq2givx1eq1/countx1eq1,"\n")
41  }
```

Note that each of the **nreps** iterations of the main **for** loop is analogous to one line in our hypothetical notebook. So, the 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.

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.

Remember, simulation results are only approximate. The larger the value we use for **nreps**, the more accurate our simulation results are likely to be. The question of how large we need to make **nreps** will be addressed in a later chapter.

Also: Keep in mind that we did NOT use (2.21) 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.

2.12.2 Rolling Dice

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

```
1  # roll d dice; find P(total = k)
2
```

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

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.

2.12.3 Improving the Code

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

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

Here the code

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

simulates tossing the die d times (the argument **replace** says this is sampling with replacement, so for instance we could get two 6s). That returns a d-element array, and we then call R's built-in function **sum()** to find the total of the d dice.

The second version of the code here is more compact and easier to read. It also eliminates one explicit loop, which is the key to writing fast code in R.

Actually, further improvements are possible. Consider this code:

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

There is quite a bit going on here.

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

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

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).[6]

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.

[6]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.

Even better:

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

probtotk <- function(d,k,nreps) {
   # do the experiment nreps times
   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.[7] Note, though, that this comes at the expense of using more memory.

2.12.4 Example: Bus Ridership, cont'd.

Consider the example in Section 2.10. 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 similated:

```
nreps <- 10000
nstops <- 10
count <- 0
for (i in 1:nreps) {
   passengers <- 0
   for (j in 1:nstops) {
      alight <- 0
      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.

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

2.12.5 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 8.

2.13 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.

2.13.1 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 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

$$P(1 \text{ king}) = \frac{4 \cdot \binom{48}{4}}{\binom{52}{5}} = 0.299 \tag{2.47}$$

The same reasoning gives us

$$P(2 \text{ hearts}) = \frac{\binom{13}{2} \cdot \binom{39}{3}}{\binom{52}{5}} = 0.274 \tag{2.48}$$

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 at your option call a user-specified function on each combination. This allows you to save a lot of computational work. See the examples in R's online documentation.

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

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

2.13.2 "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."[8] 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**

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

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.[9]. Suppose the business has a total of 20 products available for sale. What percentage of potential rules have three or fewer antecedents?[10]

For each k = 1,...,19, there are $\binom{20}{k}$ possible sets of antecedents, thus this many possible rules. The fraction of potential rules using three or fewer antecedents is then

$$\frac{\sum_{k=1}^{3} \binom{20}{k} \cdot \binom{20-k}{1}}{\sum_{k=1}^{19} \binom{20}{k} \cdot \binom{20-k}{1}} = \frac{23180}{10485740} = 0.0022 \tag{2.49}$$

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.

2.13.3 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 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

$$\frac{13!}{6!7!} \cdot \frac{7!}{5!2!} \cdot \frac{2!}{2!0!} \tag{2.50}$$

That reduces to

$$\frac{13!}{6!5!2!} \tag{2.51}$$

The same reasoning yields the following:

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

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

Multinomial Coefficients: Suppose we have c objects in r categories, with c_i objects in category i, i = 1,...,r. Then the number of ways to arrange them is

$$\frac{c!}{c_1!...c_r!} \tag{2.52}$$

2.13.4 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

$$\frac{52!}{13!13!13!13!} \tag{2.53}$$

possible deals. 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.

$$\frac{48!}{13!13!13!9!} \tag{2.54}$$

Thus the desired probability is

$$\frac{\frac{48!}{13!13!13!9!}}{\frac{52!}{13!13!13!13!}} = 0.00264 \tag{2.55}$$

Exercises

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

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

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

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

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

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

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

3. Consider the example of association rules in Section 2.13.2. How many two-antecedent, two-consequent rules are possible from 20 items? Express your answer in terms of combinatorial ("n choose k") symbols.

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

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

6. Say we choose six cards from a standard deck, one at a time WITHOUT replacement. Let N be the number of kings we get. Does N have a binomial distribution? Choose one: (i) Yes. (ii) No, since trials are not independent. (iii) No, since the probability of success is not constant from trial to trial. (iv) No, since the number of trials is not fixed. (v) (ii) and (iii). (iv) (ii) and (iv). (vii) (iii) and (iv).

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

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

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

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

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

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

$$\binom{n}{k} = \binom{n-1}{k} + \binom{c}{d} \tag{2.56}$$

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

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

(a) Find the probability of winning.

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

(c) How could we construct such a die?

14. Consider the ALOHA simulation example in Section 2.12.1.

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

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

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

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

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

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

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

18. Suppose we have the track record of a certain weather forecaster. Of the days for which he predicts rain, a fraction c actually do haverain. Among days for which he predicts no rain, he is correct a fraction is d of the time. Among all days, he predicts rain g of the time, and predicts no rain 1-g of the time. Find P(he predicted rain | it does rain), P(he predicts wrong) and P(it does rain — he was wrong). Write R simulation

code to verify. (Partial answer: For the case c = 0.8, d = 0.6 and g = 0.2, P(he predicted rain | it does rain) = 1/3.)

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

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

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

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

(a) P(you have no Wheats).

(b) P(you have seven Wheats).

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

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

20. In the board game example in Section 2.9, suppose that the telephone report is that A ended up at square 1 after his first turn. Find the probability that he got a bonus.

21. Consider the bus ridership example in Section 2.10 of the text. Suppose the bus is initially empty, and let X_n denote the number of passengers on the bus just after it has left the n^{th} stop, n = 1,2,... Find the following:

(a) $P(X_2 = 1)$

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

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

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 $\{1,2,3,4,5,6\}$. This is a finite set.

In the ALOHA example, X_1 and X_2 each take on values in the set $\{0,1,2\}$, 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 N can take on values in 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. This 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 U and V 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 Expected Value

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 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 (i.e. the mean). Yet P(H = 500) turns out to be about 0.025. In other words, we certainly should not "expect" H to be 500.

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

[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.

3.4.1 Intuitive Definition

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

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

$$\lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} \tag{3.1}$$

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

3.4.2 Computation and Properties of Expected Value

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

$$\begin{aligned} E(X) &= \lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} && (3.2)\\ &= \lim_{n\to\infty} \frac{0 \cdot K_{0n} + 1 \cdot K_{1n} + 2 \cdot K_{2n} ... + 10 \cdot K_{10,n}}{n} && (3.3)\\ &= \sum_{i=0}^{10} i \cdot \lim_{n\to\infty} \frac{K_{in}}{n} && (3.4) \end{aligned}$$

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

$$2 + 3 + 1 + 2 + 1 = 2 \times 2 + 2 \times 1 + 1 \times 3 \tag{3.5}$$

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,

$$E(X) = \sum_{i=0}^{10} i \cdot P(X = i) \tag{3.6}$$

So in general we have a

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

$$E(X) = \sum_{c \in A} cP(X = c) \tag{3.7}$$

Note that (3.7) is the formula we'll use. The preceding equations were derivation, to motivate the formula. Note too that 3.7 is not the *definition* of expected value; that was in 3.1. It is quite important to distinguish between all of these, in terms of goals.

It will be shown in Section 3.13.4 that in our example above in which X is the number of heads we get in 10 tosses of a coin,

$$P(X = i) = \binom{10}{i} 0.5^i (1 - 0.5)^{10-i} \tag{3.8}$$

So

$$E(X) = \sum_{i=0}^{10} i \binom{10}{i} 0.5^i (1 - 0.5)^{10-i} \tag{3.9}$$

It turns out that E(X) = 5.

For X in our dice example,

$$E(X) = \sum_{c=1}^{6} c \cdot \frac{1}{6} = 3.5 \tag{3.10}$$

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,

$$E(S) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + ...12 \cdot \frac{1}{36} = 7 \tag{3.11}$$

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

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

$$E(N) = \sum_{c=1}^{\infty} c \cdot \frac{1}{2^c} = 2 \tag{3.12}$$

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

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 (3.11); 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:

$$E(U + V) = E(U) + E(V) \tag{3.13}$$

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:

Property C: For any random variable U and any constants *a* and *b*,

$$E(aU + b) = aE(U) + b \tag{3.14}$$

Note also that this implies that for any constant *b*, we have

$$E(b) = b \tag{3.15}$$

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 (3.14) with $a = \frac{9}{5}$ and b = 32.

Another important point:

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

$$E(UV) = EU \cdot EV \tag{3.16}$$

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

$$E(D) = 3.5^2 = 12.25 \tag{3.17}$$

Equation (3.16) doesn't have an easy "notebook proof." It is proved in Section 5.3.1.1.

Consider a function g() of one variable, and let W = g(X). W is then a random variable too. Say X takes on values in A, as in (3.7). Then W takes on values in $B = \{g(c) : c \epsilon A\}$. Define

$$A_d = \{c : c \in A, g(c) = d\} \tag{3.18}$$

Then

$$P(W = d) = P(X \in A_d) \tag{3.19}$$

so

$$\begin{aligned} E[g(X)] &= E(W) && (3.20)\\ &= \sum_{d \in B} dP(W = d) && (3.21)\\ &= \sum_{d \in B} d \sum_{c \in A_d} P(X = c) && (3.22)\\ &= \sum_{c \in A} g(c) P(X = c) && (3.23) \end{aligned}$$

Property E:

If E[g(X)] exists, then

$$E[g(X)] = \sum_c g(c) \cdot P(X = c) \tag{3.24}$$

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, (3.24) tells us to make a list of values that W and take on, 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 (3.24), we find that

$$E(\sqrt{X}) = \frac{1}{6} \sum_{i=1}^{6} \sqrt{i} \tag{3.25}$$

3.4.3 "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 (3.13) right away.

As discussed in Section 2.4, these properties are "mailing tubes." For instance, (3.13) 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 (3.13)." Similarly, (3.24) 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 (3.24)."

You will encounter "mailing tubes" throughout this book. For instance, (3.32) below is a very important "mailing tube." Constatly remind yourself—"Remember the 'mailing tubes'!"

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

The expected value is intended as a **measure of central tendency**, 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.

(**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 W to be 91/6 in this last example; 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 (3.1) 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 (3.1) 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, (3.1) means plenty. Say X is the amount a gambler wins on a play of a roulette wheel, and suppose (3.1) is equal to \$1.88. Then after, say, 1000 plays of the wheel (not necessarily by the same gambler), the casino knows it will have paid out a total about about \$1,880. So if the casino charges, say \$1.95 per play, it will have made a profit of about \$70 over those 1000 plays. It might be a bit more or less than that amount, but the casino can be pretty sure that it will be around \$70, and they can plan their

business accordingly.

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

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

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

Nevertheless, the mean has certain mathematical properties, such as (3.13), 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.

3.5 Variance

As in Section 3.4, 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.**

3.5.1 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*

$$Var(U) = E[(U - EU)^2] \tag{3.26}$$

For X in the die example, this would be

$$Var(X) = E[(X - 3.5)^2] \tag{3.27}$$

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 (3.24) with $g(c) = (c - 3.5)^2$:

$$Var(X) = \sum_{c=1}^{6} (c - 3.5)^2 \cdot \frac{1}{6} = 2.92 \tag{3.28}$$

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.

The properties of E() in (3.13) and (3.14) can be used to show:

Property F:

$$Var(U) = E(U^2) - (EU)^2 \tag{3.29}$$

The term $E(U^2)$ is again evaluated using (3.24).

Thus for example, if X is the number of dots which come up when we roll a die. Then, from (3.29),

$$Var(X) = E(X^2) - (EX)^2 \tag{3.30}$$

Let's find that first term (we already know the second is 3.5). From (3.24),

$$E(X^2) = \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} = \frac{91}{6} \tag{3.31}$$

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

Remember, though, that (3.29) is a shortcut formula for finding the variance, not the *definition* of variance.

An important behavior of variance is:

Property G:

$$Var(cU) = c^2 Var(U) \tag{3.32}$$

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 (3.32). Define V = cU. Then

$$\begin{aligned}
Var(V) &= E[(V - EV)^2] \text{ (def.)} && (3.33)\\
&= E\{[cU - E(cU)]^2\} \text{ (subst.)} && (3.34)\\
&= E\{[cU - cEU]^2\} \text{ ((3.14))} && (3.35)\\
&= E\{c^2[U - EU]^2\} \text{ (algebra)} && (3.36)\\
&= c^2 E\{[U - EU]^2\} \text{ ((3.14))} && (3.37)\\
&= c^2 Var(U) \text{ (def.)} && (3.38)
\end{aligned}$$

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

Property H:

$$Var(U + d) = Var(U) \tag{3.39}$$

for any constant d.

Intuitively, the variance of a constant is 0—after all, it never varies! You can show this formally using (3.29):

$$Var(c) = E(c^2) - [E(c)]^2 = c^2 - c^2 = 0 \tag{3.40}$$

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 5.12.2 for details.)

As with expected values, the properties of variance discussed above, and also in Section 5.2.1 below, are key to the entire remainder of this book. You should notice immediately when you are in a setting in which they are applicable. For instance, if you see the variance of the sum of two random variables, you should instinctively think of (3.64) right away.

3.5.2 Intuition Regarding the Size of Var(X)

A billion here, a billion there, pretty soon, you're talking real money—attribted 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 suppose 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.

3.5.2.1 Chebychev's Inequality

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

$$P(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2} \tag{3.41}$$

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 (3.41) would be 1.5.

We'll prove the inequality in Section 3.18.

3.5.2.2 The Coefficient of Variation

Continuing our discussion of the magnitude of a variance, look at our remark following (3.41):

> 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:

$$\text{coef. of var.} = \frac{\sqrt{Var(X)}}{EX} \tag{3.42}$$

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.

3.6 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
>
> $$E(X) = p \tag{3.43}$$
>
> $$Var(X) = p(1-p) \tag{3.44}$$

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

$$EX = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = P(X = 1) = P(A) = p \tag{3.45}$$

The derivation for Var(X) is similar (use (3.29)).

3.7 A Combinatorial Example

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 can take on the values 4-0, 3-1, 2-2 and 1-3, i.e. 4, 2, 0 and -2. So,

$$ED = -2 \cdot P(D = -2) + 0 \cdot P(D = 0) + 2 \cdot P(D = 2) + 4 \cdot P(D = 4) \tag{3.46}$$

Now, using reasoning along the lines in Section 2.13, we have

$$P(D = -2) = P(M = 1 \text{ and } W = 3) = \frac{\binom{6}{1}\binom{3}{3}}{\binom{9}{4}} \tag{3.47}$$

After similar calculations for the other probabilities in (3.46), 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:

$$\begin{aligned} ED &= E(M - W) && (3.48)\\ &= E[M - (4 - M)] && (3.49)\\ &= E(2M - 4) && (3.50)\\ &= 2EM - 4 \text{ (from (3.14))} && (3.51) \end{aligned}$$

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

$$M = G_1 + G_2 + G_3 + G_4 \tag{3.52}$$

so

$$\begin{aligned} EM &= E(G_1 + G_2 + G_3 + G_4) && (3.53)\\ &= EG_1 + EG_2 + EG_3 + EG_4 \ [\text{ from } (3.13)] && (3.54)\\ &= P(G_1 = 1) + P(G_2 = 1) + P(G_3 = 1) + P(G_4 = 1) \ [\text{ from } (3.43)] && (3.55) \end{aligned}$$

Note carefully that the second equality here, which uses (3.13), is true in spite of the fact that the G_i are not independent. Equation (3.13) does not require independence.

Another key point is that, due to symmetry, $P(G_i = 1)$ is the same for all i. same expected value. (Note that we did not write a *conditional* probability here.) To see this, 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 is for, say, Alice, again 1/9.

Now,

$$P(G_1 = 1) = \frac{6}{9} = \frac{2}{3} \tag{3.56}$$

Thus

$$ED = 2 \cdot (4 \cdot \frac{2}{3}) - 4 = \frac{4}{3} \tag{3.57}$$

3.8 A Useful Fact

For a random variable X, consider the function

$$g(c) = E[(X - c)^2] \tag{3.58}$$

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:

$$g(c) = E[(X-c)^2] = E(X^2 - 2cX + c^2) = E(X^2) - 2cEX + c^2 \tag{3.59}$$

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

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

$$0 = -2EX + 2c \tag{3.60}$$

so the minimizing c is c = EX!

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

Moreover: Plugging c = EX into (3.59) shows that the minimum value of g(c) is $E(X-EX)^2]$, which is Var(X)!

3.9 Covariance

This is a topic we'll cover fully in Chapter 5, but at least introduce here.

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

$$Cov(U,V) = E[(U-EU)(V-EV)] \tag{3.61}$$

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

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

$$Cov(U,V) = E(UV) - EU \cdot EV \tag{3.62}$$

Also

$$Var(U+V) = Var(U) + Var(V) + 2Cov(U,V) \tag{3.63}$$

Suppose U and V are independent. Then (3.16) and (3.62) imply that Cov(U,V) = 0. In that case,

$$Var(U+V) = Var(U) + Var(V) \tag{3.64}$$

3.10 Expected Value, Etc. in the ALOHA Example

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

$$EX_1 = 0 \cdot P(X_1 = 0) + 1 \cdot P(X_1 = 1) + 2 \cdot P(X_1 = 2) = 1 \cdot 0.48 + 2 \cdot 0.52 = 1.52 \tag{3.65}$$

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) {
      numsend <- 0
      for (i in 1:2)
         if (runif(1) < p) numsend <- numsend + 1
      if (numsend == 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
         numsend <- 0
         for (i in 1:2)
            if (runif(1) < p) numsend <- numsend + 1
         if (numsend == 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 Back to the Board Game Example

Recall the board game in Section 2.9. Below is simulation code to find the probability in (2.35):

```
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) countbousngiven4 <- countbousngiven4 + 1
      }
   }
   return(countbousngiven4/count4)
}
```

3.12 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 a list of all the values U takes on, and their 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

$$\text{distribution of } X = \{(1, \frac{1}{6}), (2, \frac{1}{6}), (3, \frac{1}{6}), (4, \frac{1}{6}), (5, \frac{1}{6}), (6, \frac{1}{6})\} \tag{3.66}$$

Example: Recall the ALOHA example. There X_1 took on the values 1 and 2, with probabilities 0.48 and 0.52, respectively. So,

$$\text{distribution of } X_1 = \{(0, 0.00), (1, 0.48), (2, 0.52)\} \tag{3.67}$$

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

$$\text{distribution of } N = \{(1, \frac{1}{2}), (2, \frac{1}{4}), (3, \frac{1}{8}), ...\} \tag{3.68}$$

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*

$$p_V(k) = P(V = k) \tag{3.69}$$

for any value k which V can take on.

(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.)

Example: In (3.68),

$$p_N(k) = \frac{1}{2^k}, k = 1, 2, ... \tag{3.70}$$

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

$$p_S(k) = \begin{cases} \frac{1}{36}, & k = 2 \\ \frac{2}{36}, & k = 3 \\ \frac{3}{36}, & k = 4 \\ ... & \\ \frac{1}{36}, & k = 12 \end{cases} \tag{3.71}$$

It is important to note that there may not be some nice closed-form expression for p_V like that of (3.70). There was no such form in (3.71), nor is there in our ALOHA example for p_{X_1} and p_{X_2}.

3.13 Parameteric Families of pmfs

Consider plotting the curves sin(ct). For each c, we get the familiar sine function. For larger c, the curve is more "squished" and for c strictly between 0 and 1, we get a broadened sine curve. So we have a family

of sine curves of different proportions. We say the family is **indexed** by the **parameter** c, meaning, each c gives us a different member of the family, i.e. a different curve.

Probability mass functions, and in the next chapter, probability density functions, can also come in families, indexed by one or more parameters. We will discuss some of the famous 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.**

3.13.1 The Geometric Family of Distributions

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

$$p_N(k) = (1 - \frac{1}{2})^{k-1} \cdot \frac{1}{2}, k = 1, 2, ... \tag{3.72}$$

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 as "success."

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

$$p_N(k) = \left(1 - \frac{1}{6}\right)^{k-1} \frac{1}{6}, k = 1, 2, ... \tag{3.73}$$

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.

The tosses of the coin and the rolls of the die are known as **Bernoulli trials**, which is a sequence of independent 1-0-valued random variables B_i, i = 1,2,3,... B_i is 1 for success, 0 for failure, with success probability p. (Note that these are indicator random variables, introduced in Section 3.6.) 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

$$p_W(k) = (1 - p)^{k-1} p, k = 1, 2, ... \tag{3.74}$$

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.

It should make good intuitive sense to you that

$$E(W) = \frac{1}{p} \tag{3.75}$$

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 $r \leq s$,

$$\sum_{i=r}^{s} t^i = t^r \frac{1 - t^{s-r+1}}{1 - t} \tag{3.76}$$

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

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

$$\sum_{i=0}^{\infty} t^i = \frac{1}{1 - t} \tag{3.77}$$

To prove this, just take r = 0 and let $s \to \infty$ in (3.76).

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

$$\sum_{i=1}^{\infty} i t^{i-1} = \frac{1}{(1 - t)^2} \tag{3.78}$$

This is derived by applying $\frac{d}{dt}$ to (3.77).[3]

Deriving (3.75) is then easy, using (3.78):

[3] To be more carefully, we should differentiate (3.76) and take limits.

$$
\begin{aligned}
EW &= \sum_{i=1}^{\infty} i(1-p)^{i-1}p && (3.79)\\
&= p\sum_{i=1}^{\infty} i(1-p)^{i-1} && (3.80)\\
&= p \cdot \frac{1}{[1-(1-p)]^2} && (3.81)\\
&= \frac{1}{p} && (3.82)
\end{aligned}
$$

Using similar computations, one can show that

$$Var(W) = \frac{1-p}{p^2} \tag{3.83}$$

We can also find a closed-form expression for the cdf. For any positive integer m we have

$$
\begin{aligned}
F_W(m) &= P(W \leq m) && (3.84)\\
&= 1 - P(W > m) && (3.85)\\
&= 1 - P(\text{the first m trials are all failures}) && (3.86)\\
&= 1 - (1-p)^m && (3.87)
\end{aligned}
$$

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).

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]**).

The function **dgeom()** returns the pmf values of a geometric distribution, with the success probability being specified as the second argument. The first argument is a vector of k values for which you have a need to evaluate (3.74).

3.13.2 R Functions

Relevant functions for a geometrically distributed random variable X with success probability p are:

- **pgeom(q,p)**, to find $P(X \leq q)$
- **qgeom(q,p)**, to find c such that $P(X \leq c) = q$
- **rgeom(n,p)**, to generate n independent values of X

3.13.3 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 0.2, with the spaces being independent. Find ED.

To solve this problem, you might at first think that D follows a geometric distribution, but actually this is not the case, given that 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. If for instance the first space is occupied but the second one isn't, then N = 2. Then

$$D = \begin{cases} 11 - N, & N \leq 10 \\ N - 11, & N > 10 \end{cases} \tag{3.88}$$

Since D is a function of N, we can use (3.24):

$$ED = \sum_{i=1}^{10} (11 - i) 0.8^{i-1} 0.2 + \sum_{i=11}^{\infty} (i - 11) 0.8^{i-1} 0.2 \tag{3.89}$$

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

3.13.4 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

$$P(X=2) = \binom{5}{2} 0.5^2 (1-0.5)^3 = \binom{5}{2}/32 = 5/16 \tag{3.90}$$

For general n and p,

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.91}$$

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:

$$X = \sum_{i=1}^{n} B_i \tag{3.92}$$

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 3.6), so

$$EB_i = p \tag{3.93}$$

and

$$Var(B_i) = p(1-p) \tag{3.94}$$

Then the reader should use our earlier properties of E() and Var() in Sections 3.4 and 3.5 to fill in the details in the following derivations of the expected value and variance of a binomial random variable:

$$EX = E(B_1 + ..., +B_n) = EB_1 + ... + EB_n = np \tag{3.95}$$

and from (3.64),

$$Var(X) = Var(B_1 + ..., +B_n) = Var(B_1) + ... + Var(B_n) = np(1-p) \tag{3.96}$$

Again, (3.95) should make good intuitive sense to you.

3.13.5 R Functions

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

- **pbinom(q,k,p)**, to find $P(X \leq q)$
- **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

3.13.6 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.

Toward this end, 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

$$
\begin{aligned}
p_X(m) &= P(X = m) && (3.97)\\
&= \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} P(X = m \text{ and } Y = i) && (3.98)\\
&= \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} P(X = m|Y = i)\ P(Y = i) && (3.99)\\
&= \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} \binom{i}{m} 0.5^i \binom{k}{i} 0.5^k && (3.100)\\
&= 0.5^k \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} \frac{k!}{m!(i-m)!(k-i)!} 0.5^i && (3.101)
\end{aligned}
$$

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

3.13.7 Example: Analysis of Social Networks

One of the earliest—and now the simplest—modesl 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.

One entity of interest is the **degree distribution**, defined as follows. For each node, the number of links connected to it is called the **degree** of the node. Since degree is a random variable, we can ask about its distribution.

Clearly the degree distribution for a single node is binomial with parameters n-1 and p. But consider k nodes, and the total T of their degrees. Let's find the distribution of T.

That distribution is again binomial, but the number of trials is not $k\binom{n-1}{2}$, due to overlap. There are $\binom{k}{2}$ potential links among these k nodes, and each has $\binom{n-k}{2}$ potential links to the "outside world," i.e. to the remaining n-k nodes. So, the distribution of T is binomial with

$$k\binom{n-k}{2} + \binom{k}{2} \tag{3.102}$$

trials and success probability p.

3.13.8 The Poisson Family of Distributions

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

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, ... \tag{3.103}$$

It turns out that

$$EX = \lambda \tag{3.104}$$

$$Var(X) = \lambda \tag{3.105}$$

The derivations of these facts are similar to those for the geometric family in Section 3.13.1. One starts with the Maclaurin series expansion for e^t:

$$e^t = \sum_{i=0}^{\infty} \frac{t^i}{i!} \tag{3.106}$$

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 4.4.6.1.

3.13.9 R Functions

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

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

3.13.10 The Negative Binomial Family of Distributions

Recall that a typical example of the geometric distribution family (Section 3.13.1) 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:

$$\{N = 5\} = \{2 \text{ heads in the first } 4 \text{ tosses and head on the } 5^{th} \text{ toss}\} \tag{3.107}$$

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

$$P(2 \text{ heads in the first } 4 \text{ tosses}) = \binom{4}{2} \left(\frac{1}{2}\right)^4 \tag{3.108}$$

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

$$P(N = 5) = \binom{4}{2} \left(\frac{1}{2}\right)^5 = \frac{3}{16} \tag{3.109}$$

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

$$P(N = k) = \binom{k-1}{r-1} (1-p)^{k-r} p^r, k = r, r+1, ... \tag{3.110}$$

We can write

$$N = G_1 + ... + G_r \tag{3.111}$$

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

$$E(N) = r \cdot \frac{1}{p} \tag{3.112}$$

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

$$Var(N) = r \cdot \frac{1-p}{p^2} \tag{3.113}$$

3.13.11 The Power Law Family of Distributions

Here

$$p_X(k) = ck^{-\gamma},\ k = 1, 2, 3, ... \tag{3.114}$$

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:

$$1.0 = \sum_{k=1}^{\infty} ck^{-\gamma} \approx c \int_1^{\infty} k^{-\gamma}\, dk = c/(\gamma - 1) \tag{3.115}$$

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

$$p_X(k) = ck^{-\gamma}q^k \tag{3.116}$$

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.

3.14 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 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.

3.14.1 Example: a Coin Game

Life is unfair—former President Jimmie 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 kth 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 kth 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:

$$P(Y_6 > X_6) = \sum_{i=1}^{4} \sum_{j=0}^{i-1} P(Y_6 = i \text{ and } X_6 = j) \tag{3.117}$$

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,

$$P(Y_6 = i \text{ and } X_6 = j) = P(Y_6 = i) \cdot P(X_6 = j) \tag{3.118}$$

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

$$P(Y_6 = i) = \binom{4}{i} 0.5^i (1 - 0.5)^{4-i} \tag{3.119}$$

and

$$P(X_6 = j) = \binom{6}{j} 0.5^j (1 - 0.5)^{6-j} \tag{3.120}$$

We would then substitute (3.119) and (3.120) in (3.117). We could then evaluate it by hand, but it would be more convenient to use R's **dbinom()** function:

```
1 prob <- 0
2 for (i in 1:4)
```

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

3.14.2 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 numbr 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

$$p = \binom{4}{2} 0.5^4 = \frac{3}{8} \tag{3.121}$$

Thus using the fact that N has a geometric distribution,

$$P(N = 5) = (1 - p)^4 p = 0.057 \tag{3.122}$$

3.14.3 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.[5] For instance, the

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

probability that there is a successful transmission is equal to the probability that exactly one of the m nodes attempts to send,

$$\binom{m}{1} p(1-p)^{m-1} = mp(1-p)^{m-1} \tag{3.123}$$

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 (3.123).

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.

3.15 A Preview of Markov Chains

Here we introduce Markov chains, a topic covered in much more detail in Chapter 11. The case covered here will be that of discrete time, finite state space.

3.15.1 Example: ALOHA

A handy first example is 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, X_4 and X_5—don't have any impact. We can state that precisely:

> The quantity
>
> $$P(X_6 = j | X_1 = i_1, X_2 = i_2, X_3 = i_3, X_4 = i_4, X_5 = i) \tag{3.124}$$
>
> does not depend on $i_m, m = 1, ..., 4$. Thus we can write (3.124) 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 $P(X_{k+1} = j|X_k = i)$. We denote this probability by p_{ij}, and refer to it as the **transition probability** from state i to state j.

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

$$P = \begin{pmatrix} (1-q)^2 + 2q(1-q)p & 2q(1-q)(1-p) + 2q^2p(1-p) & q^2[p^2 + (1-p)^2] \\ (1-q)p & 2qp(1-p) + (1-q)(1-p) & q[p^2 + (1-p)^2] \\ 0 & 2p(1-p) & p^2 + (1-p)^2 \end{pmatrix} \tag{3.125}$$

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 $p^2 + (1-p)^2$.

Let N_{it} denote the number of times we have visited state i during times 1,...,t. Than as discussed in Section 11.1.2, in typical applications

$$\pi_i = \lim_{t \to \infty} \frac{N_{it}}{t} \tag{3.126}$$

exists for each state i. Under a couple more conditions, we have the stronger result,

$$\lim_{t \to \infty} P(X_t = i) = \pi_i \tag{3.127}$$

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

And it turns out that the π_i are easy to find (in the case of finite state spaces, the subject of this section here), by solving a system of linear equations, (11.8) with the constraint (11.10). R code to do all this, **findpi1()**, is provided in Section 11.1.2.2.

For the ALOHA example here, with p = 0.4 and q = 0.3, the solution is $\pi_0 = 0.47$, $\pi_1 = 0.43$ and $\pi_2 = 0.10$.

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

$$0 \cdot 0.47 + 1 \cdot 0.43 + 2 \cdot 0.10 = 0.63 \tag{3.128}$$

3.15.2 Example: Die Game

As another example of Markov chains, consider the following game. One repeatedly rolls a die, keeping a running total. Each time the total exceeds 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.

This process clearly satisfies the Markov property, and 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 π_i:

```
p <- matrix(rep(0,100),nrow=10)
onesixth <- 1/6
for (i in 1:10) {
   for (j in 1:6) {
      k <- i + j
      if (k > 10) k <- k - 10
      p[i,k] <- onesixth
   }
}
findpi1(p)
```

Well, guess what! 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.

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.

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)
onehalf <- 1/2
p[1,1] <- onehalf
p[1,2] <- onehalf
p[2,3] <- onehalf
p[2,1] <- onehalf
p[3,1] <- 1
findpi1(p)
```

It turns out that

$$\pi = (0.5714286, 0.2857143, 0.1428571) \tag{3.129}$$

So, in the long run, about 57.1% of our rolls 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 rolls we do while in state 2, half will be heads, so half will be wins. In other words, about 0.071 of our rolls will be wins. And THAT figure answers our original question, through the following reasoning:

Think of, say, 10000 rolls. There will be about 710 wins sprinkled among those 10000 rolls. Thus the average number of rolls 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 rolls.

3.15.3 Example: Bus Ridership Problem

Consider the bus ridership problem in Section 2.10. 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:

$$p_{00} = 0.5 \tag{3.130}$$

$$p_{01} = 0.4 \tag{3.131}$$

$$p_{20} = (0.2)^2(0.5) = 0.02 \tag{3.132}$$

$$p_{19,20} = 0.5 \tag{3.133}$$

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

$$\sum_{i=0}^{20} \pi_i i \tag{3.134}$$

and the long-run average number of would-be passengers who fail to board the bus,

$$1 \cdot [\pi_{19}(0.1) + \pi_{20}(0.4)] + 2 \cdot [\pi_{20}(0.1)] \tag{3.135}$$

3.16 A Cautionary Tale

3.16.1 Trick Coins, Tricky Example

Suppose we have two trick coins in a box. They look identical, but one of them, denoted coin 1, is heavily weighted toward heads, with a 0.9 probability of heads, while the other, denoted coin 2, is biased in the opposite direction, with a 0.9 probability of tails. Let C_1 and C_2 denote the events that we get coin 1 or coin 2, respectively.

Our experiment consists of choosing a coin at random from the box, and then tossing it n times. Let B_i denote the outcome of the i^{th} toss, i = 1,2,3,..., where $B_i = 1$ means heads and $B_i = 0$ means tails. Let $X_i = B_1 + ... + B_i$, so X_i is a count of the number of heads obtained through the i^{th} toss.

The question is: "Does the random variable X_i have a binomial distribution?" Or, more simply, the question is, "Are the random variables B_i independent?" To most people's surprise, the answer is No (to both questions). Why not?

The variables B_i are indeed 0-1 variables, and they have a common success probability. But they are not independent! Let's see why they aren't.

Consider the events $A_i = \{B_i = 1\}$, i = 1,2,3,... In fact, just look at the first two. By definition, they are independent if and only if

$$P(A_1 \text{ and } A_2) = P(A_1)P(A_2) \tag{3.136}$$

First, what is $P(A_1)$? **Now, wait a minute!** Don't answer, "Well, it depends on which coin we get," because this is NOT a conditional probability. Yes, the *conditional* probabilities $P(A_1|C_1)$ and $P(A_1|C_2)$ are 0.9 and 0.1, respectively, but the *unconditional* probability is $P(A_1) = 0.5$. You can deduce that either by the symmetry of the situation, or by

$$P(A_1) = P(C_1)P(A_1|C_1) + P(C_2)P(A_1|C_2) = (0.5)(0.9) + (0.5)(0.1) = 0.5 \tag{3.137}$$

You should think of all this in the notebook context. Each line of the notebook would consist of a report of three things: which coin we get; the outcome of the first toss; and the outcome of the second toss. (Note by the way that in our experiment we don't know which coin we get, but conceptually it should have a column

in the notebook.) If we do this experiment for many, many lines in the notebook, about 90% of the lines in which the coin column says "1" will show Heads in the second column. But 50% of the lines *overall* will show Heads in that column.

So, the right hand side of Equation (3.136) is equal to 0.25. What about the left hand side?

$$\begin{aligned} P(A_1 \text{ and } A_2) &= P(A_1 \text{ and } A_2 \text{ and } C_1) + P(A_1 \text{ and } A_2 \text{ and } C_2) && (3.138)\\ &= P(A_1 \text{ and } A_2|C_1)P(C_1) + P(A_1 \text{ and} A_2|C_2)P(C_2) && (3.139)\\ &= (0.9)^2(0.5) + (0.1)^2(0.5) && (3.140)\\ &= 0.41 && (3.141) \end{aligned}$$

Well, 0.41 is not equal to 0.25, so you can see that the events are not independent, contrary to our first intuition. And that also means that X_i is not binomial.

3.16.2 Intuition in Retrospect

To get some intuition here, think about what would happen if we tossed the chosen coin 10000 times instead of just twice. If the tosses were independent, then for example knowledge of the first 9999 tosses should not tell us anything about the 10000th toss. But that is not the case at all. After 9999 tosses, we are going to have a very good idea as to which coin we had chosen, because by that time we will have gotten about 9000 heads (in the case of coin C_1) or about 1000 heads (in the case of C_2). In the former case, we know that the 10000th toss is likely to be a head, while in the latter case it is likely to be tails. **In other words, earlier tosses do indeed give us information about later tosses, so the tosses aren't independent.**

3.16.3 Implications for Modeling

The lesson to be learned is that independence can definitely be a tricky thing, not to be assumed cavalierly. And in creating probability models of real systems, we must give very, very careful thought to the conditional and unconditional aspects of our models—-it can make a huge difference, as we saw above. Also, the conditional aspects often play a key role in formulating models of nonindependence.

This trick coin example is just that—tricky—but similar situations occur often in real life. If in some medical study, say, we sample people at random from the population, the people are independent of each other. But if we sample *families* from the population, and then look at children within the families, the children within a family are not independent of each other.

3.17 Why Not Just Do All Analysis by Simulation?

Now that computer speeds are so fast, one might ask why we need to do mathematical probability analysis; why not just do everything by simulation? There are a number of reasons:

- Even with a fast computer, simulations of complex systems can take days, weeks or even months.
- Mathematical analysis can provide us with insights that may not be clear in simulation.
- Like all software, simulation programs are prone to bugs. The chance of having an uncaught bug in a simulation program is reduced by doing mathematical analysis for a special case of the system being simulated. This serves as a partial check.
- Statistical analysis is used in many professions, including engineering and computer science, and in order to conduct meaningful, useful statistical analysis, one needs a firm understanding of probability principles.

An example of that second point arose in the computer security research of a graduate student at UCD, C. Senthilkumar, who was working on a way to more quickly detect the spread of a malicious computer worm. He was evaluating his proposed method by simulation, and found that things "hit a wall" at a certain point. He wasn't sure if this was a real limitation; maybe, for example, he just wasn't running his simulation on the right set of parameters to go beyond this limit. But a mathematical analysis showed that the limit was indeed real.

3.18 Proof of Chebychev's Inequality

To prove (3.41), let's first state and prove Markov's Inequality: For any nonnegative random variable Y,

$$P(Y \geq d) \leq \frac{EY}{d} \tag{3.142}$$

To prove (3.142), let Z be the indicator random variable for the event $Y \geq d$ (Section 3.6).

Now note that

$$Y \geq dZ \tag{3.143}$$

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

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

So

$$EY \geq dEZ \tag{3.144}$$

(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 (3.144) is $dP(Y \geq d)$, so (3.142) follows.

Now to prove (3.41), define

$$Y = (X - \mu)^2 \tag{3.145}$$

and set $d = c^2\sigma^2$. Then (3.142) says

$$P[(X - \mu)^2 \geq c^2\sigma^2] \leq \frac{E[(X - \mu)^2]}{c^2\sigma^2} \tag{3.146}$$

Since

$$(X - \mu)^2 \geq c^2\sigma^2 \text{ if and only if } |X - \mu| \geq c\sigma \tag{3.147}$$

the left-hand side of (3.146) is the same as the left-hand side of (3.41). The numerator of the right-hand size of (3.146) is simply Var(X), i.e. σ^2, so we are done.

3.19 Reconciliation of Math and Intuition (optional section)

Here is a more theoretical definition of probability, as opposed to the intuitive "notebook" idea in this book. The definition is an abstraction of the notions of events (the sets A in $\mathcal{W}$ below) and probabilities of those events (the values of the function P(A)):

Definition 8 *Let S be a set, and let $\mathcal{W}$ be a collection of subsets of S. Let P be a real-valued function on $\mathcal{W}$. Then S, $\mathcal{W}$ and P form a* **probability space** *if the following conditions hold:*

- $S \in \mathcal{W}$.
- *$\mathcal{W}$ is closed under complements (if a set is in $\mathcal{W}$, then the set's complement with respect to S is in $\mathcal{W}$ too) and under unions of countably many members of $\mathcal{W}$.*
- $P(A) \geq 0$ *for any A in $\mathcal{W}$.*
- *If $A_1, A_2, ... \in \mathcal{W}$ and the A_i are pairwise disjoint, then*

$$P(\cup_i A_i) = \sum_i P(A_i) \tag{3.148}$$

A **random variable** *is any function $X : S \to \mathcal{R}$.*[6]

Using just these simple axioms, one can prove (with lots of heavy math) theorems like the Strong Law of Large Numbers:

Theorem 9 *Consider a random variable U, and a sequence of independent random variables $U_1, U_2, ...$ which all have the same distribution as U, Then*

$$\lim_{n \to \infty} \frac{U_1 + ... + U_n}{n} = E(U) \text{ with probability 1} \tag{3.149}$$

In other words, the average value of U in all the lines of the notebook will indeed converge to EU.

Exercises

1. Consider a game in which one rolls a single die until one accumulates a total of at least four dots. Let X denote the number of rolls needed. Find $P(X \leq 2)$ and $E(X)$.

2. Recall the committee example in Section 3.7. Suppose now, though, that the selection protocol is that there must be at least one man and at least one woman on the committee. Find $E(D)$ and $Var(D)$.

3. Suppose a bit stream is subject to errors, with each bit having probability p of error, and with the bits being independent. Consider a set of four particular bits. Let X denote the number of erroneous bits among those four.

[6] The function must also have a property called **measurability**, which we will not discuss here.

(a) Find P(X = 2) and EX.

(b) What famous parametric family of distributions does the distribution of X belong to?

(c) Let Y denote the maximum number of consecutive erroneous bits. Find P(Y = 2) and Var(Y).

4. Derive (3.83).

5. Finish the computation in (3.89).

6. Derive the facts that for a Poisson-distributed random variable X with parameter λ, $EX = Var(X) = \lambda$. Use the hints in Section 3.13.8.

7. A civil engineer is collecting data on a certain road. She needs to have data on 25 trucks, and 10 percent of the vehicles on that road are trucks. State the famous parametric family that is relevant here, and find the probability that she will need to wait for more than 200 vehicles to pass before she gets the needed data.

8. In the ALOHA example:

(a) Find $E(X_1)$ and $Var(X_1)$, for the case p = 0.4, q = 0.8. You are welcome to use quantities already computed in the text, e.g. $P(X_1 = 1) = 0.48$, but be sure to cite equation numbers.

(b) Find P(collision during epoch 1) for general p, q.

9. Our experiment is to toss a nickel until we get a head, taking X rolls, and then toss a dime until we get a head, taking Y tosses. Find:

(a) Var(X+Y).

(b) Long-run average in a "notebook" column labeled X^2.

10. Consider the game in Section 3.14.1. Find $E(Z)$ and $Var(Z)$, where $Z = Y_6 - X_6$.

11. Say we choose six cards from a standard deck, one at a time WITHOUT replacement. Let N be the number of kings we get. Does N have a binomial distribution? Choose one: (i) Yes. (ii) No, since trials are not independent. (iii) No, since the probability of success is not constant from trial to trial. (iv) No, since the number of trials is not fixed. (v) (ii) and (iii). (iv) (ii) and (iv). (vii) (iii) and (iv).

12. Suppose we have n independent trials, with the probability of success on the ith trial being p_i. Let X = the number of successes. Use the fact that "the variance of the sum is the sum of the variance" for independent random variables to derive $Var(X)$.

13. Prove Equation (3.29).

14. Show that if X is a nonnegative-integer valued random variable, then

$$EX = \sum_{i=1}^{\infty} P(X \geq i) \tag{3.150}$$

Hint: Write $i = \sum_{j=1}^{i} 1$, and when you see an iterated sum, reverse the order of summation.

15. Suppose we toss a fair time n times, resulting in X heads. Show that the term *expected value* is a misnomer, by showing that

$$\lim_{n \to \infty} P(X = n/2) = 0 \tag{3.151}$$

Use Stirling's approximation,

$$k! \approx \sqrt{2\pi k} \left(\frac{k}{e} \right)^k \tag{3.152}$$

16. Suppose X and Y are independent random variables with standard deviations 3 and 4, respectively.

(a) Find Var(X+Y).

(b) Find Var(2X+Y).

17. Fill in the blanks in the following simulation, which finds the approximate variance of N, the number of rolls of a die needed to get the face having just one dot.

```
onesixth <- 1/6
sumn <- 0
sumn2 <- 0
for (i in 1:10000) {
   n <- 0
   while(TRUE) {
      _____________________________________
      if (_____________________________ < onesixth) break
   }
   sumn <- sumn + n
   sumn2 <- sumn2 + n^2
}
approxvarn <- _____________________________________________
cat("the approx. value of Var(N) is ",approx,"\n")
```

18. Let X be the total number of dots we get if we roll three dice. Find an upper bound for $P(X \geq 15)$, using our course materials.

19. Suppose X and Y are independent random variables, and let Z = XY. Show that $Var(Z) = E(X^2)E(Y^2) - [E(X)]^2[E(Y)]^2$.

20. This problem involves a very simple model of the Web. (Far more complex ones exist.)

Suppose we have n Web sites. For each pair of sites i and j, $i \neq j$, there is a link from i to j with probability p, and no link (in that direction) with probability 1-p. Let N_i denote the number of sites that site i is linked to; note that N_i can range from 0 to n-1. Also, let M_{ij} denote the number of outgoing links that i and j have in common, not counting the one between them, if any. Assume that each site forms its outgoing links independently of the others.

Say n = 10, p = 0.2. Find the followinga:

(a) $P(N_1 = 3)$

(b) $P(N_1 = 3 \text{ and } N_2 = 2)$

(c) $Var(N_1)$

(d) $Var(N_1 + N_2)$

(e) $P(M_{12} = 4)$

Note: There are some good shortcuts in some of these problems, making the work much easier. But you must JUSTIFY your work.

21. Let X denote the number of heads we get by tossing a coin 50 times. Consider Chebychev's Inequality for the case of 2 standard deviations. Compare the upper bound given by the inequality to the exact probability.

22. Suppose the number N of cars arriving during a given time period at a toll booth has a Poisson distribution with parameter λ. Each car has a probability p of being in a car pool. Let M be the number of car-pool cars that arrive in the given period. Show that M also has a Poisson distribution, with parameter $p\lambda$. (Hint: Use the Maclaurin series for e^x.)

23. Consider a three-sided die, as on page 29.

(a) (10) State the value of $p_X(2)$.

(b) (10) Find EX and Var(X).

(c) (15) Suppose you win $2 for each dot. Find EW, where W is the amount you win.

24. Consider the parking space problem in Section 3.13.3. Find Var(M), where M is the number of empty spaces in the first block, and Var(D).

25. Suppose X and Y are independent, with variances 1 and 2, respectively. Find the value of c that minimizes Var[cX + (1-c)Y].

26. In the cards example in Section 2.13.1, let H denote the number of hearts. Find EH and Var(H).

27. In the bank example in Section 3.13.8, suppose you observe the bank for n days. Let X denote the number of days in which at least 2 customers entered during the 11:00-11:15 observation period. Find P(X = k).

28. Find $E(X^3)$, where X has a geometric distribution with parameter p.

Chapter 4

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*. It is for such random variables that the calculus prerequisite for this book is needed.

4.1 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,

$$P(u \leq D \leq v) = v - u \tag{4.1}$$

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

The first crucial point to note is that

$$P(D = c) = 0 \tag{4.2}$$

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

$$P(D = 0.3) \leq P(0.29 \leq D \leq 0.31) = 0.02 \tag{4.3}$$

  the last equality coming from (4.1).

  So, $P(D = 0.3) \leq 0.02$. But we can replace 0.29 and 0.31 in (4.3) 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.

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. So (4.2) 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.

All this may still sound odd to you, but remember, this is an idealization. D actually cannot be just any old point in (0,1). Our dart has nonzero thickness, our measuring instrument has only finite precision, and so on. So it really is an idealization, though an extremely useful one. It's like the assumption of "massless string" in physics analyses; there is no such thing, but it's a good approximation to reality.

4.2 But This Presents a Problem

But Equation (4.2) presents a problem for us in defining the term **distribution** for variables like this. In Section 3.12, we defined this for a discrete random variable Y as a list of the values Y takes on, together with their probabilities. But that would be impossible here—all the probabilities of individual values here are 0.

Instead, we define the distribution of a random variable W which puts 0 probability on individual points in another way. To set this up, we first must define a key function:

Definition 10 *For any random variable W (including discrete ones), its* **cumulative distribution function** *(cdf),* F_W*, is defined by*

$$F_W(t) = P(W \leq t), -\infty < t < \infty \tag{4.4}$$

(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.

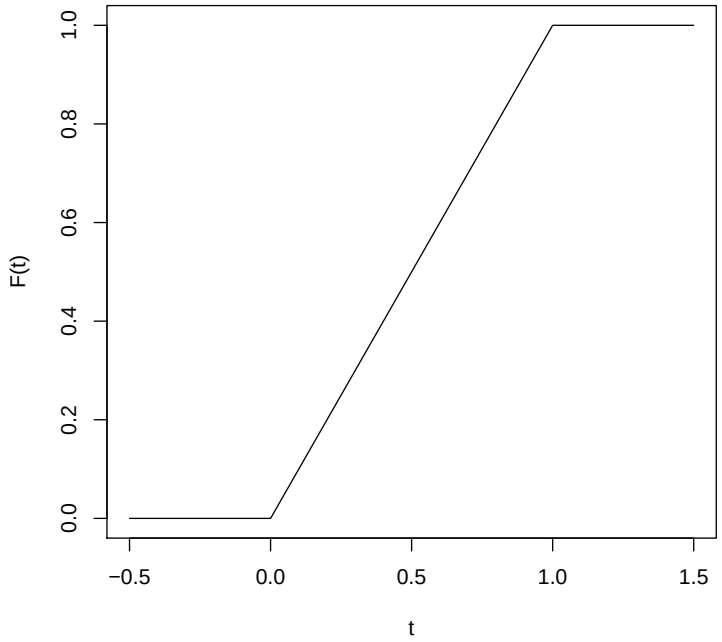
For an example of a cdf, consider our "random dart" example above. We know that, for example

$$F_D(0.23) = P(D \leq 0.23) = 0.23 \tag{4.5}$$

In general for our dart,

$$F_D(t) = \begin{cases} 0, & \text{if } t \leq 0 \\ t, & \text{if } 0 < t < 1 \\ 1, & \text{if } t \geq 1 \end{cases} \tag{4.6}$$

Here is the graph of F_D:

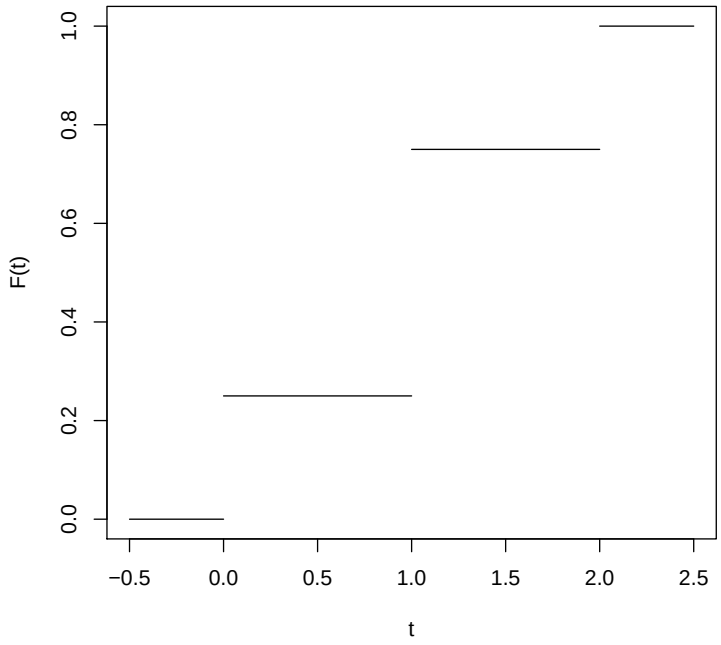


The cdf of a discrete random variable is defined as in Equation (4.4) too. For example, say Z is the number

of heads we get from two tosses of a coin. Then

$$F_Z(t) = \begin{cases} 0, & \text{if } t < 0 \\ 0.25, & \text{if } 0 \le t < 1 \\ 0.75, & \text{if } 1 \le t < 2 \\ 1, & \text{if } t \ge 2 \end{cases} \tag{4.7}$$

For instance, $F_Z(1.2) = P(Z \le 1.2) = P(Z = 0 \text{ or } Z = 1) = 0.25 + 0.50 = 0.75$. (Make sure you confirm this!) 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 (4.6) is continuous while F_Z in (4.7) has jumps. For this reason, we call random variables like D—ones which have 0 probability for individual points—**continuous random variables**.

At this level of study of probability, most random variables are either discrete or continuous, but some are not.

4.3 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.

4.3.1 Motivation, Definition and Interpretation

OK, now we have a name for random variables that have probability 0 for individual points—"continuous"—and we have solved the problem of how to describe their distribution. Now we need something which will be continuous random variables' analog of a probability mass function. (The reader may wish to review pmfs in Section 3.12.)

Think as follows. From (4.4) we can see that for a discrete random variable, its cdf can be calculated by summing is 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 11 *(Oversimplified from a theoretical math point of view.) Consider a continuous random variable W. Define*

$$f_W(t) = \frac{d}{dt} F_W(t), -\infty < t < \infty \tag{4.8}$$

wherever the derivative exists. The function f_W is called 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.)

Recall from calculus that an integral is the area under the curve, derived as the limit of the sums of areas of rectangles drawn at the curve, as the rectangles become narrower and narrower. Since the integral is a limit of sums, its symbol $\int$ is shaped like an S.

Now look at Figure 4.1, depicting a density function f_X. (It so happens that in this example, the density is an increasing function, but most are not.) A rectangle is drawn, positioned horizontally at 1.3 ± 0.1, and with height equal $f_X(1.3)$. The area of the rectangle approximates the area under the curve in that region, which in turn is a probability:

$$
\begin{aligned}
2(0.1) f_X(1.3) &\approx \int_{1.2}^{1.4} f_X(t)\, dt \quad \text{(rect. approx. to slice of area)} && (4.9)\\
&= F_X(1.4) - F_X(1.2) \quad (f_X = F_X') && (4.10)\\
&= P(1.2 < X \leq 1.4) \quad \text{(def. of } F_X) && (4.11)\\
&= P(1.2 < X < 1.4) \quad \text{(prob. of single pt. is 0)} && (4.12)
\end{aligned}
$$

In other words, for any density f_X at any point t, and for small values of c,

$$2c f_X(t) \approx P(t - c < X < t + c) \tag{4.13}$$

Thus we have:

Interpretation of Density Functions

For any density f_X and any two points r and s,

$$\frac{P(r - c < X < r + c)}{P(s - c < X < s + c)} \approx \frac{f_X(r)}{f_X(s)} \tag{4.14}$$

So, X will take on values in regions in which f_X is large much more often than in regions where it is small, with the ratio of frequencies being proportion to the values of f_X.

For our dart random variable D, $f_D(t) = 1$ for t in (0,1), and it's 0 elsewhere.[1] Again, $f_D(t)$ is NOT P(D = t), since the latter value is 0, but it is still viewable as a "relative likelihood." The fact that $f_D(t) = 1$ for all t in (0,1) can be interpreted as meaning that all the points in (0,1) are equally likely to be hit by the dart. More precisely put, you can view the constant nature of this density as meaning that all subintervals of the same length within (0,1) have the same probability of being hit.

Note too that if, say, X has the density in the previous paragraph, then $f_X(3) = 6/15 = 0.4$ and thus $P(1.99 < X < 2.01) \approx 0.008$. Using our notebook viewpoint, think of many repetitions of the experiment, with each line in the notebook recording the value of X in that repetition. Then in the long run, about 0.8% of the lines would have X in (1.99,2.01).

The interpretation of the density is, as seen above, via the relative heights of the curve at various points. The absolute heights are not important. Think of what happens when you view a histogram of grades on an exam. Here too you are just interested in relative heights. (In a later unit, you will see that a histogram is actually an estimate for a density.)

[1] The derivative does not exist at the points 0 and 1, but that doesn't matter.

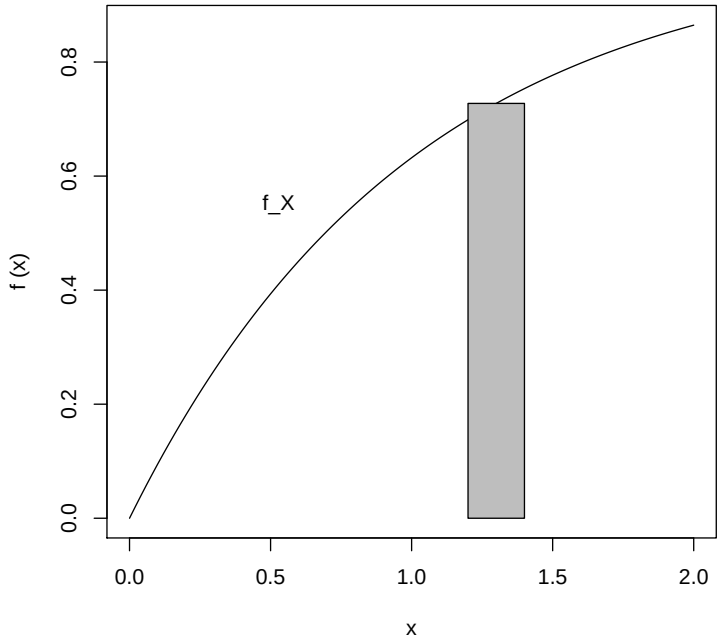


Figure 4.1: Approximation of Probability by a Rectangle

4.3.2 Properties of Densities

Equation (4.8) implies

Property A:

$$P(a < W \leq b) = F_W(b) - F_W(a) = \int_a^b f_W(t)\, dt \tag{4.15}$$

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

Property B:

$$P(a < W \leq b) = P(a \leq W \leq b) = P(a \leq W < b) = P(a < W < b) = \int_a^b f_W(t)\, dt \tag{4.16}$$

This in turn implies:

Property C:

$$\int_{-\infty}^{\infty} f_W(t)\, dt = 1 \tag{4.17}$$

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 $t > 1$.

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

$$E(W) = \sum_c c p_W(c) \tag{4.18}$$

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:

$$E(W) = \int_t t f_W(t)\, dtE \tag{4.19}$$

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

$$E(W) = \int_{-\infty}^{\infty} f_W(t)\ dtE \tag{4.20}$$

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

And of course,

$$E(W^2) = \int_t t^2 f_W(t)\ dt \tag{4.21}$$

and in general, similarly to (3.24):

Property E:

$$E[g(W)] = \int_t g(t) f_W(t)\ dt \tag{4.22}$$

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

Property F:

Equations (3.13), (3.14), (3.16), (3.29), (3.32), (3.39) still hold in the continuous case.

4.3.3 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:

$$EX = \int_1^4 t \cdot 2t/15\ dt = 2.8 \tag{4.23}$$

$$P(X > 2.5) = \int_{2.5}^4 2t/15\ dt = 0.65 \tag{4.24}$$

$$F_X(s) = \int_1^s 2t/15\ dt = \frac{s^2 - 1}{15} \quad \text{for s in (1,4) (cdf is 0 for t} < 1\text{, and 1 for t} > 4) \tag{4.25}$$

$$\begin{aligned} Var(X) &= E(X^2) - (EX)^2 \quad \text{(from (3.29)} && (4.26)\\ &= \int_1^4 t^2 2t/15 \, dt - 2.8^2 \quad \text{(from (4.23))} && (4.27)\\ &= 5.7 && (4.28) \end{aligned}$$

$$\begin{aligned} P(\text{tenths digit of X is even}) &= \sum_{i=0}^{28} P\left[1 + i/10 < X < 1 + (i+1)/10\right] && (4.29)\\ &= \sum_{i=0}^{28} \int_{1+i/10}^{1+(i+1)/10} 2t/15 \, dt && (4.30)\\ &= \text{...(integration left to the reader)} && (4.31) \end{aligned}$$

4.4 Famous Parametric Families of Continuous Distributions

4.4.1 The Uniform Distributions

4.4.1.1 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 (4.17). So, that constant must be 1 divided by the length of the interval:

$$f_D(t) = \frac{1}{r-q} \tag{4.32}$$

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).

4.4.2 R Functions

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

- **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

4.4.2.1 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 at 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**.

4.4.2.2 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.

4.4.3 The Normal (Gaussian) Family of Continuous Distributions

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

[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] 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.

4.4.3.1 Density and Properties

Density and Parameters:

The density for a normal distribution is

$$f_W(t) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-0.5\left(\frac{t-\mu}{\sigma}\right)^2}, -\infty < t < \infty \tag{4.33}$$

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).

Closure Under Affine Transformation:

The family is closed under affine transformations, meaning that if X has the distribution $N(\mu, \sigma^2)$, then Y = cX + d has the distribution $N(c\mu + d, c^2\sigma^2)$, i.e. Y too has a normal distribution.

Consider this 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 (3.39) *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 (4.33), now with the new mean and variance.

Let's derive this. For convenience, suppose $c > 0$. Then

$$\begin{aligned}
F_Y(t) &= P(Y \leq t) \text{ (definition of } F_Y) && (4.34)\\
&= P(cX + d \leq t) \text{ (definition of Y)} && (4.35)\\
&= P\left(X \leq \frac{t-d}{c}\right) \text{ (algebra)} && (4.36)\\
&= F_X\left(\frac{t-d}{c}\right) \text{ (definition of } F_X) && (4.37)
\end{aligned}$$

Therefore

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

$$
\begin{aligned}
f_Y(t) &= \frac{d}{dt} F_Y(t) \text{ (definition of } f_Y\text{)} && (4.38)\\
&= \frac{d}{dt} F_X\left(\frac{t-d}{c}\right) \text{ (from (4.37))} && (4.39)\\
&= f_X\left(\frac{t-d}{c}\right) \cdot \frac{d}{dt}\frac{t-d}{c} \text{ (definition of } f_X \text{ and the Chain Rule)} && (4.40)\\
&= \frac{1}{c} \cdot \frac{1}{\sqrt{2\pi}\sigma}\, e^{-0.5\left(\frac{\frac{t-d}{c}-\mu}{\sigma}\right)^2} \text{ (from (4.33)} && (4.41)\\
&= \frac{1}{\sqrt{2\pi}(c\sigma)}\, e^{-0.5\left(\frac{t-(c\mu+d)}{c\sigma}\right)^2} \text{ (algebra)} && (4.42)
\end{aligned}
$$

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

Evaluating the Normal cdf

The function in (4.33) 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). This one table is sufficient for the entire normal family, because if X has the distribution $N(\mu, \sigma^2)$ then

$$\frac{X-\mu}{\sigma} \tag{4.43}$$

has a N(0,1) distribution too, due to the affine transformation closure property discussed above.

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

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 signature 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.

We'll use both methods in our first couple of examples below.

4.4.3.2 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 of disk sectors she reads or writes X has a normal distribution has a mean of 500 and a standard deviation of 15. 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

$$P(X \geq 535) = P\left(Z \geq \frac{535 - 500}{15}\right) = 1 - \Phi(35/15) = 0.01 \tag{4.44}$$

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

4.4.3.3 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 deman. Then the key point is that we are given that $N \geq 25$, so

$$
\begin{aligned}
P(N \geq 30 | N \geq 25) &= \frac{P(N \geq 30 \text{ and } N \geq 25)}{P(N \geq 25)} \quad ((2.5)) && (4.45)\\
&= \frac{P(N \geq 30}{P(N \geq 25)} && (4.46)\\
&= \frac{\Phi\left[(30 - 28.8)/3.1\right]}{\Phi\left[(25 - 28.8)/3.1\right]} && (4.47)\\
&= 0.39 && (4.48)
\end{aligned}
$$

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 4.4.3.6.

4.4.3.4 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,[5] 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:[6]

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.

4.4.3.5 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.

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

[6] A more mathematically precise statement of the theorem is given in Section 4.4.3.8.

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

```
> pnorm(106,20*5.2,sqrt(20*5.2))
[1] 0.5777404
```

4.4.3.6 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 3.6):

$$T = T_1 + ... + T_n \tag{4.49}$$

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

$$P(X > 12) = P(Z > \frac{12 - 10}{\sqrt{5}}) \approx 1 - \Phi(0.894) = 0.186 \tag{4.50}$$

Or:

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

The exact answer is 0.132. Remember, the reason we could do this 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 (4.50), we might write

$$P(X > 12) = P(X \geq 13) = P(Z > \frac{13 - 10}{\sqrt{5}}) \approx 1 - \Phi(1.342) = 0.090 \tag{4.51}$$

That value of 0.090 is considerably smaller than the 0.186 we got from (4.50). We could "split the difference" this way:

$$P(X > 12) = P(X \geq 12.5) = P(Z > \frac{12.5 - 10}{\sqrt{5}}) \approx 1 - \Phi(1.118) = 0.132 \tag{4.52}$$

(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.

4.4.3.7 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.

4.4.3.8 Optional topic: Formal Statement of the CLT

Definition 13 *A sequence of random variables $L_1, L_2, L_3, ...$* **converges in distribution** *to a random variable M if*

$$\lim_{n \to \infty} P(L_n \leq t) = P(M \leq t), \textit{ for all } t \tag{4.53}$$

Note by the way, that these random variables need not be defined on the same probability space.

The formal statement of the CLT is:

Theorem 14 *Suppose $X_1, X_2, ...$ are independent random variables, all having the same distribution which has mean m and variance v^2. Then*

$$Z = \frac{X_1 + ... + X_n - nm}{v\sqrt{n}} \tag{4.54}$$

converges in distribution to a N(0,1) random variable.

4.4.3.9 Importance in Modeling

Needless to say, there are no random variable 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, 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 distributions. This should come as no surprise, knowing the CLT. In addition, the CLT tells us in many of these cases the quantities used for statistical estimation are approximately normal, even if the data they are calculated from do not.

4.4.4 The Chi-Square Family of Distributions

4.4.4.1 Density and Properties

Let $Z_1, Z_2, ..., Z_k$ be independent N(0,1) random variables. Then the distribution of

$$Y = Z_1^2 + ... + Z_k^2 \tag{4.55}$$

is called **chi-square with k degrees of freedom**. We write such a distribution as χ_k^2. Chi-square is a one-parameter family of distributions, and arises quite frequently in statistical applications, as will be seen in future chapters.

It turns out that chi-square is a special case of the gamma family in Section 4.4.7 below, with r = k/2 and $\lambda = 0.5$.

4.4.4.2 Importance in Modeling

This distribution is used widely in statistical applications. As will be seen in our chapters on statistics, many statistical methods involve a sum of squared normal random variables.[7]

4.4.5 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.

4.4.5.1 Density and Properties

The densities in this family have the form

$$f_W(t) = \lambda e^{-\lambda t}, 0 < t < \infty \tag{4.56}$$

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 (4.56)), since the latter is the mean. But this is actually quite natural, for the reason cited in the following subsection.

4.4.6 R Functions

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

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

[7] The motivation for the term *degrees of freedom* will be explained in those chapters too.

4.4.6.1 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

$$T_r = X_1 + ... + X_r \tag{4.57}$$

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 λt. 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 15 *Suppose $X_1, X_2, ...$ are i.i.d. nonnegative continuous random variables. Define*

$$T_r = X_1 + ... + X_r \tag{4.58}$$

and

$$N(t) = \max\{k : T_k \leq t\} \tag{4.59}$$

Then the distribution of N(t) is Poisson with parameter λt for all t if and only the X_i have an exponential distribution with parameter λ.

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 has lasts longer than t, then the count of burnouts at time t is 0, and vice versa. Then

$$\begin{aligned} P(X_1 > t) &= P[N(t) = 0] \text{ (see above equiv.)} && (4.60)\\ &= \frac{(\lambda t)^0}{0!} \cdot e^{-\lambda t} \text{ (exp. hyp.)} && (4.61)\\ &= e^{-\lambda t} \text{ (3.99)} && (4.62) \end{aligned}$$

Then

$$f_{X_1}(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = \lambda e^{-\lambda t} \tag{4.63}$$

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 λ, then for u nonnegative and each positive integer k,

$$P[N(u) = k] = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{4.64}$$

The proof for the case k = 0 just reverses (4.60) 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 be integrating (4.65) from u to infinity. One needs to perform k-1 integrations by parts, and eventually one arrives at (4.64), summed from 1 to k, as required.

■

The collection of random variables N(t) $t \geq 0$, is called a **Poisson process**.

The relation $E[N(t)] = \lambda t$ says that replacements are occurring at an average rate of λ per unit time. Thus λ is called the **intensity parameter** of the process. It is because of this "rate" interpretation that makes λ a natural indexing parameter in (4.56).

4.4.6.2 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.

Exponential distributions are the only continous ones that are "memoryless." This point is pursued in Chapter 6. Due to this property, exponential distributions play a central role in Markov chains (Chapter 11).

4.4.7 The Gamma Family of Distributions

4.4.7.1 Density and Properties

Recall Equation (4.57), in which the random variable T_r was defined to be 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, with density

$$f_{T_r}(t) = \frac{1}{(r-1)!}\lambda^r t^{r-1} e^{-\lambda t},\ t > 0 \tag{4.65}$$

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 (4.65).

then

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

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx \tag{4.66}$$

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

$$f_W(t) = \frac{1}{\Gamma(r)}\lambda^r t^{r-1} e^{-\lambda t},\ t > 0 \tag{4.67}$$

(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 (4.57) 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.

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 4.2 that even with r = 10 it is rather close to normal.

It also turns out that the chi-square distribution with d degrees of freedom is a gamma distribution, with r = d/2 and $\lambda = 0.5$.

4.4.7.2 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

$$P(Y > 552) = \int_{552}^{\infty} \frac{1}{4!} 0.01^5 t^4 e^{-0.01t}\, dt \tag{4.68}$$

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
```

4.4.7.3 Importance in Modeling

As seen in (4.57), 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 4.2.

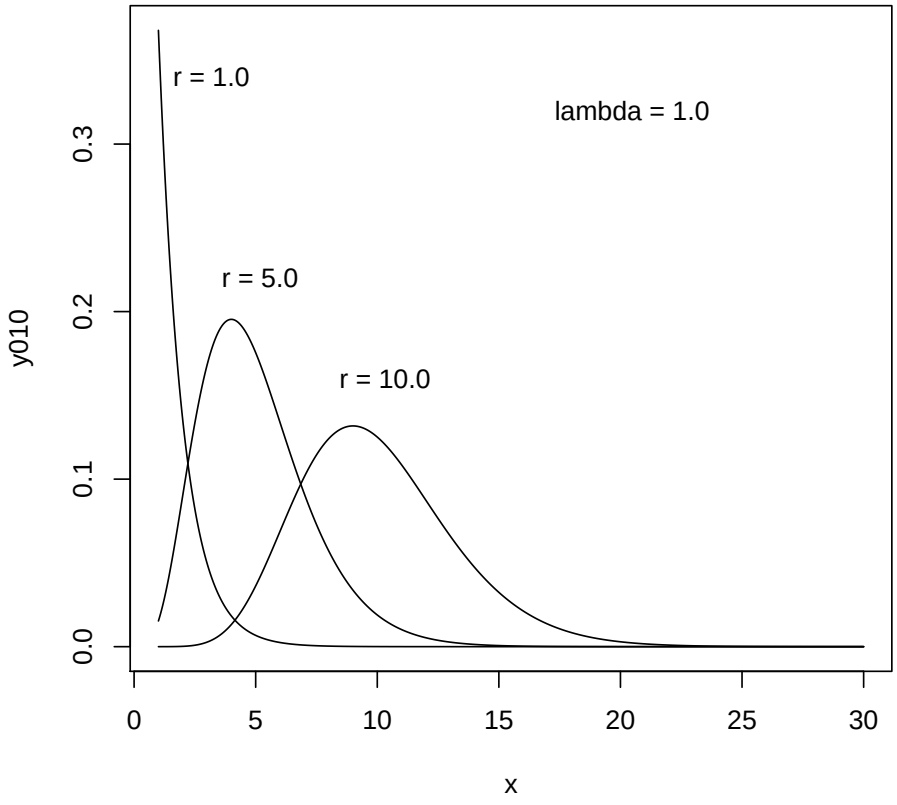


Figure 4.2: Various Gamma Densities

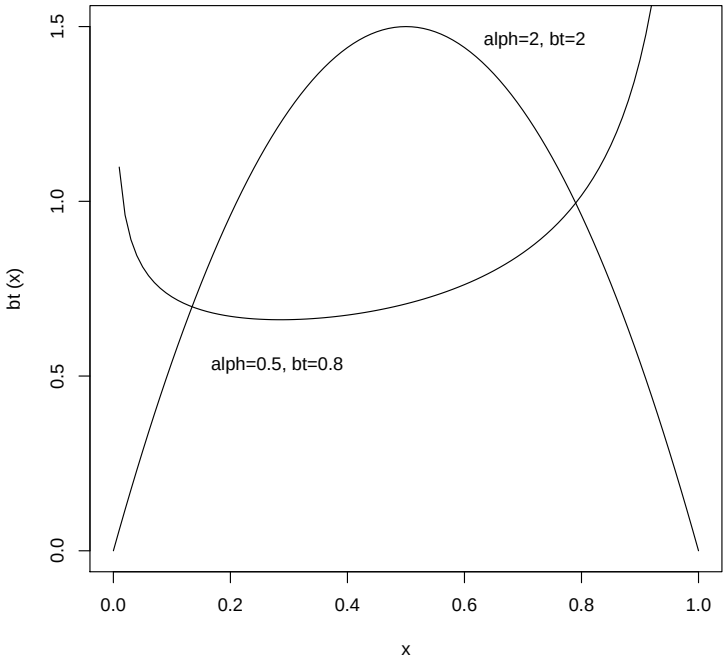
4.4.8 The Beta Family of Distributions

As seen in Figure 4.2, 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)? 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.

The densities of the family have the following form:

$$\frac{\Gamma(\alpha+\lambda)}{\Gamma(\alpha)\Gamma(\lambda)}(1-t)^{\alpha-1}t^{\beta-1} \tag{4.69}$$

There are two parameters, α and β. Here are two possibilities.



The mean and variance are

$$\frac{\alpha}{\alpha+\beta} \tag{4.70}$$

and

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4.71}$$

4.5 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 4.2, 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 to 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 9.

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**.

4.6 A General Method for Simulating a Random Variable

Suppose we wish to simulate a random variable X with cdf 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 cdf of X. Here "inverse" is in the sense that, for instance, squaring and "square-rooting," exp() and ln(), etc. are inverse operations of each other.

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 X in R as **sqrt(runif(1))**. Here's why:

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

$$\begin{aligned} P[G(U) \leq t] & \\ &= P[U \leq G^{-1}(t)] \\ &= P[U \leq H(t)] \\ &= H(t) \end{aligned} \tag{4.72}$$

In other words, the cdf of G(U) is F_X! So, G(U) has the same distribution as X.

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

4.7 "Hybrid" Continuous/Discrete Distributions

A random variable could have a distribution that it partly discrete and partly continuous. Recall our first example, from Section 4.1, in which D is the position that a dart hits when thrown at the interval (0,1). Suppose our measuring instrument is broken, and registers any value of D past 0.8 as being equal to 0.8. Let W denote the actual value recorded by this instrument.

Then P(W = 0.8) = 0.2, so W is not a continuous random variable, in which every point has mass 0. On the other hand, P(W = t) = 0 for every t before 0.8, so W is not discrete either.

In the advanced theory of probability, some very odd mixtures, beyond this simple discrete/continuous example, can occur, though primarily of theoretical interest.

Exercises

1. Fill in the blanks, in the following statements about continuous random variables. Make sure to use our book's notation.

(a) $\frac{d}{dt}P(X \leq t) =$ _____

(b) $P(a < X < b) =$ _____ $-$ _____

2. Suppose X has a uniform distribution on (-1,1), and let $Y = X^2$. Find f_Y.

3. In the network intrusion example in Section 4.4.3.2, suppose X is not normally distributed, but instead has a uniform distribution on (450,550). Find $P(X \geq 535)$ in this case.

4. Suppose X has an exponential distribution with parameter λ. Show that $EX = 1/\lambda$ and $Var(X) = 1/\lambda^2$.

5. Suppose $f_X(t) = 3t^2$ for t in (0,1) and is zero elsewhere. Find $F_X(0.5)$ and $E(X)$.

6. Suppose light bulb lifetimes X are exponentially distributed with mean 100 hours.

(a) Find the probability that a light bulb burns out before 25.8 hours.

In the remaining parts, suppose we have two light bulbs. We install the first at time 0, and then when it burns out, immediately replace it with the second.

(b) Find the probability that the first light bulb lasts less than 25.8 hours and the lifetime of the second is more than 120 hours.

(c) Find the probability that the second burnout occurs after time 192.5.

7. Suppose for some continuous random variable X, $f_X(t)$ is equal to 2(1-t) for t in (0,1) and is 0 elsewhere.

(a) Why is the constant here 2? Why not, say, 168?

(b) Find $F_X(0.2)$ and Var(X).

(c) Using the method in Section 4.6, write an R function, named **oneminust()**, that generates a random variate sampled from this distribution. Then use this function to verify your answers in (b) above.

8. The company Wrong Turn Criminal Mismanagement makes predictions every day. They tend to err on the side of overpredicting, with the error having a uniform distribution on the interval (-0.5,1.5). Find the following:

(a) The mean and variance of the error.

(b) The mean of the absolute error.

(c) The probability that exactly two errors are greater than 0.25 in absolute value, out of 10 predictions. Assume predictions are independent.

9. 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.

(a) Suppose we compute just one square root. Find the probability that it is in error by more than 0.2.

(b) Suppose we compute a sum of 50 square roots. find the approximate probability that the sum is in error by more than 2.0.

(c) Find a number c such that the probability is approximately 95% that the sum is in error by no more than c. Again assume we have a sum of 50 square roots.

10. "All that glitters is not gold," and not every bell-shaped density is normal. The family of Cauchy distributions, having density

$$f_X(t) = \frac{1}{\pi c} \frac{1}{1 + (\frac{t-b}{c})^2}, \ \infty < t < \infty \tag{4.73}$$

is bell-shaped but definitely not normal.

Here the parameters b and c correspond to mean and standard deviation in the normal case, but actual neither the mean nor standard deviation exist for Cauchy distributions. The mean's failure to exist is due to technical problems involving the theoretical definition of integration. In the case of variance, it does not exist because there is no mean, but even more significantly, $E[(X-b)^2] = \infty$.

However, a Cauchy distribution does have a median, b, so we'll use that instead of a mean. Also, instead of a standard deviation, we'll use as our measure of dispersion the interquartile range, defined (for any distribution) to be the difference between the 75th and 25th percentiles.

We will be investigating the Cauchy distribution that has b = 0 and c = 1.

(a) Find the interquartile range of this Cauchy distribution.

(b) Find the normal distribution that has the same median and interquartile range as this Cauchy distribution.

(c) Use R to plot the densities of the two distributions on the same graph, so that we can see that they are both bell-shaped, but different.

11. Consider the following game. A dart will hit the random point Y in (0,1) according to the density $f_Y(t) = 2t$. You must guess the value of Y. (Your guess is a constant, not random.) You will lose \$2 per unit error if Y is to the left of your guess, and will lose \$1 per unit error on the right. Find best guess in terms of expected loss.

12. Consider a machine that places a pin in the middle of a flat, disk-shaped object. The placement is subject to error. Let X and Y be the placement errors in the horizontal and vertical directions, respectively, and let W denote the distance from the true center to the pin placement. Suppose X and Y are independent and have normal distributions with mean 0 and variance 0.04. Find $P(W > 0.6)$.

Hint: $P(W > 0.7) = P(W^2 > 0.49)$. Find the distribution of cW^2 for a suitably chosen constant c.

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

14. A certain public parking garage charges parking fees of \$1.50 for the first hour or fraction thereof, and \$1 per hour after that. So, someone who stays 57 minutes pays \$1.50, someone who parks for one hour and 12 minutes pays \$1.70, and so on. Suppose parking times T are exponentially distributed with mean 1.5 hours. Let W denote the total fee paid. Find E(W) and Var(W).

15. Fill in the blank: Density functions for continuous random variables are analogs of the _____ functions that are used for discrete random variables.

16. Suppose for some random variable W, $F_W(t) = t^3$ for $0 < t < 1$, with $F_W(t)$ being 0 and 1 for $t < 0$ and $t > 0$, respectively. Find $f_W(t)$ for $0 < t < 1$.

17. Suppose X has a binomial distribution with parameters n and p. Then X is approximately normally distributed with mean np and variance np(1-p). For each of the following, answer either A or E, for "approximately" or "exact," respectively:

(a) the distribution of X is normal

(b) E(X) is np

(c) Var(X) is np(1-p)

18. Consider the density $f_Z(t) = 2t/15$ for $1 < t < 4$ and 0 elsewhere. Find the median of Z, as well as Z's third moment, $E(Z^3)$, and its third central moment, $E[(Z - EZ)^3]$.

19. Suppose X has a uniform distribution on the interval (20,40), and we know that X is greater than 25. What is the probability that X is greater than 32?

20. Suppose U and V have the 2t/15 density on (1,4). Let N denote the number of values among U and V that are greater than 1.5, so N is either 0, 1 or 2. Find Var(N).

21. Find the value of $E(X^4)$ if X has an N(0,1) distribution. (Give your answer as a number, not an integral.)

Chapter 5

Multivariate Probability Models

Most applications of probability and statistics involve the interaction between variables. For instance, when you buy a book at Amazon.com, the software will likely inform you of other books that people bought in conjunction with the one you selected. Amazon is relying on the fact that sales of certain pairs or groups of books are correlated.

Thus we need the notion of distributions that describe how two or more variables vary together. This chapter develops that notion.

5.1 Multivariate Distributions

Individual pmfs p_X and densities f_X don't describe these correlations. We need something more. We need ways to describe multivariate distributions.

5.1.1 Discrete Case

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 of discrete random variables U and V, as follows.

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

$$p_{Y,B}(i,j) = P(Y = i \text{ and } B = j) = \frac{\binom{2}{i}\binom{3}{j}\binom{4}{4-i-j}}{\binom{9}{4}} \tag{5.1}$$

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

i ↓, j →	0	1	2	3
0	0.002	0.024	0.036	0.008
1	0.162	0.073	0.048	0.004
2	0.012	0.024	0.006	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 all values that X takes on. We do the same thing for a pair of random variables:

Definition 16 *For discrete random variables U and V, their probability mass function is defined to be*

$$p_{U,V}(i,j) = P(U = i \text{ and } V = j) \tag{5.2}$$

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

$$p_{U,V,W}(i,j,k) = P(U = i \text{ and } V = j \text{ and } W = k) \tag{5.3}$$

So in our marble example above, $p_{Y,B}(1,2) = 0.048$, $p_{Y,B}(2,0) = 0.012$ and so on.

Just as in the case of a single discrete random variable X we have

$$P(X \in A) = \sum_{i \in A} p_X(i) \tag{5.4}$$

for any subset A of the range of X, for a discrete pair (U,V) and any subset A of the pair's range, we have

$$P[(U,V) \in A) = \sum_{(i,j) \in A} \sum p_{U,V}(i,j) \tag{5.5}$$

Again, consider our marble example. Suppose we want to find $P(Y < B)$. Doing this "by hand," we would simply sum the relevant probabilities in the table above, which are marked in bold face below:

i ↓, j →	0	1	2	3
0	0.002	**0.024**	**0.036**	**0.008**
1	0.162	0.073	**0.048**	**0.004**
2	0.012	0.024	0.006	**0.000**

The desired probability would then be 0.024+0.036+0.008+0.048+0.004 = 0.12.

Writing it in the more formal way using (5.5), we would set

$$A = \{(i, j) : i < j\} \tag{5.6}$$

and then

$$P(Y < B) = P[(Y, B) \in A) = \sum_{i=0}^{2} \sum_{j=i+1}^{3} p_{Y,B}(i, j) \tag{5.7}$$

Note that the lower bound in the inner sum is j = i+1. This reflects the common-sense point that in the event $Y < B$, B must be at least equal to Y+1.

Of course, this sum still works out to 0.12 as before, but it's important to be able to express this as a double sum of $p_{Y,B}()$, as above. We will rely on this to motivate the continuous case in the next section.

Expected values are calculated in the analogous manner. Recall that for a function g() of X

$$E[g(X] = \sum_i g(i) p_X(i) \tag{5.8}$$

So, for any function g() of two discrete random variables U and V, define

$$E[g(U, V)] = \sum_i \sum_j g(i, j) p_{U,V}(i, j) \tag{5.9}$$

For instance, if for some bizarre reason we wish to find the expected value of the product of the numbers of yellow and blue marbles above,[1], the calculation would be

$$E(YB) = \sum_{i=0}^{2} \sum_{j=0}^{3} ij \; p_{Y,B}(i, j) = 0.255 \tag{5.10}$$

[1]Not so bizarre, we'll find in Section 5.2.1.

The univariate pmfs, called *marginal pmfs*, can of course be recovered from the multivariate pmf:

$$p_U(i) = P(U = i) = \sum_j P(U = i, V = j) = \sum_j p_{U,V}(i, j) \tag{5.11}$$

5.1.2 Multivariate Densities

5.1.2.1 Motivation and Definition

Extending our previous definition of cdf for a single variable, we define the two-dimensional cdf for a pair of random variables X and Y as

$$F_{X,Y}(u, v) = P(X \leq u \text{ and } Y \leq v) \tag{5.12}$$

If X and Y were discrete, we would evaluate that cdf via a double sum of their bivariate pmf. You may have guessed by now that the analog for continuous random variables would be a double integral, and it is. The integrand is the bivariate density:

$$f_{X,Y}(u, v) = \frac{\partial^2}{\partial u \partial v} F_{X,Y}(u, v) \tag{5.13}$$

Densities in higher dimensions are defined similarly.[2]

As in the univariate case, a bivariate density shows which regions of the X-Y plane occur more frequently, and which occur less frequently.

5.1.2.2 Use of Multivariate Densities in Finding Probabilities and Expected Values

Again by analogy, for any region A in the X-Y plane,

$$P[(X, Y) \epsilon A] = \iint_A f_{X,Y}(u, v)\, du\, dv \tag{5.14}$$

So, just as probabilities involving a single variable X are found by integrating f_X over the region in question, for probabilities involving X and Y, we take the double integral of $f_{X,Y}$ over that region.

[2] Just as we noted in Section 4.7 that some random variables are neither discrete nor continuous, there are some pairs of continuous random variables whose cdfs do not have the requisite derivatives. We will not pursue such cases here.

Also, for any function g(X,Y),

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u,v) f_{X,Y}(u,v)\ du\ dv \tag{5.15}$$

where it must be kept in mind that $f_{X,Y}(u,v)$ may be 0 in some regions of the U-V plane. Note that there is no set A here as in (5.14). See (5.19) below for an example.

Finding marginal densities is also analogous to the discrete case, e.g.

$$f_X(s) = \int_t f_{X,Y}(s,t)\ dt \tag{5.16}$$

Other properties and calculations are analogous as well. For instance, the double integral of the density is equal to 1, and so on.

5.1.2.3 Example: a Triangular Distribution

Suppose (X,Y) has the density

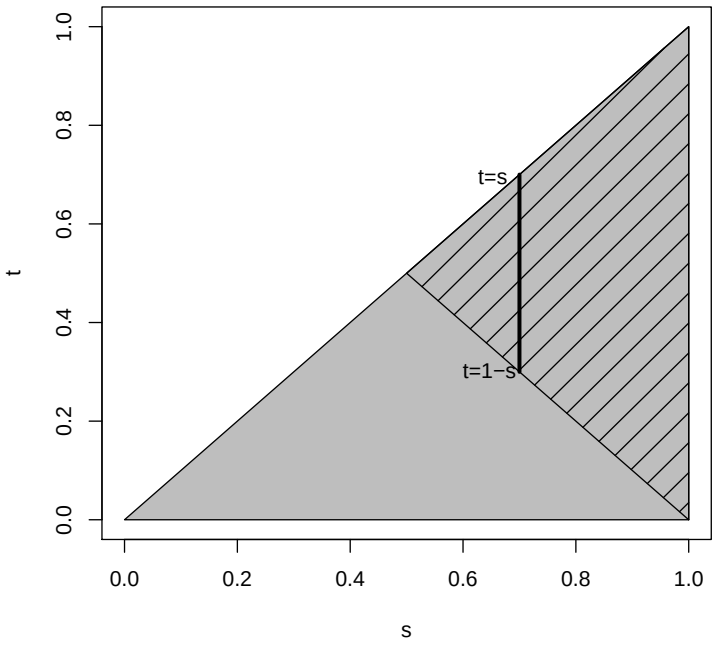
$$f_{X,Y}(s,t) = 8st, 0 < t < s < 1 \tag{5.17}$$

The density is 0 outside the region $0 < t < s < 1$.

First, think about what this means, say in our notebook context. We do the experiment many times. Each line of the notebook records the values of X and Y. Each of these (X,Y) pairs is a point in the triangular region $0 < t < s < 1$. Since the density is highest near the point (1,1) and lowest near (0,1), (X,Y) will be observed near (1,1) much more often than near (0,1), with points near, say, (1,0.5) occurring with middling frequencies.

Let's find $P(X + Y > 1)$. This calculation will involve a double integral. The region A in (5.14) is $\{(s,t) : s + t > 1, 0 < t < s < 1\}$. We have a choice of integrating in the order ds dt or dt ds. The latter will turn out to be more convenient.

To see how the limits in the double integral are obtained, first review (5.7). We use the same reasoning here, changing from sums to integrals and applying the current density, as shown in this figure:



Here s represents X and t represents Y. The gray area is the region in which (X,Y) ranges. The subregion A in (5.14), corresponding to the event X+Y > 1, is shown in the striped area in the figure.

The dark vertical line shows all the points (s,t) in the striped region for a typical value of s in the integration process. Since s is the variable in the outer integral, considered it fixed for the time being and ask where t will range *for that s*. We see that for X = s, Y will range from 1-s to s; thus we set the inner integral's limits to 1-s and s. Finally, we then ask where s can range, and see from the picture that it ranges from 0.5 to 1. Thus those are the limits for the outer integral.

$$P(X+Y>1) = \int_{0.5}^{1} \int_{1-s}^{s} 8st \; dt \; ds = \int_{0.5}^{1} 8s \cdot (s-0.5) \; ds = \frac{5}{6} \tag{5.18}$$

Following (5.15),

$$E[\sqrt{X+Y}] = \int_{0}^{1} \int_{0}^{s} \sqrt{s+t} \; 8st \; dt \; ds \tag{5.19}$$

Let's find the marginal density $f_Y(t)$. So we must "integrate out" the s in (5.17):

$$f_Y(t) = \int_t^1 8st\ ds = 4t - 4t^3 \tag{5.20}$$

for $0 < t < 1$, 0 elsewhere.

5.2 More on Co-variation of Random Variables

5.2.1 Covariance

Definition 17 *The* **covariance** *between random variables X and Y is defined as*

$$Cov(X,Y) = E[(X - EX)(Y - EY)] \tag{5.21}$$

Suppose that typically when X is larger than its mean, Y is also larger than its mean, and vice versa for below-mean values. Then (5.21) will likely be positive. In other words, if X and Y are positively correlated (a term we will define formally later but keep intuitive for now), then their covariance is positive. Similarly, if X is often smaller than its mean whenever Y is larger than its mean, the covariance and correlation between them will be negative. All of this is roughly speaking, of course, since it depends on *how much* and *how often* X is larger or smaller than its mean, etc.

It can be shown that covariance is linear in both arguments:

Property A:

$$Cov(aX + bY, cU + dV) = acCov(X,U) + adCov(X,V) + bcCov(Y,U) + bdCov(Y,V) \tag{5.22}$$

for any constants a, b, c and d.

Also:

Property B:

$$Cov(X, Y + q) = Cov(X,Y) \tag{5.23}$$

for any constant q and so on.

Note:

Property C:

$$Cov(X, X) = Var(X) \tag{5.24}$$

for any X with finite variance.

Also, here is a shortcut way to find the covariance:

Property D:

$$Cov(X, Y) = E(XY) - EX \cdot EY \tag{5.25}$$

The proof will help you review some important issues, namely (a) E(U+V) = EU + EV, (b) E(cU) = c EU and Ec = c for any constant c, and (c) EX and EY are constants in (5.25).

$$\begin{aligned} Cov(X, Y) &= E[(X - EX)(Y - EY)] \text{ (definition)} && (5.26)\\ &= E\left[XY - EX \cdot Y - EY \cdot X + EX \cdot EY\right] \text{ (algebra)} && (5.27)\\ &= E(XY) + E[-EX \cdot Y] + E[-EY \cdot X] + E[EX \cdot EY] \text{ (E[U+V]=EU+EV)} && (5.28)\\ &= E(XY) - EX \cdot EY \text{ (E[cU] = cEU, Ec = c)} && (5.29) \end{aligned}$$

Another important property:

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) \tag{5.30}$$

This comes from (5.25), the relation $Var(X) = E(X^2) - EX^2$ and the corresponding one for Y. Just substitute and do the algebra.

5.2.2 Correlation

Covariance does measure how much or little X and Y vary together, but it is hard to decide whether a given value of covariance is "large" or not. For instance, if we are measuring lengths in feet and change to inches, then (5.22) shows that the covariance will increase by $12^2 = 144$. Thus it makes sense to scale covariance according to the variables' standard deviations. Accordingly, the *correlation* between two random variables

X and Y is defined by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \tag{5.31}$$

So, correlation is unitless, i.e. does not involve units like feet, pounds, etc.

It is shown later in this chapter that

- $-1 \leq \rho(X,Y) \leq 1$
- $|\rho(X,Y)| = 1$ if and only if X and Y are exact linear functions of each other, i.e. Y = cX + d for some constants c and d

5.2.3 Example: Continuation of Section 5.1.2.3

Let's find the correlation between X and Y in the example in Section 5.1.2.3.

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^s st \cdot 8st \; dt \; ds && (5.32)\\ &= \int_0^1 8s^2 \cdot s^3/3 \; ds && (5.33)\\ &= \frac{4}{9} && (5.34) \end{aligned}$$

$$\begin{aligned} f_X(s) &= \int_0^s 8st \; dt && (5.35)\\ &= 4st^2\Big|_0^s && (5.36)\\ &= 4s^3 && (5.37) \end{aligned}$$

$$\begin{aligned} f_Y(t) &= \int_t^1 8st \; ds && (5.38)\\ &= 4t \cdot s^2\Big|_t^1 && (5.39)\\ &= 4t(1-t^2) && (5.40) \end{aligned}$$

$$EX = \int_0^1 s \cdot 4s^3 \, ds = \frac{4}{5} \tag{5.41}$$

$$E(X^2) = \int_0^1 s^2 \cdot 4s^3 \, ds = \frac{2}{3} \tag{5.42}$$

$$Var(X) = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = 0.027 \tag{5.43}$$

$$EY = \int_0^1 t \cdot (4t - 4t^3) \, ds = \frac{4}{3} - \frac{4}{5} = \frac{8}{15} \tag{5.44}$$

$$E(Y^2) = \int_0^1 t^2 \cdot (4t - 4t^3) \, dt = 1 - \frac{4}{6} = \frac{1}{3} \tag{5.45}$$

$$Var(Y) = \frac{1}{3} - \left(\frac{8}{15}\right)^2 = 0.049 \tag{5.46}$$

$$Cov(X, Y) = \frac{4}{9} - \frac{4}{5} \cdot \frac{8}{15} = 0.018 \tag{5.47}$$

$$\rho(X, Y) = \frac{0.018}{\sqrt{0.027 \cdot 0.049}} = 0.49 \tag{5.48}$$

5.2.4 Example: a Catchup Game

Consider the following simple game. There are two players, who take turns playing. One's position after k turns is the sum of one's winnings in those turns. Basically, a turn consists of generating a random U(0,1) variable, with one difference—if that player is currently losing, he gets a bonus of 0.2 to help him catch up.

Let X and Y be the total winnings of the two players after 10 turns. Intuitively, X and Y should be positively correlated, due to the 0.2 bonus which brings them closer together. Let's see if this is true.

Though very simply stated, this problem is far too tough to solve mathematically in an elementary course (or even an advanced one). So, we will use simulation. In addition to finding the correlation between X and Y, we'll also find $F_{X,Y}(5.8, 5.2)$.

```
1  taketurn <- function(a,b) {
2     win <- runif(1)
3     if (a >= b) return(win)
4     else return(win+0.2)
5  }
6
7  cdf2 <- function(xy,t1,t2) {  # 2-dim. cdf
8     tmp <- xy[xy[,1] <= t1 & xy[,2] <= t2,]
9     return(nrow(tmp)/nrow(xy))
10 }
11
12 nturns <- 10
13 xyvals <- matrix(nrow=nreps,ncol=2)
14 for (rep in 1:nreps) {
15    x <- 0
16    y <- 0
17    for (turn in 1:nturns) {
18       # x's turn
19       x <- x + taketurn(x,y)
20       # y's turn
21       y <- y + taketurn(y,x)
22    }
23    xyvals[rep,] <- c(x,y)
24 }
25 print(cor(xyvals[,1],xyvals[,2]))
26 print(cdf2(xyvals,5.8,5.2))
```

The output is 0.65 and 0.03. So, X and Y are indeed positively correlated as we had surmised.

Note the use of R's built-in function **cor()** to compute correlation, a shortcut that allows us to avoid summing all the products **xy** and so on, from (5.25).

Note too that the bonus makes the two players' winnings "leapfrog" over each other. Without it, we would have EX = EY = 5.0, and $F_{X,Y}(5.8, 5.2)$ somewhat greater than 0.25. (The latter would be the value of $F_{X,Y}(5.0, 5.0)$.) But the bonus moves the distributions of X and Y more toward 10.0.

5.3 Sets of Independent Random Variables

Recall from Section 3.3:

Definition 18 *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).*

Intuitively, though, it simply means that knowledge of the value of X tells us nothing about the value of Y, and vice versa.

Great mathematical tractability can be achieved by assuming that the X_i in a random vector $X = (X_1, ..., X_k)$ are independent. In many applications, this is a reasonable assumption.

5.3.1 Properties

In the next few sections, we will look at some commonly-used properties of sets of independent random variables. For simplicity, consider the case k = 2, with X and Y being independent (scalar) random variables.

5.3.1.1 Probability Mass Functions and Densities Factor

Property E:

If X and Y are independent, then

$$p_{X,Y} = p_X p_Y \tag{5.49}$$

in the discrete case, and

$$f_{X,Y} = f_X f_Y \tag{5.50}$$

in the continuous case. In other words, the joint pmf/density is the product of the marginal ones.

This is easily seen in the discrete case:

$$\begin{aligned} p_{X,Y}(i,j) &= P(X = i \text{ and } Y = j) \text{ (definition)} && (5.51)\\ &= P(X = i)P(Y = j) \text{ (independence)} && (5.52)\\ &= p_X(i)p_Y(j) \text{ (definition))} && (5.53) \end{aligned}$$

Here is the proof for the continuous case;

$$\begin{aligned} f_{X,Y}(u,v) &= \frac{\partial^2}{\partial u \partial v} F_{X,Y}(u,v) && (5.54)\\ &= \frac{\partial^2}{\partial u \partial v} P(X \leq u \text{ and } Y \leq v) && (5.55)\\ &= \frac{\partial^2}{\partial u \partial v} \left[P(X \leq u) \cdot P(Y \leq v)\right] && (5.56)\\ &= \frac{\partial^2}{\partial u \partial v} F_X(u) \cdot F_Y(v) && (5.57)\\ &= f_X(u) f_Y(v) && (5.58) \end{aligned}$$

5.3.1.2 Expected Values Factor

Property F:

If X and Y are independent, then

$$E(XY) = E(X)E(Y) \tag{5.59}$$

To prove this, use (5.49) and (5.50) for the discrete and continuous cases.

5.3.1.3 Covariance Is 0

Property G:

If X and Y are independent, we have

$$Cov(X,Y) = 0 \tag{5.60}$$

and thus

$\rho(X,Y) = 0$ as well.

This follows from (5.59) and (5.25).

However, the converse is false. A counterexample is the random pair (V,W) that is uniformly distributed on the unit disk, $\{(s,t) : s^2 + t^2 \leq 1\}$. Clearly 0 = E(XY) = EX = EY due to the symmetry of the distribution about (0,0), so Cov(X,Y) = 0 by (5.25).

But X and Y just as clearly are not independent. If for example we know that $X > 0.8$, say, then $Y^2 < 1 - 0.8^2$ and thus $|Y| < 0.6$. If X and Y were independent, knowledge of X should not tell us anything about Y, which is not the case here, and thus they are not independent. If we also know that X and Y are bivariate normally distributed (Section 5.8.2.1), then zero covariance does imply independence.

5.3.1.4 Variances Add

Property H:

If X and Y are independent, then we have

$$Var(X+Y) = Var(X) + Var(Y). \tag{5.61}$$

This follows from (5.30) and (5.59).

5.4 Convolution

Definition 19 *Suppose g and h are densities of continuous random variables X and Y, respectively. The* **convolution** *of g and h, denoted g*h,*[3] *is another density, defined to be that of the random variable X+Y. In other words, convolution is a binary operation on the set of all densities.*

If X and Y are nonnegative, then the convolution reduces to

$$f_Z(t) = \int_0^t g(s)h(t-s)\,ds \tag{5.62}$$

You can get intuition on this by considering the discrete case. Say U and V are nonnegative integer-valued random variables, and set W = U+V. Let's find p_W;

$$\begin{aligned}
p_W(k) &= P(W = k) \text{ (by definition)} && (5.63)\\
&= P(U + V = k) \text{ (substitution)} && (5.64)\\
&= \sum_{i=0}^{k} P(U = i \text{ and } V = k - i) \text{ (“In what ways can it happen?”)} && (5.65)\\
&= \sum_{i=0}^{k} p_{U,V}(i, k - i) \text{ (by definition)} && (5.66)\\
&= \sum_{i=0}^{k} p_U(i)p_V(k - i) \text{ (from Section 5.3.1.1)} && (5.67)
\end{aligned}$$

Review the analogy between densities and pmfs in our unit on continuous random variables, Section 4.3.1, and then see how (5.62) is analogous to (5.63) through (5.67):

- k in (5.63) is analogous to t in (5.62)
- the limits 0 to k in (5.67) are analogous to the limits 0 to t in (5.62)
- the expression k-i in (5.67) is analogous to t-s in (5.62)
- and so on

[3] The reason for the asterisk, suggesting a product, will become clear in Section 5.11.3.

5.5 Examples

5.5.1 Example: Dice

In Section 5.2.1, we speculated that the correlation between X, the number on the blue die, and S, the total of the two dice, was positive. Let's compute it.

Write S = X + Y, where Y is the number on the yellow die. Then using the properties of covariance presented above, we have that

$$\begin{aligned} Cov(X,S) &= Cov(X,X+Y) \text{ (def. of S)} && (5.68)\\ &= Cov(X,X)+Cov(X,Y) \text{ (from } (5.22)) && (5.69)\\ &= Var(X)+0 \text{ (from } (5.24),\ (5.60)) && (5.70) \end{aligned}$$

Also, from (5.61),

$$Var(S) = Var(X+Y) = Var(X) + Var(Y) \tag{5.71}$$

But Var(Y) = Var(X). So the correlation between X and S is

$$\rho(X,S) = \frac{Var(X)}{\sqrt{Var(X)}\sqrt{2Var(X)}} = 0.707 \tag{5.72}$$

Since correlation is at most 1 in absolute value, 0.707 is considered a fairly high correlation. Of course, we did expect X and S to be highly correlated.

5.5.2 Example: Variance of a Product

Suppose X_1 and X_2 are independent random variables with $EX_i = \mu_i$ and $Var(X_i) = \sigma_i^2$, i = 1,2. Let's find an expression for $Var(X_1X_2)$.

$$\begin{aligned} Var(X_1X_2) &= E(X_1^2X_2^2) - [E(X_1X_2)]^2 \quad (3.29) && (5.73)\\ &= E(X_1^2)\cdot E(X_2^2) - \mu_1^2\mu_2^2 \quad (5.59) && (5.74)\\ &= (\sigma_1^2+\mu_1^2)(\sigma_2^2+\mu_2^2) - \mu_1^2\mu_2^2 && (5.75)\\ &= \sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 && (5.76) \end{aligned}$$

5.5.3 Example: Ratio of Independent Geometric Random Variables

Suppose X and Y are independent geometrically distributed random variables with success probability p. Let Z = X/Y. We are interested in EZ and F_Z.

First, by (5.59), we have

$$EZ = \frac{1}{p} \cdot E\left[\frac{1}{Y}\right] \tag{5.77}$$

so we need to find E(1/Y):

$$E(\frac{1}{Y}) = \sum_{i=1}^{\infty} \frac{1}{i}(1-p)^{i-1}p \tag{5.78}$$

Now let's find $F_Z(m)$ for a positive integer m.

$$\begin{aligned} F_Z(m) &= P\left(\frac{X}{Y} \leq m\right) && (5.79)\\ &= P(X \leq mY) && (5.80)\\ &= \sum_{i=1}^{\infty} P(Y=i)\, P(X \leq mY | Y=i) && (5.81)\\ &= \sum_{i=1}^{\infty} (1-p)^{i-1}p\, P(X \leq mi) && (5.82)\\ &= \sum_{i=1}^{\infty} (1-p)^{i-1}p\, [1-(1-p)^{mi}] && (5.83) \end{aligned}$$

this last step coming from (3.87).

We can actually reduce (5.83) to closed form, by writing

$$(1-p)^{i-1}(1-p)^{mi} = (1-p)^{mi+i-1} = \frac{1}{1-p}\left[(1-p)^{m+1}\right]^i \tag{5.84}$$

and then using (3.77). Details are left to the reader.

5.5.4 Example: Ethernet

Consider this network, essentially Ethernet. Here nodes can send at any time. Transmission time is 0.1 seconds. Nodes can also "hear" each other; one node will not start transmitting if it hears that another has a transmission in progress, and even when that transmission ends, the node that had been waiting will wait an additional random time, to reduce the possibility of colliding with some other node that had been waiting.

Suppose two nodes hear a third transmitting, and thus refrain from sending. Let X and Y be their random backoff times, i.e. the random times they wait before trying to send. (In this model, assume that they do not do "listen before talk" after a backoff.) Let's find the probability that they clash, which is $P(|X-Y| \leq 0.1)$.

Assume that X and Y are independent and exponentially distributed with mean 0.2, i.e. they each have density $5e^{-5u}$ on $(0, \infty)$. Then from (5.50), we know that their joint density is the product of their marginal densities,

$$f_{X,Y}(s,t) = 25e^{-5(s+t)}, s, t > 0 \tag{5.85}$$

Now

$$P(|X-Y| \leq 0.1) = 1 - P(|X-Y| > 0.1) = 1 - P(X > Y + 0.1) - P(Y > X + 0.1) \tag{5.86}$$

Look at that first probability. Applying (5.14) with $A = \{(s,t) : s > t + 0.1, 0 < s, t\}$, we have

$$P(X > Y + 0.1) = \int_0^\infty \int_{t+0.1}^\infty 25e^{-5(s+t)}\, ds\, dt = 0.303 \tag{5.87}$$

By symmetry, $P(Y > X + 0.1)$ is the same. So, the probability of a clash is 0.394, rather high. We may wish to increase our mean backoff time, though a more detailed analysis is needed.

5.5.5 Example: Analysis of Seek Time

This will be an analysis of seek time on a disk. Suppose we have mapped the innermost track to 0 and the outermost one to 1, and assume that (a) the number of tracks is large enough to treat the position H of the read/write head the interval [0,1] to be a continous random variable, and (b) the track number requested has a uniform distribution on that interval.

Consider two consecutive service requests for the disk, denoting their track numbers by X and Y. In the simplest model, we assume that X and Y are independent, so that the joint distribution of X and Y is the product of their marginals, and is thus is equal to 1 on the square $0 \leq X, Y \leq 1$.

The seek distance will be $|X - Y|$. Its mean value is found by taking g(s,t) in (5.15) to be $|s - t|$.

$$\int_0^1 \int_0^1 |s - t| \cdot 1 \; ds \; dt = \frac{1}{3} \tag{5.88}$$

By the way, what about the assumptions here? The independence would be a good assumption, for instance, for a heavily-used file server accessed by many different machines. Two successive requests are likely to be from different machines, thus independent. In fact, even within the same machine, if we have a lot of users at this time, successive requests can be assumed independent. On the other hand, successive requests from a particular user probably can't be modeled this way.

As mentioned in our unit on continuous random variables, page 87, if it's been a while since we've done a defragmenting operation, the assumption of a uniform distribution for requests is probably good.

Once again, this is just scratching the surface. Much more sophisticated models are used for more detailed work.

5.5.6 Example: Backup Battery

Suppose we have a portable machine that has compartments for two batteries. The main battery has lifetime X with mean 2.0 hours, and the backup's lifetime Y has mean life 1 hours. One replaces the first by the second as soon as the first fails. The lifetimes of the batteries are exponentially distributed and independent. Let's find the density of W, the time that the system is operational (i.e. the sum of the lifetimes of the two batteries).

Recall that if the two batteries had the same mean lifetimes, W would have a gamma distribution. But that's not the case here. But we notice that the distribution of W is a convolution of two exponential densities, as it is the sum of two nonnegative independent random variables. Using (5.4), we have

$$f_W(t) = \int_0^t f_X(s) f_Y(t - s) \; ds = \int_0^t 0.5 e^{-0.5s} e^{-(t-s)} \; ds = e^{-0.5t} - e^{-t}, \; 0 < t < \infty \tag{5.89}$$

5.5.7 Example: Minima of Independent Exponentially Distributed Random Variables

The memoryless property of the exponential distribution leads to other key properties. Here's a famous one:

Theorem 20 *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*

(a) Z is exponentially distributed with parameter $\lambda_1 + ... + \lambda_k$

(b) $P(Z = W_i) = \frac{\lambda_i}{\lambda_1 + ... + \lambda_k}$

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.

- The sum $\lambda_1 + ... + \lambda_n$ in (a) should make good intuitive sense to you, for the following reasons. Recall from Section 4.4.6.1 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 $\lambda_1 + \lambda_2$.

  Similarly, (b) should be intuitively clear as well from the above "thought experiment," since for instance a proportion $\lambda_1/(\lambda_1 + \lambda_2)$ of my shouts will be in response to person 1's shouts.

  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.

Proof

Properties (a) and (b) above are easy to prove, starting with the relation

$$\begin{aligned}
F_Z(t) &= P(Z \leq t) \quad \text{(def. of cdf)} && (5.90)\\
&= 1 - P(Z > t) && (5.91)\\
&= 1 - P(W_1 > t \text{ and ... and } W_k > t) \quad (\min > t \text{ iff all } W_i > t) && (5.92)\\
&= 1 - \Pi_i \, P(W_i > t) \quad \text{(indep.)} && (5.93)\\
&= 1 - \Pi_i \, e^{-\lambda_i t} \quad \text{(expon. distr.)} && (5.94)\\
&= 1 - e^{-(\lambda_1 + ... + \lambda_n)t} && (5.95)
\end{aligned}$$

Taking $\frac{d}{dt}$ of both sides shows (a).

For (b), suppose k = 2. we have that

$$
\begin{aligned}
P(Z = W_1) &= P(W_1 < W_2) && (5.96)\\
&= \int_0^\infty \int_t^\infty \lambda_1 e^{-\lambda_1 t} \lambda_2 e^{-\lambda_2 s} \, ds \, dt && (5.97)\\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2} && (5.98)
\end{aligned}
$$

The case for general k can be done by induction, writing $W_1 + ... + W_{c+1} = (W_1 + ... + W_c) + W_{c+1}$.

■

Note carefully: Just as the probability that a continuous random variable takes on a specific value is 0, the probability that two continuous and independent random variables are equal to each other is 0. Thus in the above analysis, $P(W_1 = W_2) = 0$.

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 Section 11.4.

5.5.8 Example: Computer Worm

A computer science graduate student at UCD, C. Senthilkumar, 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 as a continuous-time Markov chain. Again, this topic is much more fully developed in Section 11.4, but all we need here is the result of Section 5.5.7.

In state i, there are i infected hosts, each trying to infect all of the g-i noninfected hots. 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. 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

$$\sum_{i=1}^{g-1} \frac{1}{i(g-i)} \tag{5.99}$$

Using a calculus approximation, we have

$$\int_1^{g-1} \frac{1}{x(g-x)}\, dx = \frac{1}{g} \int_1^{g-1} (\frac{1}{x} + \frac{1}{g-x})\, dx = \frac{2}{g} \ln(g-1) \tag{5.100}$$

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, (5.99) 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.

5.6 Matrix Formulations

(Note that there is a review of matrix algebra in Appendix A.)

In your first course in matrices and linear algebra, your instructor probably motivated the notion of a matrix by using an example involving linear equations, as follows.

Suppose we have a system of equations

$$a_{i1}x_1 + ... + a_{in}x_n = b_i, \;\; i = 1, ..., n, \tag{5.101}$$

where the x_i are the unknowns to be solved for.

This system can be represented compactly as

$$AX = B, \tag{5.102}$$

where A is nxn and X and B are nx1.

That compactness coming from the matrix formulation applies to statistics too, though in different ways, as we will see. (Linear algebra in general is used widely in statistics—matrices, rank and subspace, eigenvalues, even determinants.)

When dealing with multivariate distributions, some very messy equations can be greatly compactified through the use of matrix algebra. We will introduce this here.

Throughout this section, consider a random vector $W = (W_1, ..., W_k)'$ where $'$ denotes matrix transpose, and a vector written horizontally like this without a $'$ means a row vector.

5.6.1 Properties of Mean Vectors

Definition 21 *The expected value of W is defined to be the vector*

$$EW = (EW_1, ..., EW_k)' \tag{5.103}$$

The linearity of the components implies that of the vectors:

Property I: For any scalar constants c and d, and any random vectors V and W, we have

$$E(cV + dW) = cEV + dEW \tag{5.104}$$

where the multiplication and equality is now in the vector sense.

Also, multiplication by a constant matrix factors:

Property J: If A is a nonrandom matrix having k columns, then

$$E(AW) = AEW \tag{5.105}$$

5.6.2 Covariance Matrices

Definition 22 *The covariance matrix Cov(W) of* $W = (W_1, ..., W_k)'$ *is the k x k matrix whose* $(i, j)^{th}$ *element is* $Cov(W_i, W_j)$.

Note that that implies that the diagonal elements of the matrix are the variances of the W_i, and that the matrix is symmetric.

As you can see, in the statistics world, the Cov() notation is "overloaded." If it has two arguments, it is ordinary covariance, between two variables. If it has one argument, it is the covariance matrix, consisting of the covariances of all pairs of components in the argument. When people mean the matrix form, they always say so, i.e. they say "covariance MATRIX" instead of just "covariance."

The covariance matrix is just a way to compactly do operations on ordinary covariances. Here are some important properties:

- **Property K:** Say c is a constant scalar. Then cW is a k-component random vector like W, and

$$Cov(cW) = c^2 Cov(W) \tag{5.106}$$

- **Property L:** If A is an r x k but nonrandom matrix. Then AW is an r-component random vector, and

$$Cov(AW) = A\, Cov(W)\, A' \tag{5.107}$$

- **Property M:** Suppose V and W are independent random vectors, meaning that each component in V is independent of each component of W. (But this does NOT mean that the components within V are independent of each other, and similarly for W.) Then

$$Cov(V + W) = Cov(V) + Cov(W) \tag{5.108}$$

- **Property N:** In analogy with (3.29), for any random vector Q,

$$Cov(Q) = E(QQ') - EQ\ (EQ)' \tag{5.109}$$

5.7 Conditional Distributions

The key to good probability modeling and statistical analysis is to understand conditional probability. The issue arises constantly.

5.7.1 Conditional Pmfs and Densities

First, let's review: In many repetitions of our "experiment," P(A) is the long-run proportion of the time that A occurs. By contrast, P(A|B) is the long-run proportion of the time that A occurs, *among those repetitions in which B occurs.* Keep this in your mind at all times.

Now we apply this to pmfs, densities, etc. We define the conditional pmf as follows for discrete random variables X and Y:

$$p_{Y|X}(j|i) = P(Y = j|X = i) = \frac{p_{X,Y}(i,j)}{p_X(i)} \tag{5.110}$$

By analogy, we define the conditional density for continuous X and Y:

$$f_{Y|X}(t|s) = \frac{f_{X,Y}(s,t)}{f_X(s)} \tag{5.111}$$

5.7.2 Conditional Expectation

Conditional expectations are defined as straightforward extensions of (5.110) and (5.111):

$$E(Y|X=i) = \sum_j j p_{Y|X}(j|i) \tag{5.112}$$

$$E(Y|X=s) = \int_t t f_{Y|X}(t|s)\, dt \tag{5.113}$$

5.7.3 The Law of Total Expectation (advanced topic)

5.7.3.1 Conditional Expected Value As a Random Variable

For a random variable Y and an event A, the quantity E(Y|A) is the long-run average of Y, among the times when A occurs. Note several things about the expression E(Y|A):

- The item to the left of the | symbol is a *random variable* (Y).
- The item on the right of the | symbol is an *event* (A).
- The overall expression evaluates to a constant.

By contrast, for the quantity E(Y|W) to be defined shortly for a random variable W, it is the case that:

- The item to the left of the | symbol is a random variable (Y).
- The item to the right of the | symbol is a random variable (W).
- The overall expression itself is a random variable, not a constant.

It will be very important to keep these differences in mind.

Consider the function g(t) defined as[4]

$$g(t) = E(Y|W = t) \tag{5.114}$$

In this case, the item to the right of the | is an event, and thus g(t) is a constant (for each value of t), not a random variable.

Definition 23 *Define g() as in (5.114). Form the new random variable Q = g(W). Then the quantity E(Y|W) is defined to be Q.*

(Before reading any further, re-read the two sets of bulleted items above, and make sure you understand the difference between E(Y|W=t) and E(Y|W).)

One can view E(Y|W) as a projection in an abstract vector space. This is very elegant, and actually aids the intuition. If (and only if) you are mathematically adventurous, read the details in Section 5.12.2.

5.7.3.2 Famous Formula: Theorem of Total Expectation

An extremely useful formula, given only scant or no mention in most undergraduate probability courses, is

$$E(Y) = E[E(Y|W)] \tag{5.115}$$

for any random variables Y and W (for which the expectations are defined).

The RHS of (5.115) looks odd at first, but it's merely E[g(W)]; since Q = E(Y|W) is a random variable, we can certainly ask what its expected value is.

Equation (5.115) is a bit abstract. It's a very useful abstraction, enabling streamlined writing and thinking about the probabilistic structures at hand. Still, you may find it helpful to consider the case of discrete W, in which (5.115) has the more concrete form

$$EY = \sum_i P(W = i) \cdot E(Y|W = i) \tag{5.116}$$

To see this intuitively, think of measuring the heights and weights of all the adults in Davis. Say we measure height to the nearest inch, so that height is discrete. We look at all the adults in Davis who are 72 inches tall, and write down their mean weight. Then we write down the mean weight of all adults of height 68. Then we write down the mean weight of all adults of height 75, and so on. Then (5.115) says that if we take

[4] Of course, the t is just a placeholder, and any other letter could be used.

the average of all the numbers we write down—the average of the averages—then we get the mean weight among *all* adults in Davis.

Note carefully, though, that this is a *weighted* average. If for instance people of height 69 inches are more numerous in the population, then their mean weight will receive greater emphasis in over average of all the means we've written down. This is seen in (5.116), with the weights being the quantities P(W=i).

The relation (5.115) is proved in the discrete case in Section 5.13.

5.7.4 What About the Variance?

By the way, one might guess that the analog of the Theorem of Total Expectation for variance is

$$Var(Y) = E[Var(Y|W)] \tag{5.117}$$

But this is false. Think for example of the extreme case in which Y = W. Then Var(Y|W) would be 0, but Var(Y) would be nonzero.

The correct formula, called the Law of Total Variance, is

$$Var(Y) = E[Var(Y|W)] + Var[E(Y|W)] \tag{5.118}$$

Deriving this formula is easy, by simply evaluating both sides of **bis**, and using the relation $Var(X) = E(X^2) - (EX)^2$. This exercise is left to the reader. See also Section 5.12.2.

5.7.5 Example: Trapped Miner

(Adapted from *Stochastic Processes,* by Sheldon Ross, Wiley, 1996.)

A miner is trapped in a mine, and has a choice of three doors. Though he doesn't realize it, if he chooses to exit the first door, it will take him to safety after 2 hours of travel. If he chooses the second one, it will lead back to the mine after 3 hours of travel. The third one leads back to the mine after 5 hours of travel. Suppose the doors look identical, and if he returns to the mine he does not remember which door(s) he tried earlier. What is the expected time until he reaches safety?

Let Y be the time it takes to reach safety, and let W denote the number of the door chosen (1, 2 or 3) on the first try. Then let us consider what values E(Y|W) can have. If W = 1, then Y = 2, so

$$E(Y|W = 1) = 2 \tag{5.119}$$

If W = 2, things are a bit more complicated. The miner will go on a 3-hour excursion, and then be back in its original situation, and thus have a further expected wait of EY, since "time starts over." In other words,

$$E(Y|W=2) = 3 + EY \tag{5.120}$$

Similarly,

$$E(Y|W=3) = 5 + EY \tag{5.121}$$

In summary, now considering the *random variable* E(Y|W), we have

$$Q = E(Y|W) = \begin{cases} 2, & w.p. \frac{1}{3} \\ 3 + EY, & w.p. \frac{1}{3} \\ 5 + EY, & w.p. \frac{1}{3} \end{cases} \tag{5.122}$$

where "w.p." means "with probability." So, using (5.115) or (5.116), we have

$$EY = EQ = 2 \times \frac{1}{3} + (3 + EY) \times \frac{1}{3} + (5 + EY) \times \frac{1}{3} = \frac{10}{3} + \frac{2}{3} EY \tag{5.123}$$

Equating the extreme left and extreme right ends of this series of equations, we can solve for EY, which we find to be 10.

It is no accident that the answer, 10, is 2+3+5. This was discovered by UCD grad student Ahmed Ahmedin. Here's why (different from Ahmed's reasoning):

Let N denote the total number of attempts the miner makes before escaping (including the successful attempt at the end), and let U_i denote the time spent traveling during the i^{th} attempt, i = 1,...,N. Then

$$\begin{aligned} ET &= E(U_1 + ..., + U_N) & (5.124) \\ &= E\left[E(U_1 + ..., + U_N|N)\right] & (5.125) \end{aligned}$$

Given N, each of $U_1, ..., U_{N-1}$ takes on the values 3 and 5, with probability 0.5 each, while U_N is the constant 2. Thus

$$E(U_1 + ..., + U_N|N) = (N-1)\frac{3+5}{2} + 2 = 4N - 2 \tag{5.126}$$

N has a geometric distribution with p = 1/3, thus mean 3. Putting all this together, we have

$$ET = E(U_1 + ..., +U_N) = E(4N - 2) = 10 \tag{5.127}$$

This would be true if 2, 3 and 5 were replaced by *a*, *b* and *c*. In other words, intuitively: It takes an average of 3 attempts to escape, with mean time of (a+b+c)/3, so the mean time overall is a+b+c.

It is left to the reader to see how this would change if we assume that the miner remembers which doors he has already hit.

5.7.5.1 Example: More on Flipping Coins with Bonuses

Recall the situation of Section 3.13.6: 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. We'll compute EX.

As before, Y denote the number of heads you obtain through nonbonus flips. This is a natural situation in which to try the Theorem of Total Expectation, conditioning on Y. Reason as follows:

It would be tempting to say that, given Y = m, X has a binomial distribution with parameters m and 0.5. That is not correct, but what is true is the random variable X-m does have that distribution. Note by the way that X-Y is the number of bonus flips.

Then

$$\begin{aligned} EX &= E[E(X|Y)] && (5.128)\\ &= E\left[E(\{X - Y\} + Y|Y)\right] && (5.129)\\ &= E\left[0.5Y + Y\right] && (5.130)\\ &= 1.5EY && (5.131)\\ &= 0.75k && (5.132) \end{aligned}$$

5.7.6 Example: Analysis of Hash Tables

(Famous example, adapted from various sources.)

Consider a database table consisting of m cells, only some of which are currently occupied. Each time a new key must be inserted, it is used in a hash function to find an unoccupied cell. Since multiple keys map to the same table cell, we may have to probe multiple times before finding an unoccupied cell.

We wish to find E(Y), where Y is the number of probes needed to insert a new key. One approach to doing so would be to condition on W, the number of currently occupied cells at the time we do a search. After finding E(Y|W), we can use the Theorem of Total Expectation to find EY. We will make two assumptions (to be discussed later):

(a) Given that W = k, each probe will collide with an existing cell with probability k/m, with successive probes being independent.

(b) W is uniformly distributed on the set 1,2,...,m, i.e. P(W = k) = 1/m for each k.

To calculate E(Y|W=k), we note that given W = k, then Y is the number of independent trials until a "success" is reached, where "success" means that our probe turns out to be to an unoccupied cell. This is a **geometric** distribution, i.e.

$$P(Y = r|W = k) = (\frac{k}{m})^{r-1}(1 - \frac{k}{m}) \tag{5.133}$$

The mean of this geometric distribution is, from (3.75),

$$\frac{1}{1 - \frac{k}{m}} \tag{5.134}$$

Then

$$\begin{aligned} EY &= E[E(Y|W)] && (5.135)\\ &= \sum_{k=1}^{m-1} \frac{1}{m} E(Y|W = k) && (5.136)\\ &= \sum_{k=1}^{m-1} \frac{1}{m-k} && (5.137)\\ &= 1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{m-1} && (5.138)\\ &\approx \int_1^m \frac{1}{u} du && (5.139)\\ &= ln(m) && (5.140) \end{aligned}$$

where the approximation is something you might remember from calculus (you can picture it by drawing rectangles to approximate the area under the curve.).

Now, what about our assumptions, (a) and (b)? The assumption in (a) of each cell having probability k/m should be reasonably accurate if k is much smaller than m, because hash functions tend to distribute probes uniformly, and the assumption of independence of successive probes is all right too, since it is very unlikely that we would hit the same cell twice. However, if k is not much smaller than m, the accuracy will suffer.

Assumption (b) is more subtle, with differing interpretations. For example, the model may concern one specific database, in which case the assumption may be questionable. Presumably W grows over time, in which case the assumption would make no sense—it doesn't even *have* a distribution. We could instead think of a database which grows and shrinks as time progresses. However, even here, it would seem that W would probably oscillate around some value like m/2, rather than being uniformly distributed as assumed here. Thus, this model is probably not very realistic. However, even idealized models can sometimes provide important insights.

5.8 Parametric Families of Distributions

Since there are so many ways in which random variables can correlate with each other, there are rather few parametric families commonly used to model multivariate distributions (other than those arising from sets of independent random variables have a distribution in a common parametric univariate family). We will discuss two here.

5.8.1 The Multinomial Family of Distributions

5.8.1.1 Probability Mass Function

This is a generalization of the binomial family.

Suppose one tosses a die 8 times. What is the probability that the results consist of two 1s, one 2, one 4, three 5s and one 6? Well, if the tosses occur in that order, i.e. the two 1s come first, then the 2, etc., then the probability is

$$\left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^1 \tag{5.141}$$

But there are many different orderings, in fact

$$\frac{8!}{2!1!0!1!3!1!} \tag{5.142}$$

of them, from Section 2.13.3.

From this, we can more generally see the following. Suppose:

- we have n trials, each of which has r possible outcomes or categories
- the trials are independent
- the ith outcome has probability p_i

Let X_i denote the number of trials with outcome i, i = 1,...,r. In the die example above, for instance, r = 6 for the six possible outcomes of one trial, i.e. one roll of the die, and X_1 is the number of times we got one dot, in our n = 8 rolls.

Then we say that the vector $X = (X_1, ..., X_r)$ have a **multinomial distribution**. Since the X_i are discrete random variables, they have a joint pmf $p_{X_1,...,X_r}()$. Taking the above die example for illustration again, the probability of interest there is $p_X(2, 1, 0, 1, 3, 1)$. We then have in general,

$$p_{X_1,...,X_r}(j_1, ..., j_r) = \frac{n!}{j_1!...j_r!} p_1^{j_1} ... p_r^{j_r} \tag{5.143}$$

Note that this family of distributions has r+1 parameters.

We can simulate multinomial random vectors in R using the **sample()** function:

```
# n is the number of trial, p the vector of probabilities of the r
# categories
multinom <- function(n,p) {
   r <- length(p)
   outcome <- sample(x=1:r,size=n,replace=T,prob=p)
   counts <- vector(length=r)  # counts of the various categories
   # tabulate the counts (could be done more efficiently)
   for (i in 1:n) {
       j <- outcome[i]
       counts[j] <- counts[j] + 1
   }
   return(counts)
}
```

5.8.1.2 Mean Vectors and Covariance Matrices in the Multinomial Family

Now look at the vector $X = (X_1, ..., X_r)'$. Let's find its mean vector and covariance matrix.

First, note that the marginal distributions of the X_i are binomial! So,

$$EX_i = np_i \text{ and } Var(X_i) = np_i(1 - p_i) \tag{5.144}$$

So we know EX now:

$$EX = \begin{pmatrix} np_1 \\ ... \\ np_r \end{pmatrix} \tag{5.145}$$

We also know the diagonal elements of Cov(X)—$np_i(1-p_i)$ is the ith diagonal element, i = 1,...,r.

But what about the rest? To this end, let T_{ki} be the indicator random variable of the event that the kth trial results in outcome i, k = 1,...,n and i = 1,...,r. (Recall Section 3.6.)

Make sure you understand that

$$X_i = \sum_{k=1}^{n} T_{ki} \tag{5.146}$$

Then for $i \neq j$,

$$\begin{aligned} Cov(X_i, X_j) &= Cov(T_{1i} + ..., T_{ni}, T_{1j} + ..., T_{nj}) && (5.147) \\ &= \sum_{c=1}^{n} \sum_{d=1}^{n} Cov(T_{ci}, T_{dj}) \text{ (5.22)} && (5.148) \\ &= \sum_{c=1}^{n} Cov(T_{ci}, T_{cj}) \text{ (indep. trials, (5.60))} && (5.149) \end{aligned}$$

Note that the abbreviated reason "indep. trials" above refers to the following. In the double sum above, consider what happens in the case $c \neq d$. Since T_{ci} and T_{dj} are random variables associated with trials c and d, respectively, they are independent.

But for a fixed trial, in this case the cth, only one of the Ts can be 1, with the rest being 0. So, $T_{ci} \cdot T_{cj} = 0$! And since the Ts are indicator random variables, we have $ET_{ci} = p_i$ and the same for the j case. So, from (5.25), then for $i \neq j$,

$$Cov(T_{ci}, T_{cj}) = -p_i p_j \tag{5.150}$$

So, the above reduces to

$$Cov(X_i, X_j) = -np_i p_j \tag{5.151}$$

Putting all this together, we see that

$$Cov(X) = n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & ... & -p_1p_r \\ -p_1p_2 & p_2(1-p_2) & ... & -p_2p_r \\ ... & ... & ... & ... \\ ... & ... & ... & p_r(1-p_r) \end{pmatrix} \tag{5.152}$$

Note too that if we define R = X/n, so that R is the vector of proportions in the various categories (e.g. X_1/n is the fraction of trials that resulted in category 1), then (5.152) and (5.106), we have

$$Cov(R) = \frac{1}{n} \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & ... & -p_1p_r \\ -p_1p_2 & p_2(1-p_2) & ... & -p_2p_r \\ ... & ... & ... & ... \\ ... & ... & ... & p_r(1-p_r) \end{pmatrix} \tag{5.153}$$

Whew! That was a workout, but these formulas will become very useful later on, both in this unit and subsequent ones.

5.8.1.3 Application: Text Mining

One of the branches of computer science in which the multinomial family plays a prominent role is in text mining. One goal is automatic document classification. We want to write software that will make reasonably accurate guesses as to whether a document is about sports, the stock market, elections etc., based on the frequencies of various key words the program finds in the document.

Many of the simpler methods for this use the **bag of words model**. We have r key words we've decided are useful for the classification process, and the model assumes that statistically the frequencies of those words in a given document category, say sports, follow a multinomial distribution. Each category has its own set of probabilities $p_1, ..., p_r$. For instance, if "Barry Bonds" is considered one word, its probability will be much higher in the sports category than in the elections category, say. So, the observed frequencies of the words in a particular document will hopefully enable our software to make a fairly good guess as to the category the document belongs to.

Once again, this is a very simple model here, designed to just introduce the topic to you. Clearly the multinomial assumption of independence between trials is grossly incorrect here, most models are much more complex than this.

5.8.2 The Multivariate Normal Family of Distributions

Note to the reader: This is a more difficult section, but worth putting extra effort into, as so many statistical applications in computer science make use of it. It will seem hard at times, but in the end won't be too bad.

5.8.2.1 Densities and Properties

Intuitively, this family has densities which are shaped like multidimensional bells, just like the univariate normal has the famous one-dimensional bell shape.

Let's look at the bivariate case first. The joint distribution of X_1 and X_2 is said to be **bivariate normal** if their density is

$$f_{X,Y}(s,t) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(s-\mu_1)^2}{\sigma_1^2}+\frac{(t-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(s-\mu_1)(t-\mu 2)}{\sigma_1\sigma_2}\right]}, \quad -\infty < s,t < \infty \tag{5.154}$$

This looks horrible, and it is. But don't worry, as we won't work with this directly. It's important for conceptual reasons, as follows.

First, note parameters here: μ_1, μ_2, σ_1 and σ_2 are the means and standard deviations of X and Y, while ρ is the correlation between X and Y. So, we have a five-parameter family of distributions.

Now, let's look at some pictures, generated by R code which I've adapted from one of the entries in the R Graph Gallery, `http://addictedtor.free.fr/graphiques/graphcode.php?graph=42`.[5] Both are graphs of bivariate normal densities, with $EX_1 = EX_2 = 0$, $Var(X_1) = 10$, $Var(X_2) = 15$ and a varying value of the correlation ρ between X_1 and X_2. Figure 5.1 is for the case $\rho = 0.2$.

The surface is bell-shaped, though now in two dimensions instead of one. Again, the height of the surface at any (s,t) point the relative likelihood of X_1 being near s and X_2 being near t. Say for instance that X_1 is height and X_2 is weight. If the surface is high near, say, (70,150) (for height of 70 inches and weight of 150 pounds), it mean that there are a lot of people whose height and weight are near those values. If the surface is rather low there, then there are rather few people whose height and weight are near those values.

Now compare that picture to Figure 5.2, with $\rho = 0.8$.

Again we see a bell shape, but in this case "narrower." In fact, you can see that when X_1 (s) is large, X_2 (t) tends to be large too, and the same for "large" replaced by small. By contrast, the surface near (5,5) is much higher than near (5,-5), showing that the random vector (X_1, X_2) is near (5,5) much more often than (5,-5).

[5] There appears to be an error in their definition of the function **f()**; the assignment to **term5** should not have a negative sign at the beginning.

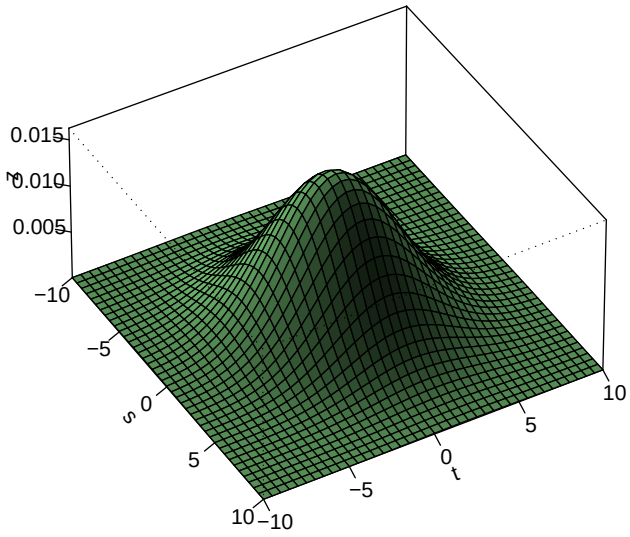


Figure 5.1: Bivariate Normal Density, $\rho = 0.2$

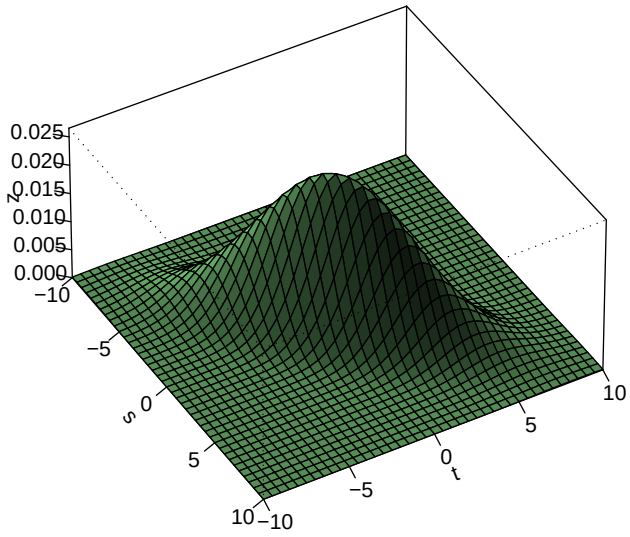


Figure 5.2: Bivariate Normal Density, $\rho = 0.8$

All of this reflects the high correlation (0.8) between the two variables. If we were to continue to increase ρ toward 1.0, we would see the bell become narrower and narrower, with X_1 and X_2 coming closer and closer to a linear relationship, one which can be shown to be

$$X_1 - \mu_1 = \frac{\sigma_1}{\sigma_2}(X_2 - \mu_2) \tag{5.155}$$

In this case, that would be

$$X_1 = \sqrt{\frac{10}{15}}X_2 = 0.82X_2 \tag{5.156}$$

The multivariate normal family of distributions is parameterized by one vector-valued quantity, the mean μ, and one matrix-valued quantity, the covariance matrix Σ. Specifically, suppose the random vector $X = (X_1, ..., X_k)'$ has a k-variate normal distribution.

The density has this form:

$$f_X(t) = ce^{-0.5(t-\mu)'\Sigma^{-1}(t-\mu)} \tag{5.157}$$

where

$$c = \frac{1}{(2\pi)^{k/2}\sqrt{det(\Sigma)}} \tag{5.158}$$

Here again $'$ denotes matrix transpose, -1 denotes matrix inversion and det() means determinant. Again, note that t is a kx1 vector.

Since the matrix is symmetric, there are k(k+1)/2 distinct parameters there, and k parameters in the mean vector, for a total of k(k+3)/2 parameters for this family of distributions.

The family has the following important properties:

Theorem 24 (Properties of Multivariate Normal Distributions)

Suppose $X = (X_1, ..., X_k)$ has a multivariate normal distribution with mean vector μ and covariance matrix Σ. Then:

(a) The contours of f_X are k-dimensional ellipsoids. In the case k = 2 for instance, where we can visualize the density of X as a three-dimensional surface, the contours for points at which the bell has the same height (think of a topographical map) are elliptical in shape. The larger the correlation (in absolute

value) between X_1 and X_2, the more elongated the ellipse. When the absolute correlation reaches 1, the ellipse degenerates into a straight line.

(b) Let A be a constant (i.e. nonrandom) matrix with k columns. Then the random vector Y = AX also has a multivariate normal distribution, with mean $A\mu$ and covariance matrix $A\Sigma A'$. (Note that the statements here about the mean vector and covariance matrix hold even if X does not have a multivariate normal distribution, by (5.105) and (5.107).)

(c) If $U_1, ..., U_m$ are each univariate normal and they are independent, then they jointly have a multivariate normal distribution. (In general, though, having a normal distribution for each U_i does not imply that they are jointly multivariate normal.)

(d) Suppose W has a multivariate normal distribution. The conditional distribution of some components of W, given other components, is again multivariate normal.

Part [(b)] has some important implications:

(i) The lower-dimensional marginal distributions are also multivariate normal. For example, if k = 3, the pair $(X_1, X_3)'$ has a bivariate normal distribution, as can be seen by setting

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5.159}$$

(ii) Scalar linear combinations of X are normal. In other words, for constant scalars $a_1, ..., a_k$, set $a = (a_1, ..., a_k)'$. Then the quantity $Y = a_1X_1 + ... + a_kX_k$ has a univariate normal distribution with mean $a'\mu$ and variance $a'\Sigma a$.

(iii) Vector linear combinations are multivariate normal. Again using the case k = 3 as our example, consider $(U, V)' = (X_1 - X_3, X_2 - X_3)$. Then set

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \tag{5.160}$$

In R the density, cdf and quantiles of the multivariate normal distribution are given by the functions **dmvnorm()**, **pmvnorm()** and **qmvnorm()** in the library **mvtnorm**. You can simulate a multivariate normal distribution by using **mvrnorm()** in the library **MASS**.

5.8.2.2 The Multivariate Central Limit Theorem

The multidimensional version of the Central Limit Theorem holds. A sum of independent identically distributed random vectors has an approximate multivariate normal distribution.

For example, since a person's body consists of many different components, the CLT (a non-independent, non-identically version of it) explains intuitively why heights and weights are approximately bivariate normal. Histograms of heights will look approximately bell-shaped, and the same is true for weights. The multivariate CLT says that three-dimensional histograms—plotting frequency along the "Z" axis against height and weight along the "X" and "Y" axes—will be approximately three-dimensional bell-shaped.

5.8.2.3 Example: Dice Game

Suppose we roll a die 50 times. Let X denote the number of rolls in which we get one dot, and let Y be the number of times we get either two or three dots. For convenience, let's also define Z to be the number of times we get four or more dots, though our focus will be on X and Y. Suppose also that we win \$5 for each roll of a one, and \$2 for each roll of a two or three.

Let's find the approximate values of the following:

- $P(X \leq 12 \text{ and } Y \leq 16)$
- P(win more than \$90)
- $P(X > Y > Z)$

The triple (X,Y,Z) has a multinomial distribution with n = 50 and three possible outcomes (1; 2 or 3; 4, 5 or 6), with $p_1 = 1/6$, $p_2 = 1/3$ and $p_3 = 1/2$.

Moreover, (X,Y,Z) is a sum of independent, identically distributed vectors. This comes from (5.146). Thus (X,Y,Z) has an approximately multivariate normal distribution.

These probabilities of interest to us here would be quite difficult to find directly. For $P(X \leq 12 \text{ and } Y \leq 16)$, for instance, we would need to sum (5.143) over many, many different cases. So, the CLT will be very valuable here.

We'll of course need to know the mean vector and covariance matrix of the random vector $(X, Y, Z)'$. We have those from (5.144) and (5.152):

$$E[(X, Y, Z)] = (50/6, 50/3, 50/2) \tag{5.161}$$

and

$$Cov[(X, Y, Z)] = 50 \begin{pmatrix} 5/36 & -1/18 & -1/12 \\ -1/18 & 2/9 & -1/6 \\ -1/12 & -1/6 & 1/4 \end{pmatrix} \tag{5.162}$$

We use the R function **pmvnorm()** introduced in Section 5.8.2.1. To account for the integer nature of X and Y, we call the function with upper limits of 12.5 and 16.5, rather than 12 and 16, which is often used to get a better approximation. Our code is

```
p1 <- 1/6
p23 <- 1/3
meanvec <- 50*c(p1,p23)
var1 <- 50*p1*(1-p1)
var23 <- 50*p23*(1-p23)
covar123 <- -50*p1*p23
covarmat <- matrix(c(var1,covar123,covar123,var23),nrow=2)
print(pmvnorm(upper=c(12.5,16.5),mean=meanvec,sigma=covarmat))
```

We find that

$$P(X \leq 12 \text{ and } Y \leq 16) \approx 0.43 \tag{5.163}$$

Now, let's find the probability that our total winnings, W, is over \$90. We know that W = 5X + 2Y, and Theorem 24(b) tells us that linear combinations of a multivariate normal random vector are (univariate) normal. In other words, W has an approximate normal distribution!

We thus need the mean and variance of W. The mean is easy:

$$EW = E(5X + 2Y) = 5EX + 2EY = 250/6 + 100/3 = 75 \tag{5.164}$$

For the variance, take the matrix A to be the row vector (5,2) in the theorem—or for that matter, in (5.107) —giving us Var(W) = 162.5. Then

$$P(W > 90) = 1 - \Phi\left(\frac{90 - 75}{162.5^{0.5}}\right) = 0.12 \tag{5.165}$$

Now to find $P(X > Y > Z)$, we need to work with $(U, V)' = (X - Y, Y - Z)$, so set

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \tag{5.166}$$

and then proceed as before to find $P(U > 0, V > 0)$.

By the way, note that the fact that Z is an exact linear function of X and Y turns out to make the covariance matrix Σ **singular**, i.e. lacking an inverse. That would create problems in (5.157).

5.8.2.4 Application: Data Mining

The multivariate normal family plays a central role in multivariate statistical methods.

For instance, a major issue in data mining is **dimension reduction**, which means trying to reduce what may be hundreds or thousands of variables down to a manageable level. One of the tools for this, called **principle components analysis** (PCA), is based on multivariate normal distributions. Google uses this kind of thing quite heavily. We'll discuss PCA in Section 10.5.1.

To see a bit of how this works, note that in Figure 5.2, X_1 and X_2 had nearly a linear relationship with each other. That means that one of them is nearly redundant, which is good if we are trying to reduce the number of variables we must work with.

In general, the method of principle components takes r original variables, in the vector X and forms r new ones in a vector Y, each of which is some linear combination of the original ones. These new ones are independent. In other words, there is a square matrix A such that the components of Y = AX are independent. (The matrix A consists of the eigenvectors of Cov(X); more on this in Section 10.5.1 of our unit on statistical relations.

We then discard the Y_i with small variance, as that means they are nearly constant and thus do not carry much information. That leaves us with a smaller set of variables that still captures most of the information of the original ones.

Many analyses in bioinformatics involve data that can be modeled well by multivariate normal distributions. For example, in automated cell analysis, two important variables are forward light scatter (FSC) and sideward light scatter (SSC). The joint distribution of the two is approximately bivariate normal.[6]

5.9 Simulation of Random Vectors

Let $X = (X_1, ..., X_k)'$ be a random vector having a specified distribution. How can we write code to simulate it? It is not always easy to do this. We'll discuss a couple of easy cases here, and illustrate what one may do in other situations.

The easiest case (and a very frequenly-occurring one) is that in which the X_i are independent. One simply simulates them individually, and that simulates X!

Another easy case is that in which X has a multivariate normal distribution. We noted in Section 5.8.2.1 that R includes the function **mvrnorm()**, which we can use to simulate our X here. The way this function works is to use the notion of principle components mentioned in Section 5.8.2.4. We construct Y = AX for the matrix A discussed there. The Y_i are independent, thus easily simulated, and then we transform back to

[6] See *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, edited by Robert Gentleman, Wolfgang Huber, Vincent J. Carey, Rafael A. Irizarry and Sandrine Dudoit, Springer, 2005.

X via $X = A^{-1}Y$

In general, though, things may not be so easy. For instance, consider the distribution in (5.17). There is no formulaic solution here, but the following strategy works.

First we find the (marginal) density of X. As in the case for Y shown in (5.20), we compute

$$f_X(s) = \int_0^s 8st\ dt = 4s^3 \tag{5.167}$$

Using the method shown in our unit on continuous probability, Section 4.6, we can simulate X as

$$X = F_X^{-1}(W) \tag{5.168}$$

where W is a U(0,1) random variable, generated as **runif(1)**. Since $F_X(u) = u^4$, $F_X^{-1}(v) = v^{0.25}$, and thus our code to simulate X is

```
runif(1)^0.25
```

Now that we have X, we can get Y. We know that

$$f_{Y|X}(t|S) = \frac{8st}{4s^3} = \frac{2}{s^2}t \tag{5.169}$$

Remember, s is considered constant. So again we use the "inverse-cdf" method here to find Y, given X, and then we have our pair (X,Y).

5.10 Mixture Models

To introduce this topic, suppose men's heights are normally distributed with mean 70 and standard deviation 3, with women's heights being normal with mean 66 and standard deviation 2.5. Let H denote the height of a randomly selected person from the entire population, and let G be the person's gender, 1 for male and 2 for female.

Then the conditional distribution of H, given G = 1, is N(70,9), and a similar statement holds for G = 2. But what about the unconditional distribution of H? We can derive it:

$$
\begin{aligned}
f_H(t) &= \frac{d}{dt} F_H(t) && (5.170)\\
&= \frac{d}{dt} P(H \leq t) && (5.171)\\
&= \frac{d}{dt} P(H \leq t \text{ and } G = 1 \text{ or } H \leq t \text{ and } G = 2) && (5.172)\\
&= \frac{d}{dt} \left[0.5P(H \leq t | G = 1) + 0.5P(H \leq t | G = 2) \right] && (5.173)\\
&= \frac{d}{dt} \left[0.5F_{H|G=1}(t) + 0.5F_{H|G=2}(t) \right] && (5.174)\\
&= 0.5 f_{H|G=1}(t) + 0.5 f_{H|G=2}(t) && (5.175)
\end{aligned}
$$

So the density of H in the grand population is the average of the densities of H in the two subpopulations. This makes intuitive sense.

In terms of shape, f_H, being the average of two bells that are space apart, will look like a two-humped camel, instead of a bell. We call the distribution of H a **mixture distribution**, with the name alluding to the fact that we mixed the two bells to ge the two-humped camel.

Another example is that of **overdispersion** in connection with Poisson models. Recall the following about the Poisson distribution family:

(a) This family is often used to model counts.

(b) For any Poisson distribution, the variance equals the mean.

In some cases in which we are modeling count data, one may try to fit a mixture of several Poisson distributions, instead of a single one. This frees us of constraint (b), as can be seen as follows:

Suppose X can equal 1,2,...,k, with probabilities $p_1, ..., p_k$ that sum to 1. Say the distribution of Y given X = i is Poisson with parameter λ_i. Then by the Law of Total Expectation,

$$
\begin{aligned}
EY &= E[E(|X)] && (5.176)\\
&= E(\lambda_X) && (5.177)\\
&= \sum_{i=1}^{k} p_i \lambda_i && (5.178)
\end{aligned}
$$

Note that in the above, the expression λ_X is a random variable, since its subscript X is random. Indeed, it is a function of X, so Equation (3.24) then applies, yielding the final equation. The random variable λ_x takes on the values $\lambda_1, ..., \lambda_k$ with probabilities $p_1, ..., p_k$, hence that final sum.

The corresponding formula for variance, (5.118), can be used to derive Var(Y).

$$\begin{aligned} Var(Y) &= E[Var(Y|X)] + Var[E(Y|X)] && (5.179) \\ &= E(\lambda_X) + Var(\lambda_X) && (5.180) \end{aligned}$$

We already evaluated the first term, in (5.176). The second term is evaluated the same way: This is the variance of a random variable that takes on the values $\lambda_1, ..., \lambda_k$ with probabilities $p_1, ..., p_k$, which is

$$\sum_{i=1}^{k} p_i(\lambda_i - \overline{\lambda})^2 \tag{5.181}$$

where

$$\overline{\lambda} = E\lambda_X = \sum_{i=1}^{k} p_i \lambda_i \tag{5.182}$$

Thus

$$EY = \overline{\lambda} \tag{5.183}$$

and

$$Var(Y) = \overline{\lambda} + \sum_{i=1}^{k} p_i(\lambda_i - \overline{\lambda})^2 \tag{5.184}$$

So, as long as the λ_i are not equal and no $p_i = 1$, we have

$$Var(Y) > EY \tag{5.185}$$

in this Poisson mixture model, in contrast to the single-Poisson case in which Var(Y) = EY. You can now see why the Poisson mixture model is called an overdispersion model.

So, if one has count data in which the variance is greater than the mean, one might try using this model.

In mixing the Poissons, there is no need to restrict to discrete X. In fact, it is not hard to derive the fact that if X has a gamma distribution with parameters r and p/(1-p) for some $0 < p < 1$, and Y given X has a Poisson distribution with mean X, then the resulting Y neatly turns out to have a negative binomial distribution.

5.11 Transform Methods (advanced topic)

We often use the idea of **transform** functions. For example, you may have seen **Laplace transforms** in a math or engineering course. The functions we will see here differ from this by just a change of variable.

Though in the form used here they involve only univariate distributions, their applications are often multivariate, as will be the case here.

5.11.1 Generating Functions

Let's start with the **generating function**. For any nonnegative-integer valued random variable V, its generating function is defined by

$$g_V(s) = E(s^V) = \sum_{i=0}^{\infty} s^i p_V(i),\ 0 \leq s \leq 1 \tag{5.186}$$

For instance, suppose N has a geometric distribution with parameter p, so that $p_N(i) = (1-p)p^{i-1}$, i = 1,2,... Then

$$g_N(s) = \sum_{i=1}^{\infty} s^i \cdot (1-p)p^{i-1} = \frac{1-p}{p} \sum_{i=1}^{\infty} s^i \cdot p^i = \frac{1-p}{p} \frac{ps}{1-ps} = \frac{(1-p)s}{1-ps} \tag{5.187}$$

Why restrict s to the interval [0,1]? The answer is that for $s > 1$ the series in (5.186) may not converge. for $0 \leq s \leq 1$, the series does converge. To see this, note that if s = 1, we just get the sum of all probabilities, which is 1.0. If a nonnegative s is less than 1, then s^i will also be less than 1, so we still have convergence.

One use of the generating function is, as its name implies, to generate the probabilities of values for the random variable in question. In other words, if you have the generating function but not the probabilities, you can obtain the probabilities from the function. Here's why: For clarify, write (5.186) as

$$g_V(s) = P(V = 0) + sP(V = 1) + s^2 P(V = 2) + ... \tag{5.188}$$

From this we see that

$$g_V(0) = P(V = 0) \tag{5.189}$$

So, we can obtain P(V = 0) from the generating function. Now differentiating (5.186) with respect to s, we have

$$\begin{aligned} g'_V(s) &= \frac{d}{ds}\left[P(V=0) + sP(V=1) + s^2P(V=2) + ...\right] \\ &= P(V=1) + 2sP(V=2) + ... \end{aligned} \tag{5.190}$$

So, we can obtain P(V = 2) from $g'_V(0)$, and in a similar manner can calculate the other probabilities from the higher derivatives.

5.11.2 Moment Generating Functions

The generating function is handy, but it is limited to discrete random variables. More generally, we can use the **moment generating function**, defined for any random variable X as

$$m_X(t) = E[e^{tX}] \tag{5.191}$$

for any t for which the expected value exists.

That last restriction is anathema to mathematicians, so they use the characteristic function,

$$\phi_X(t) = E[e^{itX}] \tag{5.192}$$

which exists for any t. However, it makes use of pesky complex numbers, so we'll stay clear of it here.

Differentiating (5.191) with respect to t, we have

$$m'_X(t) = E[Xe^{tX}] \tag{5.193}$$

We see then that

$$m'_X(0) = EX \tag{5.194}$$

So, if we just know the moment-generating function of X, we can obtain EX from it. Also,

$$m''_X(t) = E(X^2e^{tX}) \tag{5.195}$$

so

$$m''_X(0) = E(X^2) \tag{5.196}$$

In this manner, we can for various k obtain $E(X^k)$, the **k**th **moment** of X, hence the name.

5.11.3 Transforms of Sums of Independent Random Variables

Suppose X and Y are independent and their moment generating functions are defined. Let Z = X+Y. then

$$m_Z(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E(e^{tX}) \cdot E(e^{tY}) = m_X(t)m_Y(t) \tag{5.197}$$

In other words, the mgf of the sum is the product of the mgfs! This is true for other transforms, by the same reasoning.

Similarly, it's clear that the mgf of a sum of three independent variables is again the product of their mgfs, and so on.

5.11.4 Example: Network Packets

As an example, suppose say the number of packets N received on a network link in a given time period has a Poisson distribution with mean μ, i.e.

$$P(N = k) = \frac{e^{-\mu}\mu^k}{k!}, k = 0, 1, 2, 3, ... \tag{5.198}$$

5.11.4.1 Poisson Generating Function

Let's first find its generating function.

$$g_N(t) = \sum_{k=0}^{\infty} t^k \frac{e^{-\mu}\mu^k}{k!} = e^{-\mu} \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} = e^{-\mu+\mu t} \tag{5.199}$$

where we made use of the Taylor series from calculus,

$$e^u = \sum_{k=0}^{\infty} u^k/k! \tag{5.200}$$

5.11.4.2 Sums of Independent Poisson Random Variables Are Poisson Distributed

Supposed packets come in to a network node from two independent links, with counts N_1 and N_2, Poisson distributed with means μ_1 and μ_2. Let's find the distribution of $N = N_1 + N_2$, using a transform approach.

From Section 5.11.3:

$$g_N(t) = g_{N_1}(t) g_{N_2}(t) = e^{-\nu + \nu t} \tag{5.201}$$

where $\nu = \mu_1 + \mu_2$.

But the last expression in (5.201) is the generating function for a Poisson distribution too! And since there is a one-to-one correspondence between distributions and transforms, we can conclude that N has a Poisson distribution with parameter ν. We of course knew that N would have mean ν but did not know that N would have a Poisson distribution.

So: A sum of two independent Poisson variables itself has a Poisson distribution. By induction, this is also true for sums of k independent Poisson variables.

5.11.4.3 Random Number of Bits in Packets on One Link (advanced topic)

Consider just one of the two links now, and for convenience denote the number of packets on the link by N, and its mean as μ. Continue to assume that N has a Poisson distribution.

Let B denote the number of bits in a packet, with $B_1, ..., B_N$ denoting the bit counts in the N packets. We assume the B_i are independent and identically distributed. The total number of bits received during that time period is

$$T = B_1 + ... + B_N \tag{5.202}$$

Suppose the generating function of B is known to be h(s). Then what is the generating function of T?

$$\begin{aligned}
g_T(s) &= E(s^T) && (5.203)\\
&= E[E(s^T|N)] && (5.204)\\
&= E[E(s^{B_1+...+B_N}|N)] && (5.205)\\
&= E[E(s^{B_1}|N)...E(s^{B_N}|N)] && (5.206)\\
&= E[h(s)^N] && (5.207)\\
&= g_N[h(s)] && (5.208)\\
&= e^{-\mu+\mu h(s)} && (5.209)
\end{aligned}$$

Here is how these steps were made:

- From the first line to the second, we used the Theorem of Total Expectation.
- From the second to the third, we just used the definition of T.
- From the third to the fourth lines, we have used algebra plus the fact that the expected value of a product of independent random variables is the product of their individual expected values.
- From the fourth to the fifth, we used the definition of h(s).
- From the fifth to the sixth, we used the definition of g_N.
- From the sixth to the last we used the formula for the generating function for a Poisson distribution with mean μ.

We can then get all the information about T we need from this formula, such as its mean, variance, probabilities and so on, as seen previously.

5.11.5 Other Uses of Transforms

Transform techniques are used heavily in queuing analysis, including for models of computer networks. The techniques are also used extensively in modeling of hardware and software reliability.

Transforms also play key roles in much of theoretical probability, the Central Limit Theorems[7] being a good example. Here's an outline of the proof of the basic CLT, assuming the notation of Section 4.4.3.8:

First rewrite Z as

$$Z = \sum_{i=1}^{n} \frac{X_i - m}{v\sqrt{n}} \tag{5.210}$$

Then work with the characteristic function of Z:

$$\begin{aligned} c_Z(t) &= E(e^{itZ}) \quad \text{(def.)} && (5.211)\\ &= \Pi_{i=1}^{n} E[e^{it(X_i-m)/(v\sqrt{n})}] \quad \text{(indep.)} && (5.212)\\ &= \Pi_{i=1}^{n} E[e^{it(X_1-m)/(v\sqrt{n})}] \quad \text{(ident. distr.)} && (5.213)\\ &= [g(\frac{it}{\sqrt{n}})]^n && (5.214) \end{aligned}$$

[7] The plural is used here because there are many different versions, which for instance relax the condition that the summands be independent and identically distributed.

where g(s) is the characteristic function of $(X_1 - m)/v$, i.e.

$$g(s) = E[e^{is \cdot \frac{X_1 - m}{v}}] \tag{5.215}$$

Now expand (5.214) in a Taylor series around 0, and use the fact that $g'(0)$ is the expected value of $(X_1 - m)/v$, which is 0:

$$\begin{aligned} [g(\frac{t}{\sqrt{n}})]^n &= \left[1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\right]^n & (5.216) \\ &\rightarrow e^{-t^2/2} \text{ as } n \rightarrow \infty & (5.217) \end{aligned}$$

where we've also used the famous fact that $(1 - s/n)^n$ converges to e^{-s} as $n \rightarrow \infty$.

But (5.217) is the

5.12 Vector Space Interpretations (for the mathematically adventurous only)

The abstract vector space notion in linear algebra has many applications to statistics. We develop some of that material in this section.

Consider the set of all random variables associated with some "experiment," in our "notebook" sense from Section 2.2. (In more mathematical treatments, we would refer here to the set of all random variables defined on some **probability space**.) Note that some of these random variables are independent of each other, while others are not; we are simply considering the totality of all random variables that arise from our experiment.

Let $\mathcal{V}$ be the set of all such random variables having finite variance and mean 0. We can set up as a vector space. For that, we need to define a sum and a scalar product. Define the sum of any two vectors X and Y to be the random variable X+Y. For any constant c, the vector cX is the random variable cX. Note that $\mathcal{V}$ is closed under these operations, as it must be: If X and Y both have mean 0, then X+Y does too, and so on.

Define an inner product on this space:

$$(X, Y) = E(XY) = Cov(X, Y) \tag{5.218}$$

(Recall that Cov(X,Y) = E(XY) - EX EY, and that we are working with random variables that have mean 0.) Thus the norm of a vector X is

$$||X|| = (X, X)^{0.5} = \sqrt{E(X^2)} = \sqrt{Var(X)} \tag{5.219}$$

again since E(X) = 0.

5.12.1 Properties of Correlation

The famous Cauchy-Schwarz Inequality for inner products says,

$$|(X,Y)| \leq ||X||\,||Y|| \tag{5.220}$$

i.e.

$$|\rho(X,Y)| \leq 1 \tag{5.221}$$

Also, the Cauchy-Schwarz Inequality yields equality if and only if one vector is a scalar multiple of the other, i.e. Y = cX for some c. When we then translate this to random variables of nonzero means, we get Y = cX + d.

In other words, the correlation between two random variables is between -1 and 1, with equality if and only if one is an exact linear function of the other.

5.12.2 Conditional Expectation As a Projection

For a random variable X in , let $\mathcal{W}$ denote the subspace of $\mathcal{V}$ consisting of all functions h(X) with mean 0 and finite variance. (Again, note that this subspace is indeed closed under vector addition and scalar multiplication.)

Now consider any Y in $\mathcal{V}$. Recall that the *projection* of Y onto $\mathcal{W}$ is the closest vector T in $\mathcal{W}$ to Y, i.e. T minimizes $||Y - T||$. That latter quantity is

$$\Big(E[(Y-T)^2]\Big)^{0.5} \tag{5.222}$$

To find the minimizing T, consider first the minimization of

$$E[(S-c)^2] \tag{5.223}$$

with respect to constant c for some random variable S. We already solved this problem back in Section 3.59. The minimizing value is c = ES.

Getting back to (5.222), use the Law of Total Expectation to write

$$E[(Y-T)^2] = E\left(E[(Y-T)^2|X]\right) \tag{5.224}$$

From what we learned with (5.223), applied to the conditional (i.e. inner) expectation in (5.224), we see that the T which minimizes (5.224) is T = E(Y|X).

In other words, the conditional mean is a projection! Nice, but is this useful in any way? The answer is yes, in the sense that it guides the intuition. All this is related to issues of statistical prediction—here we would be predicting Y from X—and the geometry here can really guide our insight. This is not very evident without getting deeply into the prediction issue, but let's explore some of the implications of the geometry.

For example, a projection is perpendicular to the line connecting the projection to the original vector. So

$$0 = (E(Y|X), Y - E(Y|X)) = Cov[E(Y|X),\ Y - E(Y|X)] \tag{5.225}$$

This says that the prediction E(Y|X) is uncorrelated with the prediction error, Y-E(Y|X). This in turn has statistical importance. Of course, (5.225) could have been derived directly, but the geometry of the vector space intepretation is what suggested we look at the quantity in the first place. Again, the point is that the vector space view can guide our intuition.

Simlarly, the Pythagorean Theorem holds, so

$$||Y||^2 = ||E(Y|X)||^2 + ||Y - E(Y|X)||^2 \tag{5.226}$$

which means that

$$Var(Y) = Var[E(Y|X)] + Var[Y - E(Y|X)] \tag{5.227}$$

Equation (5.227) is a common theme in linear models in statistics, the decomposition of variance.

There is an equivalent form that is useful as well, derived as follows from the second term in (5.227). Since

$$E[Y - E(Y|X)] = EY - E[E(Y|X)] = EY - EY = 0 \tag{5.228}$$

we have

$$\begin{aligned} Var[Y - E(Y|X)] &= E\left[(Y - E(Y|X))^2\right] && (5.229)\\ &= E\left[Y^2 - 2Y E(Y|X) + (E(Y|X))^2\right] && (5.230) \end{aligned}$$

Now consider the middle term, $E[-2YE(Y|X)]$. Conditioning on X and using the Law of Total Expectation, we have

$$E[-2YE(Y|X)] = -2E\left[(E(Y|X))^2\right] \tag{5.231}$$

Then (5.229) becomes

$$\begin{aligned} Var[Y - E(Y|X)] &= E(Y^2) - E\left[(E(Y|X))^2\right] && (5.232)\\ &= E\left[E(Y^2|X)\right] - E\left[(E(Y|X))^2\right] && (5.233)\\ &= E\left(E(Y^2|X) - (E(Y|X))^2\right) && (5.234)\\ &= E\left[Var(Y|X)\right] && (5.235) \end{aligned}$$

the latter coming from our old friend, $Var(U) = E(U^2) - (EU)^2$, with U being Y here, under conditioning by X.

In other words, we have just derived another famous formula:

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)] \tag{5.236}$$

5.13 Proof of the Law of Total Expectation

Let's prove (5.115) for the case in which W and Y take values only in the set {1,2,3,...}. Recall that if T is an integer-value random variable and we have some function h(), then L = h(T) is another random variable, and its expected value can be calculated as[8]

$$E(L) = \sum_k h(k)P(T = k) \tag{5.237}$$

In our case here, Q is a function of W, so we find its expectation from the distribution of W:

[8] This is sometimes called The Law of the Unconscious Statistician, by nasty probability theorists who look down on statisticians. Their point is that technically $EL = \sum_k kP(L = k)$, and that (5.237) must be proven, whereas the statisticians supposedly think it's a definition.

$$\begin{aligned} E(Q) &= \sum_{i=1}^{\infty} g(i)P(W=i) \\ &= \sum_{i=1}^{\infty} E(Y|W=i)P(W=i) \\ &= \sum_{i=1}^{\infty} \left[\sum_{j=1}^{\infty} jP(Y=j|W=i) \right] P(W=i) \\ &= \sum_{j=1}^{\infty} j \sum_{i=1}^{\infty} P(Y=j|W=i)P(W=i) \\ &= \sum_{j=1}^{\infty} jP(Y=j) \\ &= E(Y) \end{aligned}$$

In other words,

$$E(Y) = E[E(Y|W)] \tag{5.238}$$

Exercises

1. Suppose the random pair (X, Y) has the density $f_{X,Y}(s,t) = 8st$ on the triangle $\{(s,t) : 0 < t < s < 1\}$.

(a) Find $\rho(X, Y)$ and $f_X(s)$.

(b) Consider the bivariate density (5.1.2.3). Find $P(X < Y/2)$.

2. Suppose packets on a network are of three types. In general, 40% of the packets are of type A, 40% have type B and 20% have type C. We observe six packets, and denote the numbers of packets of types A, B and C by X, Y and Z, respectively.

(a) Find P(X = Y = Z = 2).

(b) Find Cov(X,Y+Z).

(c) Which one of the following is the distribution of Y+Z?

(i) Binomial.

(ii) Multinomial.

(iii) Negative binomial.

(iv) Poisson.

(v) Uniform.

(vi) Exponential.

(vii) A distribution not listed above.

(viii) None; Y+Z doesn't have a distribution.

3. In the catchup game in Section 5.2.4, let V and W denote the winnings of the two players after only one turn. Find $P(V > 0.4)$.

4. Suppose X and Y are independent, each having an exponential distribution with means 1.0 and 2.0, respectively.

(a) Find $P(Y > X^2)$.

(b) Fill in the subscript for f:

$f_{\square}(t) = \int_0^t e^{-s} \cdot 0.5 e^{-0.5(t-s)} \, ds$

5. Bus lines A and B intersect at a certain transfer point, with the schedule stating that buses from both lines will arrive there at 3:00 p.m. However, they are often late, by amounts X and Y, measured in hours, for the two buses. The bivariate density is

$$f_{X,Y}(s,t) = 2 - s - t, \ 0 < s, t < 1 \tag{5.239}$$

Two friends agree to meet at the transfer point, one taking line A and the other B. Let W denote the time in minutes the person arriving on line B must wait for the friend.

(a) Show that X and Y are not independent, by evaluating $P(X \in A \text{ and } Y \in B)$, $P(X \in A)$ and $P(Y \in B)$ for some sets A and B.

(b) Find $P(W > 6)$.

(c) Find EW. Note that W is neither fully discrete nor fully continuous. We have not developed machinery for this, but the Law of Total Expectation will give you what you need here.

6. Suppose the pair (X,Y)' has a bivariate normal distribution with mean vector (0,2) and covariance matrix

$$\begin{pmatrix} 1 & 2 \\ 2 & 6 \end{pmatrix}$$

(a) Set up (but do not evaluate) the double integral for the exact value of $P(X^2 + Y^2 \leq 2.8)$.

(b) Using the matrix methods of Section 5.6, find the covariance matrix of the pair U = (X+Y,X-2Y)'. Does U have a bivariate normal distribution?

7. Show that

$$\rho(aX + b, cY + d) = \rho(X, Y) \tag{5.240}$$

for any constants a, b, c and d.

8. Suppose X are Y independent, and each has a U(0,1) distribution. Let V = X + Y.

(a) Find f_V. (Advice: It will be a "two-part function," i.e. the type we have to describe by saying something like, "The function has value 2z for z $<$ 6 and 1/z for z $>$ 6.")

(b) Verify your answer in (a) by finding EV from your answer in (a) and then using the fact that EX = EY = 0.5.

9. Suppose the following:

- In the general population of parents who have 10-year-old kids, the parent/kid weight pairs have an exact bivariate normal distribution.
- Parents' weights have mean 152.6 and standard deviation 25.0.
- Weights of kids have mean 62 and standard deviation 6.4.
- The correlation between the parents' and kids' weights is 0.4.

Use R functions (not simulation) in the following:

(a) Find the fraction of parents who weigh more than 160.

(b) Find the fraction of kids who weigh less than 56.

(c) Find the fraction of parent/child pairs in which the parent weighs more than 160 and the child weighs less than 56.

(d) Suppose a ride at an amusement park charges by weight, one cent for each pound of weight in the parent and child. State the exact distribution of the fee, and find the fraction of parent/child pairs who are charged less than $2.00.

10. Suppose X, Y and Z are "i.i.d." (independent, identically distributed) random variables, with $E(X^k)$ being denoted by ν_k, i = 1,2,3. Find Cov(XY,XZ) in terms of the ν_k.

11. Using the properties of covariance in Section 5.2.1, show that for any random variables X and Y, Cov(X+Y,X-Y) = Var(X) - Var(Y).

12. Newspapers at a certain vending machine cost 25 cents. Suppose 60% of the customers pay with quarters, 20% use two dimes and a nickel, 15% insert a dime and three nickels, and 5% deposit five nickels. When the vendor collects the money, five coins fall to the ground. Let X, Y amd Z denote the numbers of quarters, dimes and nickels among these five coins.

(a) Is the joint distribution of (X, Y, Z) a member of a parametric family presented in this chapter? If so, which one?

(b) Find $P(X = 2, Y = 2, Z = 1)$.

(c) Find $\rho(X, Y)$.

Hint: First find the proportion of quarters, among all coins deposited in this machine generally.

13. Suppose we wish to predict a random variable Y by using another random variable, X. We may consider predictors of the form $cX + d$ for constants c and d. Show that the values of c and d that minimize the mean squared prediction error, $E[(Y - cX - d)^2$ are

$$c = \frac{E(XY) - EX \cdot EY}{Var(X)} \tag{5.241}$$

$$d = \frac{E(X^2) \cdot EY - EX \cdot E(XY)}{Var(X)} \tag{5.242}$$

14. Programs A and B consist of r and s modules, respectively, of which c modules are common to both. As a simple model, assume that each module has probability p of being correct, with the modules acting independently. Let X and Y denote the numbers of correct modules in A and B, respectively. Find the correlation (X, Y) as a function of r, s, c and p.

Hint: Write $X = X_1 + ...X_r$, where X_i is 1 or 0, depending on whether module i of A is correct. Of those, let $X_1, ..., X_c$ correspond to the modules in common to A and B. Similarly, write $Y = Y_1 + ...Y_s$, for the modules in B, again having the first c of them correspond to the modules in common. Do the same for B, and for the set of common modules.

15. Show that if random variables U and V are independent,

$$Var(UV) = E(U^2) \cdot Var(V) + Var(U) \cdot (EV)^2 \tag{5.243}$$

16. Use transform methods to derive some properties of the Poisson family:

(a) Show that for any Poisson random variable, its mean and variance are equal.

(b) Suppose X and Y are independent random variables, each having a Poisson distribution. Show that $Z = X + Y$ again has a Poisson distribution.

17. Suppose one keeps rolling a die. Let S_n denote the total number of dots after n rolls, mod 8, and let T be the number of rolls needed for the event $S_n = 0$ to occur. Find $E(T)$, using an approach like that in the "trapped miner" example in Section 5.7.5.

18. In our ordinary coins which we use every day, each one has a slightly different probability of heads, which we'll call H. Say H has the distribution $N(0.5, 0.03^2)$. We choose a coin from a batch at random, then toss it 10 times. Let N be the number of heads we get. Find $Var(N)$.

19. Jack and Jill play a dice game, in which one wins \$1 per dot. There are three dice, die A, die B and die C. Jill always rolls dice A and B. Jack always rolls just die C, but he also gets credit for 90% of die B. For instance, say in a particular roll A, B and C are 3, 1 and 6, respectively. Then Jill would win \$4 and Jack would get \$6.90. Let X and Y be Jill's and Jack's total winnings after 100 rolls. Use the Central Limit Theorem to find the approximate values of $P(X > 650, Y < 660)$ and $P(Y > 1.06X)$.

Hints: This will follow a similar pattern to the dice game in Section 5.8.2.3, which we win \$5 for one dot, and \$2 for two or three dots. Remember, in that example, the key was that we noticed that the pair (X, Y) was a sum of random pairs. That meant that (X, Y) had an approximate bivariate normal distribution, so we could find probabilities if we had the mean vector and covariance matrix of (X, Y). Thus we needed to find $EX, EY, Var(X), Var(Y)$ and $Cov(X, Y)$. We used the various properties of $E(), Var()$ and $Cov()$ to get those quantities.

You will do the same thing here. Write $X = U_1 + ... + U_{100}$, where U_i is Jill's winnings on the ith roll. Write Y as a similar sum of V_i. You probably will find it helpful to define A_i, B_i and C_i as the numbers of dots appearing on dice A, B and C on the ith roll. Then find EX etc. Again, make sure to utilize the various properties for $E(), Var()$ and $Cov()$.

20. Suppose the number N of bugs in a certain number of lines of code has a Poisson distribution, with parameter L, where L varies from one programmer to another. Show that Var(N) = EL + Var(L).

21. This problem arises from the analysis of random graphs, which for concreteness we will treat here as social networks such as Facebook.

In the model here, each vertex in the graph has N friends, N being a random variable with the same distribution at every vertex. One thinks of each vertex as generating its links, unterminated, i.e. not tied yet to a second vertex. Then the unterminated links of a vertex pair off at random with those of other vertices. (Those that fail will just pair in self loops, but we'll ignore that.)

Let M denote the number of friends a friend of mine has. That is, start at a vertex A, and follow a link from A to another vertex, say B. M is the number of friends B has (we'll include A in this number).

(a) Since an unterminated link from A is more likely to pair up with a vertex that has a lot of links, a key assumption is that P(M = k) = ck P(N = k) for some constant c. Fill in the blank: This is an example of the setting we studied called _____.

(b) Show the following relation of generating functions: $g_M(s) = g'_N(s)/EN$.

22. Suppose Type 1 batteries have exponentially distributed lifetimes with mean 2.0 hours, while Type 2 battery lifetimes are exponentially distributed with mean 1.5. We have a large box containing a mixture of the two types of batteries, in proportions q and 1-q. We reach into the box, choose a battery at random, then use it. Let Y be the lifetime of the battery we choose. Use the Law of Total Variance, (5.118), to find $Var(Y)$.

23. Consider random variables X_1 and X_2, for which $Var(X_i) = 1.0$ for i = 1,2, and $Cov(X_1, X_2) = 0.5$. Find $Var(X_1 + X_2)$.

24. Suppose we have random variables X and Y, and define the new random variable Z = 8Y. Then which of the following is correct? (i) $\rho(X, Z) = \rho(X, Y)$. (ii) $\rho(X, Z) = 0$. (iii) $\rho(Y, Z) = 0$. (iv) $\rho(X, Z) = 8\rho(X, Y)$. (v) $\rho(X, Z) = \frac{1}{8}\rho(X, Y)$. (vi) There is no special relationship.

25. Suppose $f_X(t) = 2t$ for $0 < t < 1$ and the density is 0 elsewhere.

(a) Find $h_X(0.5)$.

(b) Which statement concerning this distribution is correct? (i) IFR. (ii) DFR. (iii) U-shaped failure rate. (iv) Sinusoidal failure rate. (v) Failure rate is undefined for $t > 0.5$.

26. Consider the coin game in Section 3.14.1. Find $F_{X_3,Y_3}(0, 0)$.

27. In the backup battery example in Section 5.5.6, find Var(W).

28. Consider the "8st" density example in Section 5.1.2.3. Find $P(Y > X^2)$.

29. What will be the (approximate) output of the following R code?

```
s <- 0
s2 <- 0
for (rep in 1:10000) {
   z3 <- rnorm(3)  # generate 3 N(0,1) random variates
   tot <- sum(z3^2)  # sum of the squares of the 3 variates
   s <- s + tot
   s2 <- s2 + tot^2
}
m <- s/10000
print(m)
print(s2/10000 - m^2)
```

30. Suppose the random vector $X = (X_1, X_2, X_3)'$ has mean $(2.0, 3.0, 8.2)'$ and covariance matrix

$$\begin{pmatrix} 1 & 0.4 & -0.2 \\ & 1 & 0.25 \\ & & 3 \end{pmatrix} \tag{5.244}$$

(a) Fill in the three missing entries.

(b) Find $Cov(X_1, X_3)$.

(c) Find $\rho(X_2, X_3)$.

(d) Find $Var(X_3)$.

(e) Find the covariance matrix of $(X_1 + X_2, X_2 + X_3)'$.

(f) If in addition we know that X_1 has a normal distribution, find $P(1 < X_1 < 2.5)$, in terms of $\Phi()$.

(g) Consider the random variable $W = X_1 + X_2$. Which of the following is true? (i) $Var(W) = Var(X_1 + X_2)$. (ii) $Var(W) > Var(X_1 + X_2)$. (iii) $Var(W) < Var(X_1 + X_2)$. (iv) In order to determine which of the two variances is the larger one, we would need to know whether the variables X_i have a multivariate normal distribution. (v) $Var(X_1 + X_2)$ doesn't exist.

31. What is the (approximate) output of this R code:

```
count <- 0
for (i in 1:10000) {
   count1 <- 0
   count2 <- 0
   count3 <- 0
```

```
   for (j in 1:20) {
      x <- runif(1)
      if (x < 0.2) {
         count1 <- count1 + 1
      } else if (x < 0.6) count2 <- count2 + 1 else
             count3 <- count3 + 1
   }
   if (count1 == 9 && count2 == 2 && count3 == 9) count <- count + 1
}
cat(count/10000)
```

32. Let X denote the number we obtain when we roll a single die once. Let $G_X(s)$ denote the generating function of X.

(a) Find $G_X(s)$.

(b) Suppose we roll the die 5 times, and let T denote the total number of dots we get from the 5 rolls. Find $G_T(s)$.

33. Derive (5.23). Hint: A constant, q here, is a random variable, trivially, with 0 variance.

34. Consider a three-card hand drawn from a 52-card deck. Let X and Y denote the number of hearts and diamonds, respectively. Find $\rho(X, Y)$.

35. Consider the lightbulb example in Section 4.4.6.1. Use the "mailing tubes" on Var() and Cov() to find $\rho(X_1, T_2)$.

36. Consider this model of disk seeks. For simplicity, we'll assume a very tiny number of tracks, 3. Let X_1 and X_2 denote the track numbers of two successive disk requests. Each has a uniform distribution on {1,2,3}. But given $X_1 = i$, then $X_2 = i$ with probability 0.4, with X_2 being j with probability 0.3, $j \neq i$. (Convince yourself that these last two sentences are consistent with each other.) Find the following:

(a) $P(|X_1 - X_2| \leq 1)$

(b) $E(|X_1 - X_2|)$

(c) $F_{X_1,X_2}(2, 2)$

37. Use the convolution formula (5.62) to derive (4.65) for the case r = 2. Explain your steps carefully!

38. The book, *Last Man Standing*, author D. McDonald writes the following about the practice of combining many mortgage loans into a single package sold to investors:

> Even if every single [loan] in the [package] had a 30 percent risk of default, the thinking went, the odds that most of them would default at once were arguably infinitesimal...What [this argument] missed was the auto-synchronous relationship of many loans...[If several of them] are all

> mortgage for houses sitting next to each other on a beach...one strong hurricane and the [loan package] would be decimated.

Fill in the blank with a term from this book: The author is referring to an unwarranted assumption of _____.

39. Find the following quantities for the dice example in Section 5.5.1:

(a) Cov(X,2S)

(b) Cov(X,S+Y)

(c) Cov(X+2Y,3X-Y)

(d) $p_{X,S}(3,8)$

40. Consider the computer worm example in Section 5.5.8. Let R denote the time it takes to go from state 1 to state 3. Find $f_R(v)$. (Leave your answer in integral form.)

41. Suppose (X,Y) has a bivariate normal distribution, with EX = EY = 0, Var(X) = Var(Y) = 1, and $\rho(X,Y) = 0.2$. Find the following, in integral forms:

(a) $E(X^2 + XY^{0.5})$

(b) $P(Y > 0.5X)$

(c) $F_{X,Y}(0.6, 0.2)$

42. Suppose X_i, i = 1,2,3,4,5 are independent and each have mean 0 and variance 1. Let $Y_i = X_{i+1} - X_i$, i = 1,2,3,4. Using the material in Section 5.6, find the covariance matrix of $Y = (Y_1, Y_2, Y_3, Y_4)$.

Chapter 6

Describing "Failure"

In addition to density functions, another useful description of a distribution is its **hazard function**. Again think of the lifetimes of light bulbs, not necessarily assuming an exponential distribution. Intuitively, the hazard function states the likelihood of a bulb failing in the next short interval of time, given that it has lasted up to now. To understand this, let's first talk about a certain property of the exponential distribution family.

6.1 Memoryless Property

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**.

6.1.1 Derivation and Intuition

What the term *memoryless* means is that for positive t and u

$$P(W > t + u|W > t) = P(W > u) \tag{6.1}$$

Let's derive this:

$$
\begin{aligned}
P(W > t+u | W > t) &= \frac{P(W > t+u \text{ and } W > t)}{P(W > t)} && (6.2)\\
&= \frac{P(W > t+u)}{P(W > t)} && (6.3)\\
&= \frac{\int_{t+u}^{\infty} \lambda e^{-\lambda s}\, ds}{\int_{t}^{\infty} \lambda e^{-\lambda s}\, ds} && (6.4)\\
&= e^{-\lambda u} && (6.5)\\
&= P(W > u) && (6.6)
\end{aligned}
$$

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 (6.1) 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.

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 (6.1), we would have

$$R(t+u)]/R(t) = R(u) \tag{6.7}$$

or

$$R(t+u) = R(t)R(u) \tag{6.8}$$

Differentiating both sides with respect to t, we'd have

$$R\prime(t+u) = R\prime(t)R(u) \tag{6.9}$$

Setting t to 0, this would say

$$R\prime(u) = R\prime(0)R(u) \tag{6.10}$$

This is a well-known differential equation, whose solution is

$$R(u) = e^{-cu} \tag{6.11}$$

which is exactly 1 minus the cdf for an exponentially distributed random variable.

6.1.2 Continuous-Time Markov Chains

The memorylessness of exponential distributions implies that a Poisson process N(t) also has a "time starts over" property: Recall our example in Section 4.4.6.1 in which N(t) was the number of light bulb burnouts up to time t. The memorylessness property means that if we start counting afresh from time, say z, then the numbers of burnouts after time z, i.e. Q(u) = N(z+u) - N(z), also is a Poisson process. In other words, Q(u) has a Poisson distribution with parameter λ. Moreover, Q(u) is independent of N(t) for any $t < z$.

All this should remind you of Markov chains, which we introduced in Section 3.15—and it should. **Continuous time** Markov chains are defined in the same way as the discrete-time ones in Section 3.15, but with the process staying in each state for a random amount of time. From the considerations here, you can now see that time must have an exponential distribution. This will be discussed at length in Chapter 11.

6.2 Hazard Functions

6.2.1 Basic Concepts

Suppose the lifetimes of light bulbs L were discrete. Suppose a particular bulb has already lasted 80 hours. The probability of it failing in the next hour would be

$$P(L = 81|L > 80) = \frac{P(L = 81 \text{ and } L > 80)}{P(L > 80)} = \frac{P(L = 81)}{P(L > 80)} = \frac{p_L(81)}{1 - F_L(80)} \tag{6.12}$$

In general, for discrete L, we define its **hazard function** as

$$h_L(i) = \frac{p_L(i)}{1 - F_L(i-1)} \tag{6.13}$$

By analogy, for continuous L we define

$$h_L(t) = \frac{f_L(t)}{1 - F_L(t)} \tag{6.14}$$

Again, the interpretation is that $h_L(t)$ is the likelihood of the item failing very soon after t, given that it has lasted t amount of time.

Note carefully that the word "failure" here should not be taken literally. In our Davis railroad crossing example above, "failure" means that the train ends—a "failure" which those of us who are waiting will welcome!

Since we know that exponentially distributed random variables are memoryless, we would expect intuitively that their hazard functions are constant. We can verify this by evaluating (6.14) for an exponential density with parameter λ; sure enough, the hazard function is constant, with value λ.

The reader should verify that in contrast to an exponential distribution's constant failure rate, a uniform distribution has an increasing failure rate (IFR). Some distributions have decreasing failure rates, while most have non-monotone rates.

Hazard function models have been used extensively in software testing. Here "failure" is the discovery of a bug, and with quantities of interest include the mean time until the next bug is discovered, and the total number of bugs.

People have what is called a "bathtub-shaped" hazard function. It is high near 0 (reflecting infant mortality) and after, say, 70, but is low and rather flat in between.

You may have noticed that the right-hand side of (6.14) is the derivative of $-ln[1 - F_L(t)]$. Therefore

$$\int_0^t h_L(s)\, ds = -\ln[1 - F_L(t)] \tag{6.15}$$

so that

$$1 - F_L(t) = e^{-\int_0^t h_L(s)\, ds} \tag{6.16}$$

and thus[1]

$$f_L(t) = h_L(t) e^{-\int_0^t h_L(s)\, ds} \tag{6.17}$$

In other words, just as we can find the hazard function knowing the density, we can also go in the reverse direction. This establishes that there is a one-to-one correspondence between densities and hazard functions.

This may guide our choice of parametric family for modeling some random variable. We may not only have a good idea of what general shape the density takes on, but may also have an idea of what the hazard function looks like. These two pieces of information can help guide us in our choice of model.

6.2.2 Example: Software Reliability Models

Hazard function models have been used successfully to model the "arrivals" (i.e. discoveries) of bugs in software. Questions that arise are, for instance, "When are we ready to ship?", meaning when can we believe with some confidence that most bugs have been found?

Typically one collects data on bug discoveries from a number of projects of similar complexity, and estimates the hazard function from that data. Some investigations, such as Ohishia *et al*, Gompertz Software Reliability Model: Estimation Algorithm and Empirical Validation, *Journal of Systems and Software*, 82, 3, 2009, 535-543.

See *Accurate Software Reliability Estimation*, by Jason Allen Denton, Dept. of Computer Science, Colorado State University, 1999, and the many references therein.

6.3 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

[1] Recall that the derivative of the integral of a function is the original function!

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.*

6.3.1 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.

6.3.2 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 back, 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.

$$p_X(2) = p_X(6) = 0.1 \tag{6.18}$$

Intuitively, one would then reason that

$$p_Y(6) = 3p_Y(2) \tag{6.19}$$

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

$$p_Y(j) \propto jp_X(j) \tag{6.20}$$

where $\propto$ is the "is proportional to" symbol. Thus

$$p_Y(j) = cjp_X(j) \tag{6.21}$$

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

$$1 = \sum_j p_Y(j) = \sum_j cjp_X(j) \tag{6.22}$$

That last term is c E(X)! So, c = 1/EX, and

$$p_Y(j) = \frac{1}{EX} \cdot jp_X(j) \tag{6.23}$$

The continuous analog of (6.23) is

$$f_Y(t) = \frac{1}{EX} \cdot tf_X(t) \tag{6.24}$$

So, for our bus example, in which $f_X(t) = 0.1e^{-0.1t},\ t > 0$ and EX = 10,

$$f_Y(t) = 0.01te^{-0.1t} \tag{6.25}$$

You may recognize this as an Erlang density with r = 2 and $\lambda = 0.1$. That distribution does indeed have mean 20.

6.4 Residual-Life Distribution

In the bus-paradox example, if we had been working with light bulbs instead of buses, the analog of the time we wait for the next bus would be the remaining lifetime of the current light bulb. The time from a fixed time point t until the next bulb replacement, is known as the **residual life**. (Another name for it is the **forward recurrence time.**)

Our aim here is to derive the distribution of renewal times. To do this, let's first bring in some terminology from **renewal theory**.

6.4.1 Renewal Theory

Recall the light bulb example of Section 4.4.6.1. Every time a light bulb burns out, we immediately replace it with a new one. The time of the rth replacement is denoted by T_r, and satisfies the relation

$$N(t) = \max\{k : T_k \leq t\} \tag{6.26}$$

where N(t) is the number of replacements that have occurred by time t and X_i is the lifetime of the ith bulb. The random variables $X_1, X_2, ...$ are assumed independent and identically distributed (i.i.d.); we will NOT assume that their common distribution is exponential, though.

Note that for each t > 0, N(t) is a random variable, and since we have a collection of random variables indexed by t. This collection is called a **renewal process**, the name being motivated by the idea of "renewals" occurring when light bulbs burn out. We say that N(t) is the number of renewals by time t.

In the bus paradox example, we can think of bus arrivals as renewals too, with the interbus times sbeing analogous to the light bulb lifetimes, and with N(t) being the number of buses that have arrived by time t.

Note the following for general renewal processes:

Duality Between "Lifetime Domain" and "Counts Domain":

A very important property of renewal processes is that

$$N(t) \geq k \text{ if and only if } T_k \leq t \tag{6.27}$$

This is just a formal mathematical of common sense: There have been at least k renewals by now if and only if the k^{th} renewal has already occurred! But it is a very important device in renewal analysis.

Equation (6.27) might be described as relating the "counts domain" (left-hand side of the equation) to the "lifetimes domain" (right-hand side).

There is a very rich theory of renewal processes, but let's move on to our goal of finding the distribution of residual life.

6.4.2 Intuitive Derivation of Residual Life for the Continuous Case

Here is a derivation for the case of continuous X_i. For concreteness think of the bus case, but the derivation is general.

Denote by V the length of the interbus arrival that we happen to hit when we arrive at the bus stop, and let D denote the residual life, i.e. the time until the next bus. The key point is that, given V, D is uniformly

distributed on (0,V). To see this, think of the stick example. If the stick that we happen to touch first has length V, the point on which we touched it could be anywhere from one end to the other with equal likelihood. So,

$$f_{D|V}(s,t) = \frac{1}{t}, \ \ 0 < s < t \tag{6.28}$$

Thus (5.111) yields

$$f_{D,V}(s,t) = \frac{1}{t} \cdot f_V(t), \ \ 0 < s < t \tag{6.29}$$

Then (5.16) shows

$$\begin{aligned} f_D(s) &= \int_s^\infty \frac{1}{t} \cdot f_V(t) \, dt & (6.30) \\ &= \int_s^\infty \frac{1}{EX} \cdot f_X(t) \, dt & (6.31) \\ &= \frac{1 - F_X(s)}{EX} & (6.32) \end{aligned}$$

This is a classic result, of central importance and usefulness, as seen in our upcoming examples later in this section.[2]

It should be noted that all of this assume a "long-run" situation. In our bus example, for instance, it implicitly assumes that when we arrive at the bus stop at 5:00, the buses have been running for quite a while. To state this more precisely, let's let D depend on t: D(t) will be the residual life at time t, e.g. the time we must wait for the next bus if we arrive at the stop at time t. Then (6.30) is really the limiting density of $f_{D(t)}$ as $t \to \infty$.

6.4.3 Age Distribution

Analogous to the residual lifetime D(t), let A(t) denote the **age** (sometimes called the **backward recurrence time**) of the current light bulb, i.e. the length of time it has been in service. (In the bus-paradox example,

[2] If you are wondering about that first equality in (6.30), it is basically a continuous analog of

$$P(A) = P(A \text{ and } B_1 \text{ or } A \text{ and } B_2 \text{ or ...}) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + ...)$$

for disjoint events $B_1, B_2, ...$ This is stated more precisely in Section 5.7.3.

A(t) would be the time which has elapsed since the last arrival of a bus, to the current time t.) Using an approach similar to that taken above, one can show that

$$\lim_{t\to\infty} f_{A(t)}(w) = \frac{1 - F_L(w)}{E(L)} \tag{6.33}$$

In other words, A(t) has the same long-run distribution as D(t)!

Here is a derivation for the case in which the X_i are discrete. (We'll call the L_i here, with L being the generic random variable.) Remember, our fixed observation point t is assumed large, so that the system is in steady-state. Let W denote the lifetime so far for the current bulb. Say we have a new bulb at time 52. Then W is 0 at that time. If the total lifetime turns out to be, say, 12, then W will be 0 again at time 64.

Then we have a Markov chain in which our state at any time is the value of W. In fact, the transition probabilities for this chain are the values of the hazard function of L:

First note that when we are in state i, i.e. $W = i$, we know that the current bulb's lifetime is at least i+1. If its lifetime is exactly i+1, our next state will be 0. So,

$$p_{i,0} = P(L = i + 1 | L > i) = \frac{p_L(i + 1)}{1 - F_L(i)} \tag{6.34}$$

$$p_{i,i+1} = \frac{1 - F_L(i + 1)}{1 - F_L(i)} \tag{6.35}$$

Define

$$q_i = \frac{1 - F_L(i + 1)}{1 - F_L(i)} \tag{6.36}$$

and write

$$\pi_{i+1} = \pi_i q_i \tag{6.37}$$

Applying (6.37) recursively, we have

$$\pi_{i+1} = \pi_0 q_i q_{i-1}) \cdots q_0 \tag{6.38}$$

But the right-hand side of (6.38) telescopes down to

$$\pi_{i+1} = \pi_0 [1 - F_L(i + 1)] \tag{6.39}$$

Then

$$1 = \sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} [1 - F_L(i)] = \pi_0 E(L) \tag{6.40}$$

Thus

$$\pi_i = \frac{1 - F_L(i+1)}{EL} \tag{6.41}$$

in analogy to (6.33).

6.4.4 Mean of the Residual and Age Distributions

Taking the expected value of (6.30) or (6.33), we get a double integral. Reversing the order of integration, we find that the mean residual life or age is given by

$$\frac{E(L^2)}{2EL} \tag{6.42}$$

6.4.5 Example: Estimating Web Page Modification Rates

My paper, Estimation of Internet File-Access/Modification Rates, *ACM Transactions on Modeling and Computer Simulation*, 2005, 15, 3, 233-253, concerns the following problem.

Suppose we are interested in the rate of modfication of a file in some FTP repository on the Web. We have a spider visit the site at regular intervals. At each visit, the spider records the time of last modification to the site. We do not observe how MANY times the site was modified. The problem then is how to estimate the modification rate from the last-modification time data that we do have.

I assumed that the modifications follow a renewal process. Then the difference between the spider visit time and the time of last modfication is equal to the age A(t). I then applied a lot of renewal theory to develop statistical estimators for the modfication rate.

6.4.6 Example: Disk File Model

Suppose a disk will store backup files. We place the first file in the first track on the disk, then the second file right after the first in the same track, etc. Occasionally we will run out of room on a track, and the file we

are placing at the time must be split between this track and the next. Suppose the amount of room X taken up by a file (a continuous random variable in this model) is uniformly distributed between 0 and 3 tracks.

Some tracks will contain data from only one file. (The file may extend onto other tracks as well.) Let's find the long-run proportion of tracks which have this property.

Think of the disk as consisting of a Very Long Line, with the end of one track being followed immediately by the beginning of the next track. The points at which files begin then form a renewal process, with "time" being distance along the Very Long Line. If we observe the disk at the end of the k^{th} track, this is observing at "time" k. That track consists entirely of one file if and only if the "age" A of the current file—i.e. the distance back to the beginning of that file—is greater than 1.0.

Then from Equation (6.33), we have

$$f_A(w) = \frac{1 - \frac{w}{3}}{1.5} = \frac{2}{3} - \frac{2}{9}w \tag{6.43}$$

Then

$$P(A > 1) = \int_1^3 \left(\frac{2}{3} - \frac{2}{9}w \right) \, dw = \frac{4}{9} \tag{6.44}$$

6.4.7 Example: Memory Paging Model

(Adapted from *Probabiility and Statistics, with Reliability, Queuing and Computer Science Applicatiions*, by K.S. Trivedi, Prentice-Hall, 1982 and 2002.)

Consider a computer with an address space consisting of n pages, and a program which generates a sequence of memory references with addresses (page numbers) D_1, D_2, ... In this simple model, the D_i are assumed to be i.i.d. integer-valued random variables.

For each page i, let T_{ij} denote the time at which the j^{th} reference to page i occurs. Then for each fixed i, the T_{ij} form a renewal process, and thus all the theory we have developed here applies.[3] Let F_i be the cumulative distribution function for the interrenewal distribution, i.e. $F_i(m) = P(L_{ij} \leq m)$, where $L_{ij} = T_{ij} - T_{i,j-1}$ for m = 0, 1, 2, ...

Let $W(t, \tau)$ denote the working set at time t, i.e. the collection of page numbers of pages accessed during the time $(t - \tau, t)$, and let $S(t, \tau)$ denote the size of that set. We are interested in finding the value of

$$s(\tau) = \lim_{t \to \infty} E[S(t, \tau)] \tag{6.45}$$

[3]Note, though, tht all random variables here are discrete, not continuous.

Since the definition of the working set involves looking backward τ amount of time from time t, a good place to look for an approach to finding $s(\tau)$ might be to use the limiting distribution of backward-recurrence time, given by Equation (6.41).

Accordingly, let $A_i(t)$ be the age at time t for page i. Then

Page i is in the working set if and only if it has been accessed after time $t - \tau$., i.e. $A_i(t) < \tau$.

Thus, using (6.41) and letting 1_i be 1 or 0 according to whether or not $A_i(t) < \tau$, we have that

$$\begin{aligned} s(\tau) &= \lim_{t \to \infty} E(\sum_{i=1}^{n} 1_i) \\ &= \lim_{t \to \infty} \sum_{i=1}^{n} P(A_i(t) < \tau) \\ &= \sum_{i=1}^{n} \sum_{j=0}^{\tau-1} \frac{1 - F_i(j)}{E(L_i)} \end{aligned} \tag{6.46}$$

Exercises

1. Use R to plot the hazard functions for the gamma distributions plotted in Figure 4.2, plus the case r = 0.5. Comment on the implications for trains at 8th and J Streets in Davis.

2. Consider the "random bucket" example in Section 6.3. Suppose bucket diameter D, measured in meters, has a uniform distribution on (1,2). Let W denote the diameter of the bucket in which the tossed ball lands.

(a) Find the density, mean and variance of W, and also $P(W > 1.5)$

(b) Write an R function that will generate random variates having the distribution of W.

3. In Section 6.1, we showed that the exponential distribution is memoryless. In fact, it is the only continuous distribution with that property. Show that the U(0,1) distribution does NOT have that property. To do this, evaluate both sides of (6.1).

4. Suppose $f_X(t) = 1/t^2$ on $(1, \infty)$, 0 elsewhere. Find $h_X(2.0)$

5. Consider the three-sided die on page 29. Find the hazard function $h_V(t)$, where V is the number of dots obtained on one roll (1, 2 or 3).

Chapter 7

Introduction to Statistical Inference

Consider the following problems:

- Suppose you buy a ticket for a raffle, and get ticket number 68. Two of your friends bought tickets too, getting numbers 46 and 79. Let c be the total number of tickets sold. You don't know the value of c, but hope it's small, so you have a better chance of winning. How can you estimate the value of c, from the data, 68, 46 and 79?
- It's presidential election time. A poll says that 56% of the voters polled support candidate X, with a margin of error of 2%. The poll was based on a sample of 1200 people. How can a sample of 1200 people out of more than 100 million voters have a margin of error that small? And what does the term *margin of error* really mean, anyway?
- A satellite detects a bright spot in a forest. Is it a fire? How can we design the software on the satellite to estimate the probability that this is a fire?

If you think that statistics is nothing more than adding up columns of numbers and plugging into formulas, you are badly mistaken. Actually, statistics is an application of probability theory. We employ probabilistic models for the behavior of our sample data, and *infer* from the data accordingly—hence the name, **statistical inference**.

Arguably the most powerful use of statistics is prediction. This has applications from medicine to marketing to movie animation. We will study prediction in Chapter 10.

7.1 Sampling Distributions

We first will set up some infrastructure, which will be used heavily throughout the next few chapters.

7.1.1 Random Samples

Definition 25 *Random variables $X_1, X_2, X_3, ...$ are said to be* **i.i.d.** *if they are independent and identically distributed. The latter term means that p_{X_i} or f_{X_i} is the same for all i.*

For i.i.d. $X_1, X_2, X_3, ...$, we often use X to represent a generic random variable having the common distribution of the X_i.

Definition 26 *We say that $X_1, X_2, X_3, ..., X_n$ is a* **random sample** *of size n from a population if the X_i are i.i.d. and their common distribution is that of the population.*

If the sampled population is finite,[1] then a random sample must be drawn in this manner. Say there are k entities in the population, e.g. k people, with values $v_1, ..., v_k$. If we are interested in people's heights, for instance, then $v_1, ..., v_k$ would be the heights of all people in our population. Then a random sample is drawn this way:

(a) The sampling is done with replacement.

(b) Each X_i is drawn from $v_1, ..., v_k$, with each v_j having probability $\frac{1}{k}$ of being drawn.

Condition (a) makes the X_i independent, while (b) makes them identically distributed.

If sampling is done without replacement, we call the data a **simple random sample**. Note how this implies lack of independence of the X_i. If for instance $X_1 = v_3$, then we know that no other X_i has that value, contradicting independence; if the X_i were independent, knowledge of one should not give us knowledge concerning others.

But we assume true random sampling from here onward.

Note most carefully that *each X_i has the same distribution as the population*. If for instance a third of the population, i.e. a third of the v_j, are less than 28, then $P(X_i < 28)$ will be 1/3. This point is easy to see, but keep it in mind at all times, as it will arise again and again.

We will often make statements like, "Let X be distributed according to the population." This simply means that $P(X = v_j) = \frac{1}{k}$, j = 1,...,k.

What about drawing from an infinite population? This may sound odd at first, but it relates to the fact, noted at the outset of Chapter 4, that although continuous random variables don't really exist, they often make a good approximation. In our human height example above, for instance, heights do tend to follow a bell-shaped curve which which is well-approximated by a normal distributiion.

[1] You might wonder how it could be infinite. This will be discussed shortly.

In this case, each X_i is modeled as having a continuum of possible values, corresponding to a theoretically infinite population. Each X_i then has the same density as the population density.

7.1.2 The Sample Mean—a Random Variable

A large part of this chapter will concern the **sample mean**,

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} \tag{7.1}$$

Since $X_1, X_2, X_3, ..., X_n$ are random variables, $\overline{X}$ is a random variable too.

Make absolutely sure to distinguish between the sample mean and the population mean.

The point that $\overline{X}$ is a random variable is another simple yet crucial concept. Let's illustrate it with a tiny example. Suppose we have a population of three people, with heights 69, 72 and 70, and we draw a random sample of size 2. Here $\overline{X}$ can take on six values:

$$\frac{69+69}{2} = 69, \frac{69+72}{2} = 70.5, \frac{69+70}{2} = 69.5, \frac{70+70}{2} = 70, \frac{70+72}{2} = 71, \frac{72+72}{2} = 72 \tag{7.2}$$

The probabilities of these values are 1/9, 2/9, 2/9, 1/9, 2/9 and 1/9, respectively. So,

$$p_{\overline{X}}(69) = \frac{1}{9},\ \ p_{\overline{X}}(70.5) = \frac{2}{9},\ \ p_{\overline{X}}(69.5) = \frac{2}{9},\ \ p_{\overline{X}}(70) = \frac{1}{9},\ \ p_{\overline{X}}(71) = \frac{2}{9},\ \ p_{\overline{X}}(72) = \frac{1}{9} \tag{7.3}$$

Viewing it in "notebook" terms, we might have, in the first three lines:

notebook line	X_1	X_2	$\overline{X}$
1	70	70	70
2	69	70	69.5
3	72	70	71

Again, the point is that all of X_1, X_2 and $\overline{X}$ are random variables.

Now, returning to the case of general n and our sample $X_1, ..., X_n$, since $\overline{X}$ is a random variable, we can ask about its expected value and variance.

Let μ denote the population mean. Remember, each X_i is distributed as is the population, so $EX_i = \mu$.

This then implies that the mean of $\overline{X}$ is also μ. Here's why:

$$
\begin{aligned}
E(\overline{X}) &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \quad \text{(def. of } \overline{X}\text{)} && (7.4)\\
&= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) \quad \text{(for const. c, } E(cU) = cEU) && (7.5)\\
&= \frac{1}{n}\sum_{i=1}^{n} EX_i \quad (E[U+V] = EU + EV) && (7.6)\\
&= \frac{1}{n}n\mu \quad (EX_i = \mu) && (7.7)\\
&= \mu && (7.8)
\end{aligned}
$$

$$
\begin{aligned}
Var(\overline{X}) &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] && (7.9)\\
&= \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) \quad \text{(for const. c, , } Var[cU] = c^2 Var[U]) && (7.10)\\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \quad \text{(for U,V indep., , } Var[U+V] = Var[U] + Var[V]) && (7.11)\\
&= \frac{1}{n^2}n\sigma^2 && (7.12)\\
&= \frac{1}{n}\sigma^2 && (7.13)
\end{aligned}
$$

The Central Limit Theorem tells us that the numerator in (7.1) has an approximate normal distribution. So:

Approximate distribution of (centered and scaled) $\overline{X}$:

The quantity

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{7.14}$$

has an approximately N(0,1) distribution.

Make sure you understand why it is the "N" that is approximate here, not the 0 or 1.

7.1.3 The Sample Variance—Another Random Variable

Later we will be using the sample mean $\overline{X}$, a function of the X_i, to estimate the population mean μ. What other function of the X_i can we use to estimate the population variance σ^2?

Let X denote a generic random variable having the distribution of the X_i, which, note again, is the distribution of the population. Because of that property, we have

$$Var(X) = \sigma^2 \quad (\sigma^2 \text{ is the population variance}) \tag{7.15}$$

Recall that by definition

$$Var(X) = E[(X - EX)^2] \tag{7.16}$$

Let's estimate $Var(X) = \sigma^2$ by taking sample analogs in (7.16). Here are the correspondences:

pop. entity	samp. entity
EX	$\overline{X}$
X	X_i
E[]	$\frac{1}{n}\sum_{i=1}^{n}$

The sample analog of μ is $\overline{X}$. What about the sample analog of the "E()"? Well, since E() averaging over the whole population of Xs, the sample analog is to average over the sample. So, our sample analog of (7.16) is

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{7.17}$$

In other words, just as it is natural to estimate the population mean of X by its sample mean, the same holds for Var(X):

> The population variance of X is the mean squared distance from X to its population mean, as X ranges over all of the population. Therefore it is natural to estimate Var(X) by the average squared distance of X from its sample mean, among our sample values X_i, shown in (7.17).[2]

We use s^2 as our symbol for this estimate of population variance.[3] It should be noted that it is common to divide by n-1 instead of by n in (7.17). Though we will not take that approach here, it will be discussed in Section 8.2.2.

[2] Note the similarity to (3.29).

[3] Though I try to stick to the convention of using only capital letters to denote random variables, it is conventional to use lower case in this instance.

By the way, it can be shown that (7.17) is equal to

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \tag{7.18}$$

This is a handy way to calculate s^2, though it is subject to more roundoff error. Note that (7.18) is a sample analog of (3.29).

7.1.4 A Good Time to Stop and Review!

The material we've discussed in this section, that is since page 183, is absolutely key, forming the very basis of statistics. It will be used throughout all our chapters here on statistics. It would be highly worthwhile for the reader to review this section before continuing.

7.2 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:

7.2.1 How Long Should We Run a Simulation?

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 2.12.1. Now we will finally address this issue.

As our example, recall from the Bus Paradox in Section 6.3: 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 for the next bus?

We later 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 we 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:

```
1  doexpt <- function(opt) {
2     lastarrival <- 0.0
3     while (lastarrival < opt)
4        lastarrival <- lastarrival + rexp(1,0.1)
```

```
5     return(lastarrival-opt)
6   }
7
8   observationpt <- 240
9   nreps <- 1000
10  waits <- vector(length=nreps)
11  for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
12  cat("approx. mean wait = ",mean(waits),"\n")
```

Running the program yields

```
approx. mean wait = 9.653743
```

Was 1000 iterations enough? How close is this value 9.653743 to the true expected value of waiting time?[4]

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.

7.2.2 Confidence Intervals for Means

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?

7.2.2.1 Our First Confidence Interval

So, suppose we have a random sample $W_1, ..., W_n$ from some population with mean μ and variance σ^2.

Recall that (7.14) 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 $P(-0.96 < T < 1.96) = 0.95$.

[4] 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.

Thus

$$0.95 \approx P\left(-1.96 < \frac{\overline{W} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \tag{7.19}$$

(Note the approximation sign.) Doing a bit of algebra on the inequalities yields

$$0.95 \approx P\left(\overline{W} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{W} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{7.20}$$

Now remember, not only do we not know μ, we also don't know σ. But we can estimate it, as we saw, via (7.17). One can show (the details will be given in Section 8.5) that (7.20) is still valid if we substitute s for σ, i.e.

$$0.95 \approx P\left(\overline{W} - 1.96\frac{s}{\sqrt{n}} < \mu < \overline{W} + 1.96\frac{s}{\sqrt{n}}\right) \tag{7.21}$$

In other words, we are about 95% sure that the interval

$$(\overline{W} - 1.96\frac{s}{\sqrt{n}}, \overline{W} + 1.96\frac{s}{\sqrt{n}}) \tag{7.22}$$

contains μ. This is called a 95% **confidence interval** for μ. The quantity $1.96\frac{s}{\sqrt{n}}$ is the margin of error.

We could add this feature to our program:

```
1  doexpt <- function(opt) {
2     lastarrival <- 0.0
3     while (lastarrival < opt)
4        lastarrival <- lastarrival + rexp(1,0.1)
5     return(lastarrival-opt)
6  }
7
8  observationpt <- 240
9  nreps <- 10000
10 waits <- vector(length=nreps)
11 for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
12 wbar <- mean(waits)
13 cat("approx. mean wait =",wbar,"\n")
14 s2 <- mean(waits^2) - wbar^2
15 s <- sqrt(s2)
16 radius <- 1.96*s/sqrt(nreps)
17 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. 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 7.2.3.

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 7.2.8.

7.2.3 Meaning of Confidence Intervals

7.2.3.1 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.[5]

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, $\overline{W}$. But we don't know how accurate that estimate might be. That's the reason we form the confidence interval, as a gauge of the accuracy of $\overline{W}$ as an estimate of μ.

Say our interval (7.22) 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.

[5] 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 7.2.2, 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*.

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.

7.2.3.2 One More Point 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. This is clear in our program above—the 10 is constant, while **wbar** and **s** vary from interval to interval.

So, it was reasonable for teachers to warn students not to think μ is a random variable. But later on, some idiot must have then decided that it is incorrect to say "μ is in the interval," and other idiots then followed suit. They continue to this day, sadly.

7.2.4 General Formation of Confidence Intervals from Approximately Normal Estimators

Recall that the idea of a confidence interval is really simple: We report our estimate, plus or minus a margin of error. In (7.22),

margin of error = $1.96\times$ estimated standard deviation of $\overline{W}$ = $1.96 \times \frac{s}{\sqrt{n}}$

Remember, $\overline{W}$ is a random variable. In our Davis people example, each line of the notebook would correspond to a different sample of 1000 people, and thus each line would have a different value for $\overline{W}$. Thus it makes sense to talk about $Var(\overline{W})$, and to refer to the square root of that quantity, i.e. the standard deviation of $\overline{W}$. In (7.13), we found this to be $\sigma/\sqrt{n}$ and decided to estimate it by $s/\sqrt{n}$. The latter is called the **standard error of the estimate** (or just **standard error**, s.e.), meaning the estimate of the standard deviation of the estimate $\overline{W}$. (The word *estimate* was used twice in the preceding sentence. Make sure to understand the two different settings that they apply to.)

That gives us a general way to form confidence intervals, as long as we use approximately normally distributed estimators:

Definition 27 *Suppose $\widehat{\theta}$ is a sample-based estimator of a population quantity θ.*[6] *The sample-based estimate of the standard deviation of $\widehat{\theta}$ is called the standard error of $\widehat{\theta}$.*

We can see from (7.22) what to do in general:

> 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.
>
> Then an approximate 95% confidence interval for θ is
>
> $$\widehat{\theta} \pm 1.96 \cdot \text{s.e.}(\widehat{\theta}) \tag{7.23}$$
>
> 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.

[6] The quantity is pronounced "theta-hat." The "hat" symbol is traditional for "estimate of."

7.2.5 Confidence Intervals for Proportions

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:

```
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) - 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
cat("approx. P(W > 6.4) =",prop,"\n")
```

The value printed out for the probability is 0.516. We again ask the question, how can we gauge the accuracy of this number as an estimator of the true probability $P(W > 6.4)$?

7.2.5.1 Derivation

It turns out that we already have our answer, from Section 3.6. We found there that if

$$Y = \begin{cases} 1, & \text{if } W > 6.4 \\ 0, & \text{otherwise} \end{cases} \tag{7.24}$$

then setting $p = P(W > 6.4)$, we have

$$E(Y) = P(W > 6.4) = p \tag{7.25}$$

Let Y_i be the value of Y for our ith data point. Then

$$\widehat{p} = \overline{Y} \tag{7.26}$$

For instance, say our sample size is 3, with values 1, 1 and 0. Then $\overline{Y} = 2/3$, which is exactly the same as $\widehat{p}$.

So, a sample proportion is really a special case of a sample mean estimating a population mean, and thus we can use (7.22)!

Actually, let's depart from (7.22), and view things in terms of standard errors. Since $\widehat{p}$ comes from a sum, we can use (7.23), with θ being p and $\widehat{\theta}$ being $\widehat{p}$. All we need is the standard error of $\widehat{p}$.

The latter is, by definition, the estimated standard deviation of $\widehat{p}$. So, let's find the variance of $\widehat{p}$, then estimate that quantity, then take the square roo.

Again due to (7.26), we can use (7.8):

$$Var(\widehat{p}) = Var(Y)/n \tag{7.27}$$

But in Section 3.6, we found that Var(Y) = p(1-p). So

$$\text{s.e.}(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n} \tag{7.28}$$

Equation (7.23) becomes

$$\left(\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}, \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\right) \tag{7.29}$$

And note again that $\sqrt{\widehat{p}(1-\widehat{p})/n}$ is the standard error of $\widehat{p}$.

7.2.5.2 Examples

We incorporate that into our program:

```
1  doexpt <- function(opt) {
2     lastarrival <- 0.0
3     while (lastarrival < opt)
4        lastarrival <- lastarrival + rexp(1,0.1)
5     return(lastarrival-opt)
6  }
7
8  observationpt <- 240
9  nreps <- 1000
10 waits <- vector(length=nreps)
11 for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
12 wbar <- mean(waits)
13 cat("approx. mean wait =",wbar,"\n")
14 s2 <- (mean(waits^2) - mean(wbar)^2)
```

```
15 s <- sqrt(s2)
16 radius <- 1.96*s/sqrt(nreps)
17 cat("approx. CI for EW =",wbar-radius,"to",wbar+radius,"\n")
18 prop <- length(waits[waits > 6.4]) / nreps
19 s2 <- prop*(1-prop)
20 s <- sqrt(s2)
21 radius <- 1.96*s/sqrt(nreps)
22 cat("approx. P(W > 6.4) =",prop,", with a margin of error of",radius,"\n")
```

In this case, we get 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."

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 (7.29) 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 (7.29).

7.2.5.3 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 7.2.5.2, 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 computer has a 95% chance of containing the population value, since 95% of all intervals will contain it.

7.2.5.4 (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.

7.2.5.5 Planning Ahead

Now, why do the pollsters sample 1200 people?

First, note that the maximum possible value of $\widehat{p}(1-\widehat{p})$ is 0.25.[7] 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.

7.2.6 Confidence Intervals for Differences of Means or Proportions

7.2.6.1 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 $\overline{X} - \overline{Y}$.

[7] Use calculus to find the maximum value of f(x) = x(1-x).

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 (7.23). Here:

- θ is $\mu_1 - \mu_2$
- $\widehat{\theta}$ is $\overline{X} - \overline{Y}$

So, we need to find the standard error of $\overline{X} - \overline{Y}$.

Let's find the standard deviation of $\overline{X} - \overline{Y}$, and then estimate it from the data. We have

$$\begin{aligned}
\text{std.dev.}(\overline{X} - \overline{Y}) &= \sqrt{Var[\overline{X} - \overline{Y}]} \quad \text{(def.)} && (7.30)\\
&= \sqrt{Var[\overline{X} + (-1)\overline{Y}]} \quad \text{(algebra)} && (7.31)\\
&= \sqrt{Var(\overline{X}) + Var[(-1)\overline{Y}]} \quad \text{(indep.)} && (7.32)\\
&= \sqrt{Var(\overline{X}) + Var(\overline{Y})} \quad \text{(3.32.)} && (7.33)\\
&= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{(7.13)} && (7.34)
\end{aligned}$$

Note that we used the fact that $\overline{X}$ and $\overline{Y}$ are independent, as they come from separate people.

Replacing the σ_i^2 values by their sample estimates,

$$s_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \overline{X})^2 \tag{7.35}$$

and

$$s_2^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2 \tag{7.36}$$

we finally have

$$\text{s.e.}(\overline{X} - \overline{Y}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{7.37}$$

Thus (7.23) tells us that an approximate 95% confidence interval for $\mu_1 - \mu_2$ is

$$\left(\overline{X} - \overline{Y} - 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \overline{X} - \overline{Y} + 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) \tag{7.38}$$

What about confidence intervals for the difference in two population proportions $p_1 - p_2$? Recalling that in Section 7.2.5 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 (7.38). 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, (7.38) reduces to

$$\left(\widehat{p}_1 - \widehat{p}_2 - 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \widehat{p}_1 - \widehat{p}_2 + 1.96\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_2}} \right) \tag{7.39}$$

Example: In a network security application, C. Mano *et al*[8] 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

$$11.520 - 2.000 \pm 1.96\sqrt{\frac{9.939^2}{344} + \frac{6.299^2}{436}} = 9.52 \pm 1.22 \tag{7.40}$$

So you can see that there is a big difference between the two networks, even after allowing for sampling variation.

[8] 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.

7.2.6.2 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 (7.38).

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 (7.38), 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 (7.22) on these values, so that our approximate 95% confidence interval is

$$(\overline{T} - 1.96\frac{s}{\sqrt{n}}, \overline{T} + 1.96\frac{s}{\sqrt{n}}) \tag{7.41}$$

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_Z 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 (7.39). 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 (5.30):

$$Var(\widehat{p}_A - \widehat{p}_B) = Var(\widehat{p}_A) + Var(\widehat{p}_B) - 2Cov(\widehat{p}_A, \widehat{p}_B) \tag{7.42}$$

Or, we could have taken a matrix approach, using (5.107) with A equal to the row vector (1,-1,0).

So, using (5.153), the standard error of $\widehat{p}_A - \widehat{p}_B$ is

$$\sqrt{0.001\widehat{p}_A(1-\widehat{p}_A) + 0.001\widehat{p}_B(1-\widehat{p}_B) + 0.002\widehat{p}_A\widehat{p}_B} \tag{7.43}$$

7.2.7 Example: Machine Classification of Forest Covers

Remote sensing is machine classification of type from variables observed aerially, typically by satellite. In 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 12 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 (7.38), we have that the interval is

$$223.8 - 226.3 \pm 1.96\sqrt{\frac{15.3^2}{226} + \frac{14.3^2}{585}} = -2.5 \pm 2.3 = (-4.8, -0.3) \tag{7.44}$$

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 7.4.

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

$$\widehat{p}_1 - \widehat{p}_2 = 0.226 - 0.585 = -0.359 \tag{7.45}$$

The standard error of this quantity, from (7.43), is

$$\sqrt{0.001 \cdot 0.226 \cdot 0.7740.001 \cdot 0.585 \cdot 0.415 + 002 \cdot 0.226 \cdot 0.585} = 0.019 \tag{7.46}$$

That gives us a confidence interval of

$$-0.359 \pm 1.96 \cdot 0.019 = (-0.397, -0.321) \tag{7.47}$$

7.2.8 And What About the Student-t Distribution?

Another thing we are not doing here is to use the **Student t-distribution**. That is the name of the distribution of the quantity

$$T = \frac{\overline{W} - \mu}{\tilde{s}/\sqrt{n}} \tag{7.48}$$

where $\tilde{s}^2$ is the version of the sample variance in which we divide by n-1 instead of by n, i.e. (7.18).

Note carefully that we are assuming that the W_i themselves—not just $\overline{W}$—have a normal distribution. The exact distribution of T is called the **Student t-distribution with n-1 degrees of freedom**. These distributions thus form a one-parameter family, with the degrees of freedom being the parameter.

This distribution has been tabulated. In R, for instance, the functions **dt()**, **pt()** and so on play the same roles as **dnorm()**, **pnorm()** etc. do for the normal family. The call **qt(0.975,9)** returns 2.26. This enables

us to get an for μ from a sample of size 10, at EXACTLY a 95% confidence level, rather than being at an APPROXIMATE 95% level as we have had here, as follows.

We start with (7.19), replacing 1.96 by 2.26, $(\bar{W} - \mu)/(\sigma/\sqrt{n})$ by T, and $\approx$ by $=$. Doing the same algebra, we find the following confidence interval for μ:

$$(\overline{W} - 2.26\frac{\tilde{s}}{\sqrt{10}}, \overline{W} + 2.26\frac{\tilde{s}}{\sqrt{10}}) \tag{7.49}$$

Of course, for general n, replace 2.26 by $t_{0.975,n-1}$, the 0.975 quantile of the t-distribution with n-1 degrees of freedom. The distribution is tabulated by the R functions **dt()**, **p(t)** and so on.

I do not use the t-distribution here because:

- It depends on the parent population having an exact normal distribution, which is never really true. In the Davis case, for instance, people's weights are approximately normally distributed, but definitely not exactly so. For that to be exactly the case, some people would have to have weights of say, a billion pounds, or negative weights, since any normal distribution takes on all values from $-\infty$ to ∞.
- For large n, the difference between the t-distribution and N(0,1) is negligible anyway.

7.2.9 Other Confidence Levels

We have been using 95% as our confidence level. This is common, but of course not unique. We can for instance use 90%, which gives us a narrower interval (in (7.22),we multiply by 1.65 instead of by 1.96, which the reader should check), at the expense of lower confidence.

A confidence interval's error rate is usually denoted by $1 - \alpha$, so a 95% confidence level has $\alpha = 0.05$.

7.2.10 Real Populations and Conceptual Populations

In our example in Section 7.2.3.1, we were sampling from a real population. However, in many, probably most applications of statistics, either the population or the sampling is more conceptual.

Consider an experiment we will discuss in Section 10.1.2, in which we compare the programmability of three scripting languages. (You need not read ahead.) We divide our programmers into three groups, and assign each group to program in one of the languages. We then compare how long it took the three groups to finish writing and debugging the code, and so on.

We think of our programmers as being a random sample from the population of all programmers, but that is probably an idealization. We probably did NOT choose our programmers randomly; we just used whoever

we had available. But we can think of them as a "random sample" from the rather conceptual "population" of all programmers who *might* work at this company.[9]

You can see from this that if one chooses to apply statistics carefully—which you absolutely should do—there sometimes are some knotty problems of interpretation to think about.

7.2.11 One More Time: Why Do We Use Confidence Intervals?

After all the variations on a theme in the very long Section 7.2, it is easy to lose sight of the goal, so let's review:

Almost everyone is familiar with the term "margin of error," given in every TV news report during elections. The report will say something like, "In our poll, 62% stated that they plan to vote for Ms. X. The margin of error is 3%." Those two numbers, 62% and 3%, form the essence of confidence intervals:

- The 62% figure is our estimate of p, the true population fraction of people who plan to vote for Ms. X.
- Recognizing that that 62% figure is only a sample estimate of p, we wish to have a measure of how accurate the figure is—our margin of error. Though the poll reports don't say this, what they are actually saying is that we are 95% sure that the true population value p is in the range 0.62 ± 0.03.

So, a confidence interval is nothing more than the concept of the $a \pm b$ range that we are so familiar with.

7.3 Significance Testing

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 (7.29). 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**

[9] You're probably wondering why we haven't discussed other factors, such as differing levels of experience among the programmers. This will be dealt with in our unit on regression analysis, Chapter 10.

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 be counter to the traditional usage of statistics. Most users of statistics would use the toss data to test the **null hypothesis**

$$H_0 : p = 0.5 \tag{7.50}$$

against the **alternate hypothesis**

$$H_A : p \neq 0.5 \tag{7.51}$$

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 7.4.

7.3.1 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...?"

We form the **test statistic**

$$Z = \frac{\widehat{p} - 0.5}{\sqrt{\frac{1}{n} \cdot 0.5(1 - 0.5)}} \tag{7.52}$$

Using the material in Section 7.2.5.1, we have that if H_0 were true, $\widehat{p}$ would have an approximately normal distribution with mean 0.5 and variance $\frac{1}{n} \cdot 0.5(1 - 0.5)$. Thus the random variable Z above would have an approximate N(0,1) distribution. The basic idea is that if Z turns out to have a value which is rare for that distribution, we say, "Rather than believe we've observed a rare event, we choose instead to abandon our assumption that H_0 is true."

So, what do we take for our cutoff value for "rareness"? This probability is called the **significance level**, denoted by α. The classical value for α is 0.05. As mentioned above, if H_0 were true, Z would have an

approximate N(0,1) distribution. As we know from our derivation of confidence intervals, in that distribution there is 2.5% area to the left of -1.96 and 2.5% to the right of 1.96. Thus if H_0 were true, Z would be less than -1.96 or greater than 1.96 only 5% of the time (i.e. in 5% of all samples from this population), a "rare event."

So, if Z does stray that far (i.e. 1.96 or more in either direction) from 0, we reject H_0, and decide that $p \neq 0.5$. We say, "The value of p is significantly different from 0.5."

Let X be the number of heads we get from our 100 tosses. Note that our rule for decision making formulated above is equivalent (do the algebra to see this for yourself) to saying that we will accept H_0 if $40 \leq X \leq 60$, and reject it otherwise.

7.3.2 General Testing Based on Normally Distributed Estimators

In Section 7.2.4, 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

$$Z = \frac{\widehat{\theta} - c}{s.e.(\widehat{\theta})} \tag{7.53}$$

where $s.e.(\widehat{\theta})$ is the standard error of $\widehat{\theta}$ (Section 7.2.4), and proceed as before:

> Reject $H_0 : \theta = c$ at the significance level of $\alpha = 0.05$ if $|Z| \geq 1.96$.

7.3.3 Example: Network Security

Let's look at the network security example in Section 7.2.6.1 again. Here $\widehat{\theta} = \overline{X} - \overline{Y}$, and c is presumably 0 (depending on the goals of Mano *et al*). From 7.37, the standard error works out to 0.61. So, our test statistic (7.53) is

$$Z = \frac{\overline{X} - \overline{Y} - 0}{0.61} = \frac{11.52 - 2.00}{0.61} = 15.61 \tag{7.54}$$

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.

7.3.4 The Notion of "p-Values"

In that example above, the Z value, 15.61, was far larger than the cutoff for rejection of H_0, 1.96. You might say that we "resoundingly" rejected H_0. When data analysts encounter such a situation, they want to indicate it in their reports. This is done through something called the **observed significance level**, more often called the **p-value**.

To illustrate this, let's look at a somewhat milder case, say in which Z = 2.14. By checking a table of the N(0,1) distribution, or by calling **pnorm(2.14)** in R, we would find that the N(0,1) distribution has area 0.016 to the right of 2.14, and of course by symmetry there is an equal area to the left of -2.14. In other words, in the general formulation in Section 7.3.2, we would be able to reject H_0 even at the much more stringent significance level of 0.032 instead of 0.05. So, a Z = 2.14 would be considered even more significant than Z = 1.96. In the research community it is customary to say, "The p-value was 0.032."[10] The smaller the p-value, the more significant the results are considered.

In our 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.

It is customary to denote small p-values by asterisks. This 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.

7.3.5 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

$$H_A : p > 0.5 \tag{7.55}$$

A "rare event" which could make us abandon our belief in H_0 would now be if Z in (7.52) 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] 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.

7.3.6 Exact Tests

Remember, the tests we've seen so far are all approximate. In (7.52), 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.[11]

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.

Example:

Let's consider the coin example again, with the one-sided alternative (7.55) . 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?

$$\alpha = \sum_{i=8}^{10} P(X = i) = \sum_{i=8}^{10} \left(\begin{array}{c} 10 \\ i \end{array} \right) \left(\frac{1}{2} \right)^{10} = 0.055 \tag{7.56}$$

That's not the usual 0.05. Clearly we cannot get an exact significance level of 0.05,[12] but our α is 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.

Example:

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.

Example:

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

$$H_0 : \mu = 1000 \tag{7.57}$$

[11] 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.

[12] Actually, it could be done by introducing some randomization to our test.

vs.

$$H_A : \mu > 1000 \tag{7.58}$$

It is natural to have our test take the form in which we reject H_0 if

$$\overline{X} > w \tag{7.59}$$

for some constant w chosen so that

$$P(\overline{X} > w) = 0.05 \tag{7.60}$$

under H_0. Suppose we want an exact test, not one based on a normal approximation.

Recall that $100\overline{X}$, 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.

7.4 What's Wrong with Significance Testing

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.

7.4.1 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. (Many are forced to do so, e.g. to comply with government standards in pharmaceutical testing.) 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 Nobel laureates Richard Feynman (physics) and Kenneth Arrow (economics).
- 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.[13]
- 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.

7.4.2 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.

Consider the coin example, for instance. No coin is absolutely perfectly balanced, with p = 0.5000000000000000000000000 We know that before even collecting any data. In other words, we know beforehand 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 (7.52) as the sample size n grows. If we have a large enough sample, eventually the denominator in (7.52) will be small enough, and $\widehat{p}$ will be close enough to

[13] *Statistics*, third edition, by David Freedman, Robert Pisani, Roger Purves, pub. by W.W. Norton, 1997.

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 (7.52) 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.

7.4.3 What to 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

$$H_0 : 0.49 \leq p \leq 0.51 \tag{7.61}$$

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?

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.

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 10). 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 9.

7.4.4 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.

7.4.5 Example: the Forest Cover Data

In Section 7.2.7, we found that an approximate 95% confidence interval for $\mu_1 - \mu_2$ was

$$223.8 - 226.3 \pm 2.3 = (-4.8, -0.3) \tag{7.62}$$

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

$$H_0 : \mu_1 = \mu_2 \tag{7.63}$$

and say that the two means are "significantly" different, which sounds like there is an important difference—which there is not.

7.4.6 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.**

Exercises

1. Consider Equation (7.21). In each of the entries in the table below, fill in either R for random, or NR for nonrandom:

quantity	R or NR?
$\overline{W}$	
s	
μ	
n	

2. Consider $\hat{p}$, the estimator of a population proportion p, based on a sample of size n. Give the expression for the standard error of $\hat{p}$.

3. Suppose we take a simple random sample of size 2 from a population consisting of just three values, 66, 67 and 69. Let $\overline{X}$ denote the resulting sample mean. Find $p_{\overline{X}}(67.5)$.

4. Suppose we have a random sample $W_1, ..., W_n$, and we wish to estimate the population mean μ, as usual. But we decide to place double weight on W_1, so our estimator for μ is

$$U = \frac{2W_1 + W_2 + ... + W_n}{n+1} \tag{7.64}$$

Find E(U) and Var(U) in terms of μ and the population variance σ^2.

5. Suppose a random sample of size n is drawn from a population in which, unknown to the analyst, X actually has an exponential distribution with mean 10. Suppose the analyst forms an approximate 95% confidence interval for the mean, using (7.21). Use R simulation to find the true confidence level, for n = 10, 25, 100 and 500.

6. Suppose we draw a sample of size 2 from a population in which X has the values 10, 15 and 12. Find $p_{\overline{X}}$, first assuming sampling with replacement, then assuming sampling without replacement.

7. We ask 100 randomly sampled programmers whether C++ is their favorite language, and 12 answer yes. Give a numerical expression for an approximate 95% confidence interval for the population fraction of programmers who have C++ as their favorite language.

8. In Equation (7.22), suppose 1.96 is replaced by 1.88 in both instances. Then of course the confidence level will be smaller than 95%. Give a call to an R function (not a simulation), that will find the new confidence level.

9. Candidates A, B and C are vying for election. Let p_1, p_2 and p_3 denote the fractions of people planning to vote for them. We poll n people at random, yielding estimates $\widehat{p_1}$, $\widehat{p_2}$ and $\widehat{p_3}$. Y claims that she has more supporters than the other two candidates combined. Give a formula for an approximate 95% confidence interval for $p_2 - (p_1 + p_3)$.

Hint: There is a multinomial distribution involved.

10. In the light bulb example on page 208, suppose the actual observed value of $\overline{X}$ turns out to be 15.88. Find the p-value.

11. Suppose Jack and Jill each collect random samples of size n from a population having unknown mean μ but KNOWN variance σ^2. They each form an approximate 95% confidence interval for μ, using (7.22) but with s replaced by σ. Find the approximate probability that their intervals do not overlap. Express your answer in terms of Φ, the cdf of the N(0,1) distribution.

12. In the example of the population of three people, page 185, find the following:

(a) $p_{X_1}(70)$

(b) $p_{X_1,X_2}(69,70)$

(c) $F_{\overline{X}}(69.5)$

(d) probability that $\overline{X}$ overestimates the population mean μ

(e) $p_{\overline{X}}(69)$ if our sample size is three rather than two (remember, we are sampling with replacement)

13. In the derivation (7.8), suppose instead we have a simple random sample. Which one of the following statements is correct?

(a) $E(\overline{X})$ will still be equal to μ.

(b) $E(\overline{X})$ will not exist.

(c) $E(\overline{X})$ will exist, but may be less than μ.

(d) $E(\overline{X})$ will exist, but may be greater than μ.

(e) None of the above is necessarily true.

Chapter 8

General Statistical Estimation and Inference

In the last chapter, we often referred to certain estimators as being "natural." For example, if we are estimating a population mean, an obvious choice of estimator would be the sample mean. But in many applications, it is less clear what a "natural" estimate for a parameter of interest would be.[1] We will present general methods for estimation in this section.

We will also discuss advanced methods of inference.

8.1 General Methods of Parametric Estimation

Let's begin with a simple motivating example.

8.1.1 Example: Guessing the Number of Raffle Tickets Sold

You've just bought a raffle ticket, and find that you have ticket number 68. You check with a couple of friends, and find that their numbers are 46 and 79. Let c be the total number of tickets. How should we estimate c, using our data 68, 46 and 79?

It is reasonable to assume that each of the three of you is equally likely to get assigned any of the numbers 1,2,...,c. In other words, the numbers we get, X_i, i = 1,2,3 are uniformly distributed on the set $\{1,2,...,c\}$.

[1] Recall that we are using the term *parameter* to mean any population quantity, rather an an index into a parametric family of distributions.

We can also assume that they are independent; that's not exactly true, since we are sampling without replacement, but for large c—or better stated, for n/c small—it's close enough.

So, we are assuming that the X_i are independent and identically distributed—famously written as **i.i.d.** in the statistics world—on the set {1,2,...,c}. How do we use the X_i to estimate c?

8.1.2 Method of Moments

One approach, an intuitive one, would be to reason as follows. Note first that

$$E(X) = \frac{c+1}{2} \tag{8.1}$$

Let's solve for c:

$$c = 2EX - 1 \tag{8.2}$$

We know that we can use

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{8.3}$$

to estimate EX, so by (8.2), $2\overline{X} - 1$ is an intuitive estimate of c. Thus we take our estimator for c to be

$$\widehat{c} = 2\overline{X} - 1 \tag{8.4}$$

This estimator is called the Method of Moments estimator of c.

Let's step back and review what we did:

- We wrote our parameter as a function of the population mean EX of our data item X. Here, that resulted in (8.2).
- In that function, we substituted our sample mean $\overline{X}$ for EX, and substituted our estimator $\widehat{c}$ for the parameter c, yielding (8.4). We then solved for our estimator.

We say that an estimator $\widehat{\theta}$ of some parameter θ is **consistent** if

$$\lim_{n\to\infty} \widehat{\theta} = \theta \tag{8.5}$$

where n is the sample size. In other words, as the sample size grows, the estimator eventually converges to the true population value.

Of course here $\overline{X}$ is a consistent estimator of EX . Thus you can see from (8.2) and (8.4) that $\widehat{c}$ is a consistent estimator of c. In other words, the Method of Moments generally gives us consistent estimators.

What if we have more than one parameter to estimate? We generalize what we did above:

- Suppose we are estimating a parametric distribution with parameters $\theta_1, ..., \theta_r$.
- Let η_i denote the ith **moment** of X, $E(X^i)$.
- For i = 1,...,r we write η_i as a function g_i of all the θ_k.
- For i = 1,...,r set

$$\widehat{\eta_i} = \frac{1}{n} \sum_{j=1}^{n} X_j^i \tag{8.6}$$

- Substitute the $\widehat{\theta_k}$ in the g_i and then solve for them.

In the above example with the raffle, we had r = 1, $\theta_1 = c$, $g_1(c) = (c+1)/2$ and so on. A two-parameter example will be given below.

8.1.3 Method of Maximum Likelihood

Another method, much more commonly used, is called the **Method of Maximum Likelihood**. In our example above, it means asking the question, "What value of c would have made our data—68, 46, 79—most likely to happen?" Well, let's find what is called the **likelihood**, i.e. the probability of our particular data values occurring:

$$L = P(X_1 = 68, X_2 = 46, X_3 = 79) = \begin{cases} (\frac{1}{c})^3, & \text{if } c \geq 79 \\ 0, & \text{otherwise} \end{cases} \tag{8.7}$$

Now keep in mind that c is a fixed, though unknown constant. It is not a random variable. What we are doing here is just asking "What if" questions, e.g. "If c were 85, how likely would our data be? What about c = 91?"

Well then, what value of c maximizes (8.7)? Clearly, it is c = 79. Any smaller value of c gives us a likelihood of 0. And for c larger than 79, the larger c is, the smaller (8.7) is. So, our maximum likelihood estimator (MLE) is 79. In general, if our sample size in this problem were n, our MLE for c would be

$$\check{c} = \max_i X_i \tag{8.8}$$

8.1.4 Example: Estimation the Parameters of a Gamma Distribution

As another example, suppose we have a random sample $X_1, ..., X_n$ from a gamma distribution.

$$f_X(t) = \frac{1}{\Gamma(c)} \lambda^c t^{c-1} e^{-\lambda t},\ t > 0 \tag{8.9}$$

for some unknown c and λ. How do we estimate c and λ from the X_i?

8.1.4.1 Method of Moments

Let's try the Method of Moments, as follows. We have two population parameters to estimate, c and λ, so we need to involve two moments of X. That could be EX and $E(X^2)$, but here it would more conveniently be EX and Var(X). We know from our previous unit on continuous random variables, Chapter 4, that

$$EX = \frac{c}{\lambda} \tag{8.10}$$

$$Var(X) = \frac{c}{\lambda^2} \tag{8.11}$$

In our earlier notation, this would be r = 2, $\theta_1 = c$, $\theta_2 = \lambda$ and $g_1(c, \lambda) = c/\lambda$ and $g_2(c, \lambda) = c/\lambda^2$.

Switching to sample analogs and estimates, we have

$$\begin{matrix} \widehat{c} \\ \widehat{\lambda} \end{matrix} = \overline{X} \tag{8.12}$$

$$\frac{\widehat{c}}{\widehat{\lambda}^2} = s^2 \tag{8.13}$$

Dividing the two quantities yields

$$\widehat{\lambda} = \frac{\overline{X}}{s^2} \tag{8.14}$$

which then gives

$$\widehat{c} = \frac{\overline{X}^2}{s^2} \tag{8.15}$$

8.1.4.2 MLEs

What about the MLEs of c and λ? Remember, the X_i are continuous random variables, so the likelihood function, i.e. the analog of (8.7), is the product of the density values:

$$\begin{aligned} L &= \Pi_{i=1}^{n} \left[\frac{1}{\Gamma(c)} \lambda^c {X_i}^{c-1} e^{-\lambda X_i} \right] && (8.16) \\ &= [\lambda^c/\Gamma(c)]^n \left(\Pi_{i=1}^{n} X_i\right)^{c-1} e^{-\lambda \sum_{i=1}^{n} X_i} && (8.17) \end{aligned}$$

In general, it is usually easier to maximize the log likelihood (and maximizing this is the same as maximizing the original likelihood):

$$l = (c-1) \sum_{i=1}^{n} \ln(X_i) - \lambda \sum_{i=1}^{n} X_i + nc \ln(\lambda) - n \ln(\Gamma(c)) \tag{8.18}$$

One then takes the partial derivatives of (8.18) with respect to c and λ, and sets the derivatives to zero. The solution values, $\check{c}$ and $\check{\lambda}$, are then the MLEs of c and λ. Unfortunately, in this case, these equations do not have closed-form solutions. So the equations must be solved numerically. (In fact, numerical methods are needed even more in this case, because finding the derivative of $\Gamma(c)$ is not easy.)

8.1.4.3 R's mle() Function

R provides a function, **mle()**, for finding MLEs in mathematically intractable situations such as the one in the last section. Here's an example in the that context. We'll simulate some data from a gamma distribution with given parameter values, then pretend we don't know those, and find the MLEs from the data:

```
x <- rgamma(100,shape=2)  # Erlang, r = 2
n <- length(x)

ll <- function(c,lambda) {
   loglik <- (c-1) * sum(log(x)) - sum(x)*lambda + n*c*log(lambda) -
      n*log(gamma(c))
   return(-loglik)
}

summary(mle(minuslogl=ll,start=list(c=2,lambda=2)))
Maximum likelihood estimation

Call:
mle(minuslogl = ll, start = list(c = 1, lambda = 1))

Coefficients:
       Estimate Std. Error
c      1.993399  0.1770996
lambda 1.027275  0.1167195

-2 log L: 509.8227
```

How did this work? The main task we have is to write a function that calculates negative the log likelihood, with that function's arguments will be the parameters to be estimated. (Note that in R, **log()** calculates the natural logarithm by default.) Fortunately for us, **mle()** calculates the derivatives numerically too, so we didn't need to specify them in the log likelihood function. (Needless to say, this function thus cannot be used in a problem in which derivatives cannot be used, such as the lottery example above.)

We also need to supply **mle()** with initial guesses for the parameters. That's done in the **start** argument. I more or less arbitrarily chose 1.0 for these values. You may have to experiment, though, as some sets of initial values may not result in convergence.

The standard errors of the estimated parameters are also printed out, enabling the formation of confidence intervals and significance tests. See for instance Section 7.2.4. In fact, you can get the estimated covariance matrix for the vector of estimated parameters. In our case here:

```
> mleout <- mle(minuslogl=ll,start=list(c=2,lambda=2))
Warning messages:
1: In log(lambda) : NaNs produced
2: In log(lambda) : NaNs produced
3: In log(lambda) : NaNs produced
> solve(mleout@details$hessian)
                c     lambda
c      0.08434476 0.04156666
lambda 0.04156666 0.02582428
```

By the way, there were also some warning messages, due to the fact that during the iterative maximization process, some iterations generated guesses for $\check{\lambda}$ were 0 or near it, causing problems with **log()**.

8.1.5 More Examples

Suppose $f_W(t) = ct^{c-1}$ for t in (0,1), with the density being 0 elsewhere, for some unknown $c > 0$. We have a random sample $W_1, ..., W_n$ from this density.

Let's find the Method of Moments estimator.

$$EW = \int_0^1 tct^{c-1}\, dt = \frac{c}{c+1} \tag{8.19}$$

So, set

$$\overline{W} = \frac{\widehat{c}}{\widehat{c}+1} \tag{8.20}$$

yielding

$$\widehat{c} = \frac{\overline{W}}{1-\overline{W}} \tag{8.21}$$

What about the MLE?

$$L = \Pi_{i=1}^n cW_i^{c-1} \tag{8.22}$$

so

$$l = n \ln c + (c-1) \sum_{i=1}^n \ln W_i \tag{8.23}$$

Then set

$$0 = \frac{n}{\widehat{c}} + \sum_{i=1}^n \ln W_i \tag{8.24}$$

and thus

$$\widehat{c} = -\frac{1}{\frac{1}{n}\sum_{i=1}^n \ln W_i} \tag{8.25}$$

As in Section 8.1.3, not every MLE can be determined by taking derivatives. Consider a continuous analog of the example in that section, with $f_W(t) = \frac{1}{c}$ on (0,c), 0 elsewhere, for some $c > 0$.

The likelihood is

$$\left(\frac{1}{c}\right)^n \tag{8.26}$$

as long as

$$c \geq \max_i W_i \tag{8.27}$$

and is 0 otherwise. So,

$$\widehat{c} = \max_i W_i \tag{8.28}$$

as before.

Now consider a different problem. Suppose the random variable X is equal to 1, 2 and 3, with probabilities c, c and 1-2c. The value c is thus a population parameter. We have a random sample $X_1, ..., X_n$ from this population. Let's find the Method of Moments Estimator of c, and its bias.

First,

$$EX = c \cdot 1 + c \cdot 2 + (1 - 2c) \cdot 3 = 3 - 3c \tag{8.29}$$

Thus

$$c = (3 - EX)/3 \tag{8.30}$$

and so set

$$\hat{c} = (3 - \overline{X})/3 \tag{8.31}$$

Next,

$$
\begin{aligned}
E\hat{c} &= E[(3-\overline{X})/3) && (8.32)\\
&= \frac{1}{3}\cdot(3-E\overline{X}) && (8.33)\\
&= \frac{1}{3}[3-EX] && (8.34)\\
&= \frac{1}{3}[3-(3-3c)] && (8.35)\\
&= c && (8.36)
\end{aligned}
$$

So, the bias is 0.

8.1.6 What About Confidence Intervals?

Usually we are not satisfied with simply forming estimates (called **point estimates**). We also want some indication of how accurate these estimates are, in the form of confidence intervals (**interval estimates**).

In many special cases, finding confidence intervals can be done easily on an *ad hoc* basis. Look, for instance, at the Method of Moments Estimator in Section 8.1.2. Our estimator (8.4) is a linear function of $\overline{X}$, so we easily obtain a confidence interval for c from one for EX.

Another example is (8.25). Taking the limit as $n \to \infty$ the equation shows us (and we could verify) that

$$c = \frac{1}{E[\ln W]} \tag{8.37}$$

Defining $X_i = \ln W_i$ and $\overline{X} = (X_1 + ... + X_n)/$, we can obtain a confidence interval for EX in the usual way. We then see from (8.37) that we can form a confidence interval for c by simply taking the reciprocal of each endpoint of the interval, and swapping the left and right endpoints.

What about in general? For the Method of Moments case, our estimators are functions of the sample moments, and since the latter are formed from sums and thus are asymptotically normal, the delta method (Section 8.6) can be used to show that our estimators are asymptotically normal and to obtain asymptotic variances for them.

There is a well-developed asymptotic theory for MLEs, which under certain conditions shows asymptotic normality with a certain asymptotic variance, thus enabling confidence intervals. The theory also establishes that MLEs are in a certain sense optimal among all estimators. We will not pursue this here, but will note that **mle()** does give standard errors for the estimates, thus enabling the formation of confidence intervals.

8.2 Bias and Variance

The notions of **bias** and **variance** play central roles in the evaluation of goodness of estimators.

8.2.1 Bias

Definition 28 *Suppose $\hat{\theta}$ is an estimator of θ. Then the* **bias** *of $\hat{\theta}$ is*

$$bias = E(\hat{\theta}) - \theta \tag{8.38}$$

If the bias is 0, we say that the estimator is **unbiased**.

It's very important to note that, in spite of the perjorative-sounding name, bias is not an inherently bad property for an estimator to have. Indeed, most good estimators are at least slightly biased. We'll explore this in the next section.

8.2.2 Why Divide by n-1 in s^2?

It should be noted that it is customary in (7.17) to divide by n-1 instead of n, for reasons that are largely historical. Here's the issue:

If we divide by n, as we have been doing, then it turns out that s^2 is biased.

$$E(s^2) = \frac{n-1}{n} \cdot \sigma^2 \tag{8.39}$$

Think about this in the Davis people example, once again in the notebook context. Remember, here n is 1000, and each line of the notebook represents our taking a different random sample of 1000 people. Within each line, there will be entries for W_1 through W_{1000}, the weights of our 1000 people, and for $\overline{W}$ and s. For convenience, let's suppose we record that last column as s^2 instead of s.

Now, say we want to estimate the population variance σ^2. As discussed earlier, the natural estimator for it would be the sample variance, s^2. What (8.39) says is that after looking at an infinite number of lines in the notebook, the average value of s^2 would be just...a...little...bit...too...small. All the s^2 values would average out to $0.999\sigma^2$, rather than to σ^2. We might say that s^2 has a little bit more tendency to underestimate σ^2 than to overestimate it.

So, (8.39) implies that s^2 is a biased estimator of the population variance σ^2, with the amount of bias being

$$\frac{n-1}{n} \cdot \sigma^2 - \sigma^2 = -\frac{1}{n} \cdot \sigma^2 \tag{8.40}$$

Let's prove (8.39). As before, let W be a random variable distributed as the population, and let $W_1, ..., W_n$ be a random sample from that population. So, $EW_i = \mu$ and $Var(W_i) - \sigma^2$, where again μ and σ^2 are the population mean and variance.

It will be more convenient to work with ns^2 than s^2, since it will avoid a lot of dividing by n. So, write

$$\begin{aligned} ns^2 &= \sum_{i=1}^{n}(W_i - \overline{W})^2 \quad \text{(def.)} && (8.41)\\ &= \sum_{i=1}^{n}\left[(W_i - \mu) + (\mu - \overline{W})\right]^2 \quad \text{(alg.)} && (8.42)\\ &= \sum_{i=1}^{n}(W_i - \mu)^2 + 2(\mu - \overline{W})\sum_{i=1}^{n}(W_i - \mu) + n(\mu - \overline{W})^2 \quad \text{(alg.)} && (8.43) \end{aligned}$$

But that middle sum is

$$\sum_{i=1}^{n}(W_i - \mu) = \sum_{i=1}^{n} W_i - n\mu = n\overline{W} - n\mu \tag{8.44}$$

So,

$$ns^2 = \sum_{i=1}^{n}(W_i - \mu)^2 - n(\overline{W} - \mu)^2 \tag{8.45}$$

Now let's take the expected value of (8.45). First,

$$
\begin{aligned}
E\left(\sum_{i=1}^{n}(W_i-\mu)^2\right) &= \sum_{i=1}^{n} E[(W_i-\mu)^2] \quad \text{(E is lin.)} && (8.46)\\
&= \sum_{i=1}^{n} E[(W_i-EW_i)^2] \quad (W_i \text{ distr. as pop.}) && (8.47)\\
&= \sum_{i=1}^{n} Var(W_i) \quad \text{(def. of Var())} && (8.48)\\
&= \sum_{i=1}^{n} \sigma^2 \quad (W_i \text{ distr. as pop.}) && (8.49)\\
&= n\sigma^2 && (8.50)
\end{aligned}
$$

Also,

$$
\begin{aligned}
E[(\overline{W}-\mu)^2] &= E[(\overline{W}-E\overline{W})^2] \quad ((7.8)) && (8.51)\\
&= Var(\overline{W}) \quad \text{(def. of Var())} && (8.52)\\
&= \frac{\sigma^2}{n} \quad (7.13) && (8.53)
\end{aligned}
$$

Applying these last two findings to (8.45), we get (8.39).

$$E(s^2) = \frac{n-1}{n}\sigma^2 \tag{8.54}$$

The earlier developers of statistics were bothered by this bias, so they introduced a "fudge factor" by dividing by n-1 instead of n in (7.17). We will call that $\tilde{s}^2$:

$$\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i-\overline{W})^2 \tag{8.55}$$

This is the "classical" definition of sample variance, in which we divide by n-1 instead of n.

But we will use n. After all, when n is large—which is what we are assuming by using the Central Limit Theorem in the entire development so far—it doesn't make any appreciable difference. Clearly it is not important in our Davis example, or our bus simulation example.

Moreover, speaking generally now rather than necessarily for the case of s^2 there is no particular reason to insist that an estimator be unbiased anyway. An alternative estimator may have a little bias but much smaller

variance, and thus might be preferable. And anyway, even though the classical version of s^2, i.e. $\tilde{s}^2$, is an unbiased estimator for σ^2, s is not an unbiased estimator for σ, the population standard deviation. In other words, unbiasedness is not such an important property.

The R functions **var()** and **sd()** calculate the versions of s^2 and s, respectively, that have a divisor of n-1.

8.2.2.1 Example of Bias Calculation

Let's find the bias of the estimator (8.28).

The bias is $E\widehat{C} - c$. To get $E\widehat{c}$ we need the density of that estimator, which we get as follows:

$$P(\widehat{c} \leq t) = P(\text{all } W_i \leq t) \text{ (definition)} \tag{8.56}$$

$$= \left(\frac{t}{c}\right)^n \text{ (density of } W_i) \tag{8.57}$$

So,

$$f_{\widehat{c}}(t) = \frac{n}{c^n} t^{n-1} \tag{8.58}$$

Integrating against t, we find that

$$E\widehat{c} = \frac{n}{n+1}\, c \tag{8.59}$$

So the bias is c/(n+1), not bad at all.

8.2.3 Tradeoff Between Variance and Bias

Consider a general estimator Q of some population value b. Then a common measure of the quality (of course there are many others) of the estimator Q is the **mean squared error** (MSE),

$$E[(Q-b)^2] \tag{8.60}$$

Of course, the smaller the MSE, the better.

One can break (8.60) down into variance and (squared) bias components, as follows:[2]

[2] In reading the following derivation, keep in mind that EQ and b are constants.

$$
\begin{aligned}
MSE(Q) &= E[(Q-b)^2] \text{ (definition)} && (8.61)\\
&= E[\{(Q-EQ)+(EQ-b)\}^2] \text{ (algebra)} && (8.62)\\
&= E[(Q-EQ)^2] + 2E\left[(Q-EQ)(EQ-b)\right] + E[(EQ-b)^2] \text{ (alg., E props.)} && (8.63)\\
&= E[(Q-EQ)^2] + E[(EQ-b)^2] \text{ (factor out constant EQ-b)} && (8.64)\\
&= Var(Q) + (EQ-b)^2 \text{ (def. of Var(), fact that EQ-b is const.)} && (8.65)\\
&= \text{variance + squared bias} && (8.66)
\end{aligned}
$$

In other words, in discussing the accuracy of an estimator—especially in comparing two or more candidates to use for our estimator—the average squared error has two main components, one for variance and one for bias. In building a model, these two components are often at odds with each other; we may be able to find an estimator with smaller bias but more variance, or vice versa.

We also see from (8.66) that a little bias in an estimator may be quite tolerable, as long as the variance is low. This is good, because as mentioned earlier, most estimators are in fact biased.

These point will become central in Chapters 9 and 10.

8.3 More on the Issue of Independence/Nonindependence of Samples

In Section 7.2.6.1, we derived confidence intervals for the difference between two population means (or proportions). The derivation depended crucially on the fact that the two sample means, $\overline{X}$ and $\overline{Y}$, were independent. This in turn stemmed from the fact that the corresponding sample data sets were separate.

On the other hand, in Section 7.2.6.2, we had an example in which the two sample means, $\overline{X}$ and $\overline{Y}$, were not independent, as they came from the same set of kids. The confidence intervals derived in Section 7.2.6.1 were thus invalid, and new ones were derived, based on differences.

Note that in both cases, the observations *within* a sample were also independent. In the example of children's heights in Section 7.2.6.2, for instance, the fact that Mary was chosen as the first child in the sample had no effect on whether Jane was chosen as the second one. This was important for the derivations too, as they used (7.13), which assumed independence.

In this section, we will explore these points further, with our aim being to state the concepts in precise random variable terms.

As our concrete example, consider an election survey, in a small city. Say there are equal numbers of men and women in the city, 5000 each. We wish to estimate the population proportion of people who plan to vote for candidate A. We take a random sample of size n from the population, Define the following:

- Let V denote the indicator variable for the event that the person plans to vote for A.
- We might be interested in differences between men and women in A's support, so let G be 1 for male, 2 for female.
- Let p denote the population proportion of people who plan to vote for A.
- Let p_1 and p_2 denote the population proportions planning to vote for A, among men and women respectively. Note that

$$p = 0.5p_1 + 0./5p_2 \tag{8.67}$$

- Denote our data by $(V_1, G_1), ..., (V_n, G_n)$, recording both the planned vote and gender for each person in the sample.
- For convenience, relabel the data by gender, with $M_1, ..., M_{N_1}$ and $F_1, ..., F_{N_2}$ denoting the planned votes of the men and women.

Clearly, the male data and female data are independent. The fact that Jack is chosen in the male sample has no impact on whether Jill is chosen in the female one.

But what about data *within* a gender group? For example, are M_1 and M_2, the planned votes of the first two men in our male sample, independent? Or are they correlated, since these two people have the same gender?

The answer is that M_1 and M_2 are indeed independent. The first man could be any of the 5000 men in the city, with probability 1/5000 each, and the same is true of the second man. Moreover, the choice of the first man has no effect at all on the choice of the second one. (Remember, in random samples we sample *with* replacement.)

Our estimate of p is our usual sample proportion,

$$\widehat{p} = \frac{V_1 + ... + V_n}{n} \tag{8.68}$$

Then we can use (7.29) to find a confidence interval for p. But again, the reader might question this, saying something like, "What if G_1 and G_2 are both 1, i.e. the first two people in our sample are both men? Won't V_1 and V_2 then be correlated?" The answer is no, because the reader would be referring to the conditional distribution of V given G, whereas our use of (7.29) does not involve gender, i.e. it concerns the unconditional distribution of V.

This point is subtle, and is difficult for the beginning modeler to grasp. It is related to issues in our first discussions of probability in Chapter 2. In the ALOHA model there, for instance, beginning students who are asked to find $P(X_2 = 1)$ often object, "Well, it depends on what X_1 is." That is incorrect thinking, because they are confusing $P(X_2 = 1)$ with $P(X_2 = 1|X_1 = i)$. That confusion is resolved by thinking in

"notebook" terms, with $P(X_2 = 1)$ meaning the long-run proportion of notebook lines in which $X_2 = 1$, *regardless* of the value of X_1. In our case here, the reader must avoid confusing $P(V = 1)$ (which is p) with $P(V = 1|G = i)$ (which is p_i).

Continuing this point a bit more, note that our $\widehat{p}$ above *is* an unbiased estimate of p:

$$\begin{aligned} E\widehat{p} &= E(V_1) \quad ((7.8)) && (8.69)\\ &= P(V_1 = 1) \quad ((3.43)) && (8.70)\\ &= P(G_1 = 1)P(V_1 = 1|G_1 = 1) + P(G_1 = 2)P(V_1 = 1|G_1 = 2) \quad \text{(Chapter 2)} && (8.71)\\ &= 0.5p_1 + 0.5p_2 && (8.72)\\ &= p \quad ((8.67)) && (8.73) \end{aligned}$$

Due to the independence of the male and female samples, we can use (7.38) to find a confidence interval for $p_1 - p_2$, so that we can compare male and female support of A. Note by the way that $\overline{M}$ will be an unbiased estimate of p_1, with a similar statement holding for the women.

Now, contrast all that with a different kind of sampling, as follows. We choose a gender group at random, and then *sample n people from that gender group*. Let R denote the group chosen, so that $G_i = R$ for all i. So, what about the answers to the above questions in this new setting?

Conditionally on R, the V_i and again independent, using the same argument as we used to show that M_1 and M_2 were independent above. And (8.69) still works, so our $\widehat{p}$ is still unbiased.

However: The V_i are no longer unconditionally independent:

$$P(V_1 = 1 \text{ and } V_2 = 1) = 0.5p_1^2 + 0.5p_2^2 \tag{8.74}$$

(the reader should fill in the details, with a conditioning argument like that in (8.69)), while

$$P(V_1 = 1) \cdot P(V_2 = 1) = p^2 = (0.5p_1 + 0.5p_2)^2 \tag{8.75}$$

So,

$$P(V_1 = 1 \text{ and } V_2 = 1) \neq P(V_1 = 1) \cdot P(V_2 = 1) \tag{8.76}$$

and thus V_1 and V_2 are not unconditionally independent.

This setting is very common. We might, for instance, choose k trees at random, and then collect data on r leaves in each tree.

8.4 Nonparametric Distribution Estimation

Here we will be concerned with estimating distribution functions and densities in settings in which we do not assume our distribution belongs to some parametric model.

8.4.1 The Empirical cdf

Recall that F_X, the cdf of X, is defined as

$$F_X(t) = P(X \leq t), \ -\infty < t < \infty \tag{8.77}$$

Define its sample analog, called the **empirical distribution function**, by

$$\widehat{F}_X(t) = \frac{\#\ of\ X_i\ in\ (-\infty, t)}{n} \tag{8.78}$$

In other words, $F_X(t)$ is the proportion of X that are below t in the population, and $\widehat{F}_X(t)$ is the value of that proportion in our sample. $\widehat{F}_X(t)$ estimates $F_X(t)$ for each t.

Graphically, $\widehat{F}_X$ is a step function, with jumps at the values of the X_i. Specifically, let Y_j, j = 1,...,n denote the sorted version of the X_I.[3] Then

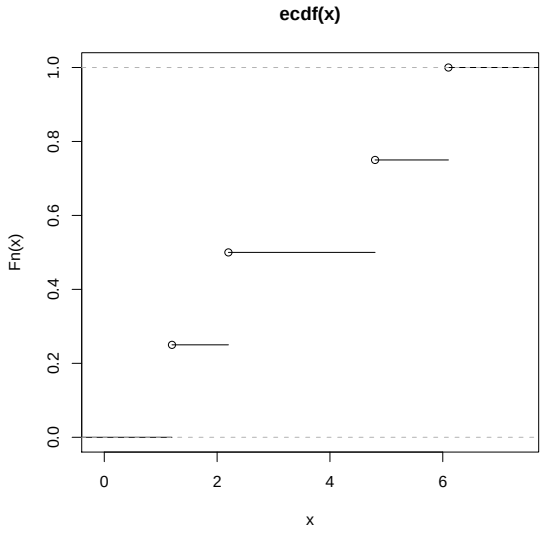
$$\widehat{F}_X(t) = \begin{cases} 0, & \text{for } t < Y_1 \\ \frac{j}{n}, & \text{for } Y_j \leq t < Y_{j+1} \\ 1, & \text{for } t > Y_n \end{cases} \tag{8.79}$$

Here is a simple example. Say n = 4 and our data are 4.8, 1.2, 2.2 and 6.1. We can plot the empirical cdf by calling R's **ecdf()** function:

```
> plot(ecdf(x))
```

Here is the graph:

[3] A common notation for this is $Y_j = X_{(j)}$, meaning that Y_j is the j^{th} smallest of the X_i. These are called the **order statistics** of our sample.



Consider the Bus Paradox example again. Recall that W denoted the time until the next bus arrives. This is called the **forward recurrence time**. The **backward recurrence time** is the time since the last bus was here, which we will denote by R.

Suppose we are interested in estimating the density of R, $f_R()$, based on the sample data $R_1, ..., R_n$ that we gather in our simulation in Section 7.2.1, where n = 1000. How can we do this?[4]

We could, of course, assume that f_R is a member of some parametric family of distributions, say the two-parameter gamma family. We would then estimate those two parameters as in Section 8.1, and possibly check our assumption using goodness-of-fit procedures, discussed in our unit on modeling, Chapter 9. On the other hand, we may wish to estimate f_R without making any parametric assumptions. In fact, one reason we may wish to do so is to visualize the data in order to search for a suitable parametric model.

If we do not assume any parametric model, we have in essence change our problem from estimating a finite number of parameters to an infinite-parameter problem; the "parameters" are the values of $f_X(t)$ for all the different values of t. Of course, we probably are willing to assume *some* structure on f_R, such as continuity, but then we still would have an infinite-parameter problem.

We call such estimation **nonparametric**, meaning that we don't use a parametric model. However, you can see that it is really infinite-parametric estimation.

Again discussed in our unit on modeling, Chapter 9, the more complex the model, the higher the variance of its estimator. **So, nonparametric estimators will have higher variance than parametric ones.** The nonparametric estimators will also generally have smaller bias, of course.

[4] Actually, one can prove that R has an exponential distribution. However, here we'll pretend we don't know that.

8.4.2 Basic Ideas in Density Estimation

Recall that

$$f_R(t) = \frac{d}{dt} F_R(t) = \frac{d}{dt} P(R \leq t) \tag{8.80}$$

From calculus, that means that

$$\begin{aligned} f_R(t) &\approx \frac{P(R \leq t+h) - P(R \leq t-h)}{2h} && (8.81) \\ &= \frac{P(t-h < R \leq t+h)}{2h} && (8.82) \end{aligned}$$

if h is small. We can then form an estimate $\widehat{f_R}(t)$ by plugging in sample analogs in the right-hand side of (8.81):

$$\begin{aligned} \widehat{f_R}(t) &\approx \frac{\#(t-h, t+h))/n}{2h} && (8.83) \\ &= \frac{\#(t-h, t+h))}{2hn} && (8.84) \end{aligned}$$

where the notation $\#(a, b)$ means the number of R_i in the interval (a,b).

There is an important issue of how to choose the value of h here, but let's postpone that for now. For the moment, let's take

$$h = \frac{\max_i R_i - \min_i R_i}{100} \tag{8.85}$$

i.e. take h to be 0.01 of the range of our data.

At this point, we'd then compute (8.84) at lots of different points t. Although it would seem that theoretically we must compute (8.84) at infinitely many such points, the graph of the function is actually a step function. Imagine t moving to the right, starting at $\min_i R_i$. The interval $(t-h, t+h)$ moves along with it. Whenever the interval moves enough to the right to either pick up a new R_i or lose one that it had had, (8.84) will change value, but not at any other time. So, we only need to evaluate the function at about $2n$ values of t.

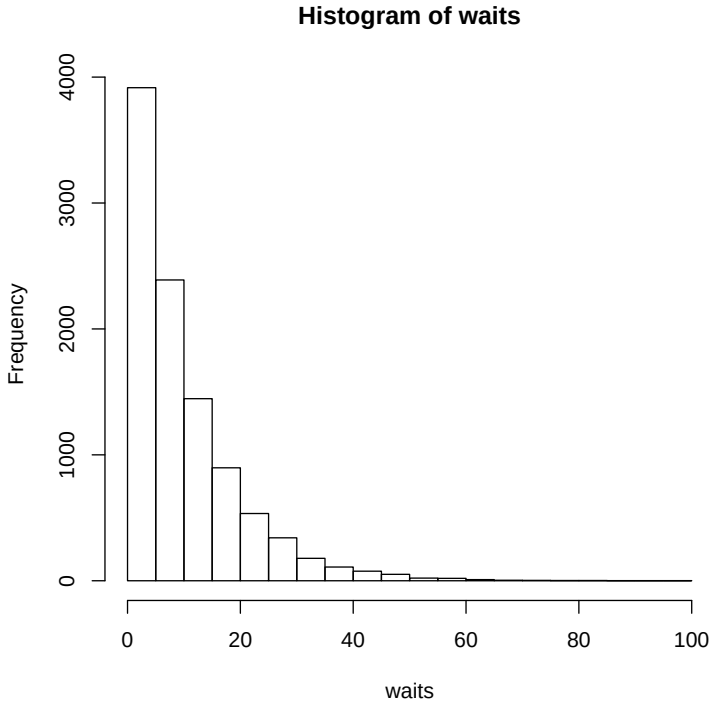
8.4.3 Histograms

If for some reason we really want to save on computation, let's say that we first break the interval $(\min_i R_i, \max_i R_i$ into 100 subintervals of size h given by (8.85). We then compute (8.84) only at the midpoints of those intervals, and pretend that the graph of $\widehat{f}_R(t)$ is constant within each subinterval. Do you know what we get from that? A histogram! Yes, a histogram is a form of density estimation. (Usually a histogram merely displays counts. We do so here too, but we have scaled things so that the total area under the curve is 1.)

Let's see how this works with our Bus Paradox simulation. We'll use R's **hist()** to draw a histogram. First, here's our simulation code:

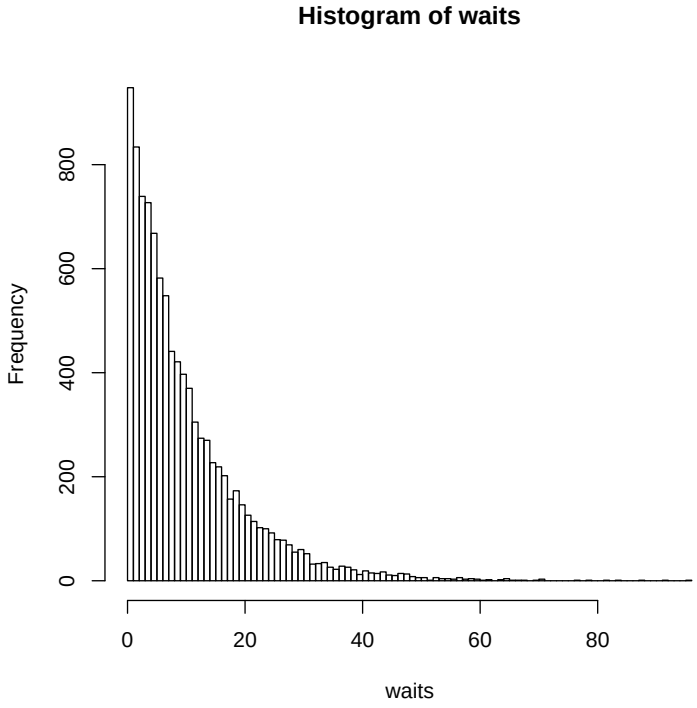
```
1  doexpt <- function(opt) {
2     lastarrival <- 0.0
3     while (TRUE) {
4        newlastarrival = lastarrival + rexp(1,0.1)
5        if (newlastarrival > opt)
6           return(opt-lastarrival)
7        else lastarrival <- newlastarrival
8     }
9  }
10
11 observationpt <- 240
12 nreps <- 10000
13 waits <- vector(length=nreps)
14 for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
15 hist(waits)
```

Note that I used the default number of intervals, 20. Here is the result:



The density seems to have a shape like that of the exponential parametric family. (This is not surprising, because it *is* exponential, but remember we're pretending we don't know that.)

Here is the plot with 100 intervals:



Again, a similar shape, though more raggedy.

8.4.4 Kernel-Based Density Estimation

No matter what the interval width is, the histogram will consist of a bunch of rectanges, rather than a curve. That is basically because, for any particular value of t, $\widehat{f_X(t)}$, depends only on the X_i that fall into that interval. We could get a smoother result if we used all our data to estimate $f_X(t)$ but put more weight on the data that is closer to t. One way to do this is called **kernel-based** density estimation, which in R is handled by the function **density()**.

We need a set of weights, more precisely a weight function k, called the **kernel**. Any nonnegative function which integrates to 1—i.e. a density function in its own right—will work. Our estimator is then

$$\widehat{f_R}(t) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{t - R_i}{h}\right) \tag{8.86}$$

To make this idea concrete, take k to be the uniform density on (-1,1), which has the value 0.5 on (-1,1) and 0 elsewhere. Then (8.86) reduces to (8.84). Note how the parameter h, called the **bandwidth**, continues to control how far away from to t we wish to go for data points.

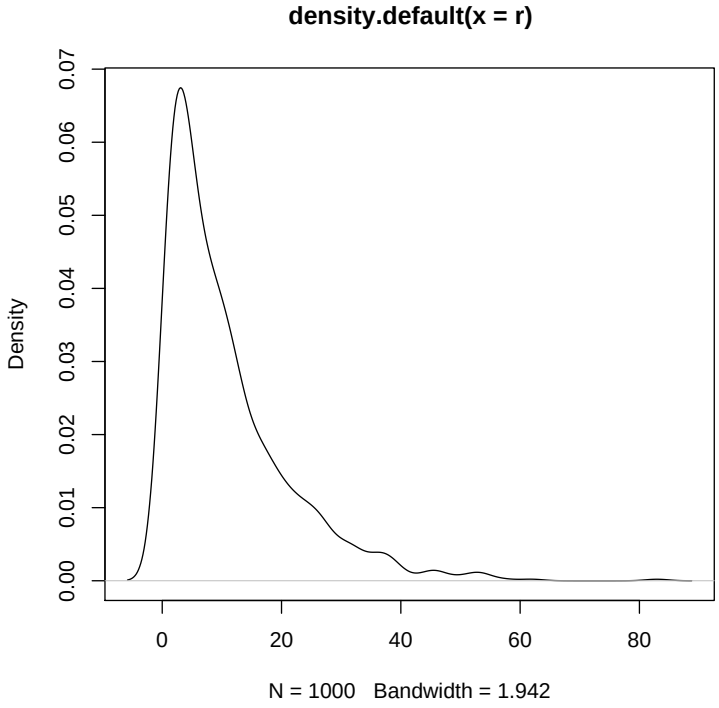


Figure 8.1: Kernel estimate, default bandwidth

But as mentioned, what we really want is to include all data points, so we typically use a kernel with support on all of $(-\infty, \infty)$. In R, the default kernel is that of the N(0,1) density. The bandwidth h controls how much smoothing we do; smaller values of h place heavier weights on data points near t and much lighter weights on the distant points. The default bandwidth in R is taken to the the standard deviation of k.

For our data here, I took the defaults:

```
plot(density(r))
```

The result is seen in Figure 8.1.

I then tried it with a bandwidth of 0.5. See Figure 8.2. This curve oscillates a lot, so an analyst might think 0.5 is too small. (We are prejudiced here, because we know the true population density is exponential.)

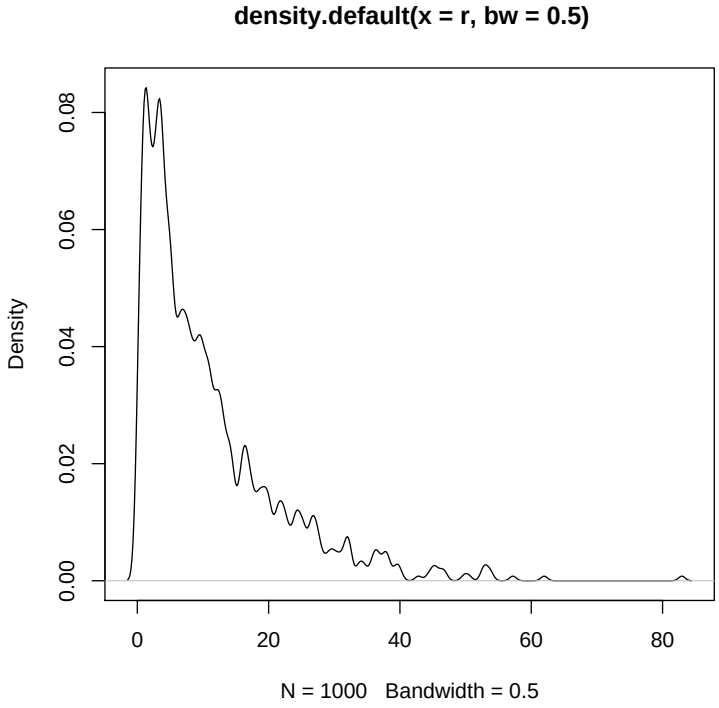


Figure 8.2: Kernel estimate, bandwidth 0.5

8.4.5 Proper Use of Density Estimates

There is no good, practical way to choose a good bin width or bandwdith. Moreover, there is also no good way to form a reasonable confidence band for a density estimate.

So, density estimates should be used as exploratory tools, not as firm bases for decision making. You will probably find it quite unsettling to learn that there is no exact answer to the problem. But that's real life!

8.5 Slutsky's Theorem

(The reader should review Section 4.4.3.8 before continuing.)

Since one generally does not know the value of σ in (7.20), we replace it by s, yielding (7.21). Why was that legitimate?

The answer depends on the theorem below. First, we need a definition.

Definition 29 *We say that a sequence of random variables L_n* **converges in probability** *to the random variable L if for every $\epsilon > 0$,*

$$\lim_{n \to \infty} P(|L_n - L| > \epsilon) = 0 \tag{8.87}$$

This is a little weaker than convergence with probability 1, as in the Strong Law of Large Numbers (SLLN, Section 3.19). Convergence with probability 1 implies convergence in probability but not vice versa.

So for example, if $Q_1, Q_2, Q_3, ...$ are i.i.d. with mean ω, then the SLLN implies that

$$L_n = \frac{Q_1 + ... + Q_n}{n} \tag{8.88}$$

converges with probability 1 to ω, and thus L_n converges in probability to ω too.

8.5.1 The Theorem

Theorem 30 Slutsky's Theorem *(abridged version): Consider random variables X_n, Y_n, and X, such that X_n converges in distribution to X and Y_n converges in probability to a constant c with probability 1,*

Then:

(a) $X_n + Y_n$ converges in distribution to $X + c$.

(b) X_n / Y_n converges in distribution to X/c.

8.5.2 Why It's Valid to Substitute s for σ

We now return to the question raised above. In our context here, that we take

$$X_n = \frac{\overline{W} - \mu}{\sigma / \sqrt{n}} \tag{8.89}$$

$$Y_n = \frac{s}{\sigma} \tag{8.90}$$

We know that (8.89) converges in distribution to N(0,1) while (8.90) converges in to 1. Thus for large n, we have that

$$\frac{\overline{W} - \mu}{s/\sqrt{n}} \tag{8.91}$$

has an approximate N(0,1) distribution, so that (7.21) is valid.

8.5.3 Example: Confidence Interval for a Ratio Estimator

Again consider the example in Section 7.2.3.1 of weights of men and women in Davis, but this time suppose we wish to form a confidence interval for the *ratio* of the means,

$$\gamma = \frac{\mu_1}{\mu_2} \tag{8.92}$$

Again, the natural estimator is

$$\widehat{\gamma} = \frac{\overline{X}}{\overline{Y}} \tag{8.93}$$

How can we construct a confidence interval from this estimator? If it were a linear combination of $\overline{X}$ and $\overline{Y}$, we'd have no problem, since a linear combination of multivariate normal random variables is again normal.

That is not exactly the case here, but it's close. Since $\overline{Y}$ converges in probability to μ_2, Slutsky's Theorem (Section 8.5) tells us that the problem here really is one of such a linear combination. We can form a confidence interval for μ_1, then divide both endpoints of the interval by $\overline{Y}$, yielding a confidence interval for γ.

8.6 The Delta Method: Confidence Intervals for General Functions of Means or Proportions

The **delta method** is a great way to derive asymptotic distributions of quantities that are functions of random variables whose asymptotic distributions are already known.

8.6.1 The Theorem

Theorem 31 *Suppose $R_1, ..., R_k$ are estimators of $\eta_1, ..., \eta_k$ based on a random sample of size n. Let R denote the vector whose components are the R_i, and let η denote the corresponding vector for the η_i. Suppose the random vector*

$$\sqrt{n}(R - \eta) = \sqrt{n} \begin{pmatrix} R_1 - \eta_1 \\ R_2 - \eta_2 \\ ... \\ R_k - \eta_k \end{pmatrix} \tag{8.94}$$

is known to have an asymptotically multivariate normal distribution with mean 0 and nonsingular covariance matrix $\Sigma = (\sigma_{ij})$.

Let h be a smooth scalar function[5] of k variables, with h_i denoting its i^{th} partial derivative. Consider the random variable

$$Y = h(R_1, ..., R_k) \tag{8.96}$$

Then $\sqrt{n}[Y - h(\eta_1, ..., \eta_k)]$ converges in distribution to a normal distribution with mean 0 and variance

$$[\nu_1, ..., \nu_k]' \Sigma [\nu_1, ..., \nu_k] \tag{8.97}$$

provided not all of

$$\nu_i = h_i(\eta_1, ..., \eta_k),\ i = 1, ..., k \tag{8.98}$$

are 0.

Informally, the theorem says, with R, η, Σ, h() and Y defined above:

[5]The word "smooth" here refers to mathematical conditions such as existence of derivatives, which we will not worry about here.

Similarly, the reason that we multiply by $\sqrt{n}$ is also due to theoretical considerations we will not go into here, other than to note that it is related to the formal statement of the Central Limit Theorem in Section 4.4.3.8. If we replace $X_1 + ..., +X_n$ in (4.54), by $n\overline{X}$, we get

$$Z = \sqrt{n} \cdot \frac{\overline{X} - m}{v} \tag{8.95}$$

> Suppose R is asymptotically multivariate normally distributed with mean η and covariance matrix Σ/n. Y will be approximately normal with mean $h(\eta_1, ..., \eta_k)$ and covariance matrix 1/n times (8.97).

Note carefully that the theorem is not saying, for example, that $E[h(R) = h(\eta)$ for fixed, finite n, which is not true. Nor is it saying that h(R) is normally distributed, which is definitely not true; recall for instance that if X has a N(0,1) distribution, then X^2 has a chi-square distribution with one degree of freedom, hardly the same as N(0,1). But the theorem says that for the purpose of asymptotic distributions, we can operate as if these things were true.

The theorem can be used to form confidence intervals for $h(\eta_1, ..., \eta_k)$, because it provides us with a standard error (Section 7.2.4):

$$\text{std. err. of } h(R) = \sqrt{\frac{1}{n} [\nu_1, ..., \nu_k]' \Sigma [\nu_1, ..., \nu_k]} \tag{8.99}$$

Of course, these quantities are typically estimated from the sample, e.g.

$$\widehat{\nu}_i = h_i(R_1, ..., R_k) \tag{8.100}$$

So, our approximate 95% confidence interval for $h(\eta_1, ..., \eta_k)$ is

$$h(R_1, ..., R_k) \pm 1.96 \sqrt{\frac{1}{n} [\widehat{\nu}_1, ..., \widehat{\nu}_k]' \widehat{\Sigma} [\widehat{\nu}_1, ..., \widehat{\nu}_k]} \tag{8.101}$$

Note that here we are considering scalar functions h(), but the theorem can easily be extended to vector-valued h().

Now, how is theorem derived?

Proof

We'll cover the case k = 1 (dropping the subscript 1 for convenience).

The intuitive version of the proof cites the fact from calculus[6] that a curve is close to its tangent line if we are close to the point of tangency. Here that means

$$h(R) \approx h(\eta) + h'(\eta)(R - \eta) \tag{8.102}$$

[6] This is where the "delta" in the name of the method comes from, an allusion to the fact that derivatives are limits of difference quotients.

if R is near η, which will be the case for large n. Note that in the right-hand side of (8.102), the only random quantity is R; the rest are constants. In other words, the right-hand side has the form c+dQ, where Q is approximately normal. Since a linear function of a normally distributed random variable itself has a normal distribution, (8.102) implies that h(R) is approximately normal with mean $h(\eta)$ and variance $[h'(\eta)]^2 Var(R)$.

Reasoning more carefully, recall the Mean Value Theorem from calculus:

$$h(R) = h(\eta) + h'(W)(R - \eta) \tag{8.103}$$

for some W between η and R. Rewriting this, we have

$$\sqrt{n}[h(R) - h(\eta)] = \sqrt{n}\, h'(W)(R - \eta) \tag{8.104}$$

It can be shown—and should be intuitively plausible to you—that if a sequence of random variables converges in distribution to a constant, the convergence is in probability too. So, $R - \eta$ converges in probability to 0, forcing W to converge in probability to $h(\eta)$. Then from Slutsky's Theorem, the asymptotic distribution of (8.104) is the same as that of $\sqrt{n}\, h'(\eta)(R - \eta)$. The result follows.

■

8.6.2 Example: Square Root Transformation

Here is an example of the delta method with k = 1. It will be a rather odd example, in that our goal is actually not to form a confidence interval for anything, but it will illustrate how the delta method is used.

It is used to be common, and to some degree is still common today, for statistical analysts to apply a square-root transformation to Poisson data. The delta method sheds light on the motivation for this, as follows.

First, note that we cannot even apply the delta method unless we have approximately normally distributed inputs, i.e. the R_i in the theorem. But actually, any Poisson-distributed random variable T is approximately normally distributed if its mean, λ, is large. To see this, recall from Section 5.11.4.2 that sums of independent Poisson random variables are themselves Poisson distributed. So, if for instance, ET is an integer k, then T has the same distribution as

$$U_1 + ... + U_m \tag{8.105}$$

where the U_i are i.i.d. Poisson random variables each having mean 1. By the Central Limit Theorem, T then has an approximate normal distribution, with mean and variance λ. (This is not quite a rigorous argument, so our treatment here is informal.)

Now that we know that T is approximately normal, we can apply the delta method. So, what h() should we use? The pioneers of statistics chose $h(t) = \sqrt{t}$. Let's see why.

Set $Y = h(T) = \sqrt{T}$ (so that T is playing the role of R in the theorem). Here η is $ET = \lambda$.

We have $h'(t) = 1/(2\sqrt{t})$. Then the delta method says that since T is approximately normally distributed with mean λ and variance λ, Y too has an approximate normal distribution, with mean

$$h(\eta) = \sqrt{\lambda} \tag{8.106}$$

What about the variance? Well, in one dimension, (8.97) reduces to

$$\nu^2 Var(R) \tag{8.107}$$

so we have

$$[h'(\eta)]^2 Var(R) = \left(\frac{1}{2\sqrt{t}} \Big|_{t=\lambda} \right)^2 \cdot \lambda = \frac{1}{4\lambda} \lambda = \frac{1}{4} \tag{8.108}$$

So, the (asymptotic) variance of $\sqrt{T}$ is a constant, independent of λ, and we say that the square root function is a **variance stabilizing transformation.** This becomes relevant in regression analysis, where, as we will discuss in Chapter 10, a classical assumption is that a certain collection of random variables all have the same variance. If those random variables are Poisson-distributed, then their square roots will all have approximately the same variance.

8.6.3 Example: Confidence Interval for σ^2

Recall that in Section 8.2.2 we noted that (7.22) is only an approximate confidence interval for the mean. An exact interval is available using the Student t-distribution, if the population is normally distributed. We pointed out that (7.22) is very close to the exact interval for even moderately large n anyway, and since no population is exactly normal, (7.22) is good enough. Note that one of the implications of this and the fact that (7.22) did not assume any particular population distribution is that a Student-t based confidence interval works well even for non-normal populations. We say that the Student-t interval is **robust** to the normality assumption.

But what about a confidence interval for a variance? It can be shown that one can form an exact interval based on the chi-square distribution, if the population is normal. In this case, though, the interval does NOT work well for non-normal populations; it is NOT robust to the normality assumption. So, let's derive an interval that doesn't assume normality; we'll use the delta method. (Warning: This will be a lengthy derivation, but it will cause you to review many concepts, which is good.)

As before, say we have $W_1, ..., W_n$, a random sample from our population, and with W representing a random variable having the population distribution.) Write

$$\sigma^2 = E(W^2) - (EW)^2 \tag{8.109}$$

and from (7.18) write our estimator of σ^2 as

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} W_i^2 - \overline{W}^2 \tag{8.110}$$

This suggests how we can use the delta method. We define

$$R_1 = \overline{W} \tag{8.111}$$

$$R_2 = \frac{1}{n} \sum_{i=1}^{n} W_i^2 \tag{8.112}$$

R_1 is an estimator of EW, and R_2 estimates $E(W^2)$. Furthermore, we'll see below that R_1 and R_2 are approximately bivariate normal, by the multivariate Central Limit Theorem, so we can use the delta method.

And most importantly, our estimator of interest, s^2, is a function of R_1 and R_2:

$$s^2 = R_2 - R_1^2 \tag{8.113}$$

So, we take our function h to be

$$h(u, v) = -u^2 + v \tag{8.114}$$

Now we must find Σ in the theorem. That means we'll need we'll need the covariance matrix of R_1 and R_2. But since

$$\begin{pmatrix} R_1 \\ R_2 \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} W_i \\ W_i^2 \end{pmatrix} \tag{8.115}$$

we can derive the covariance matrix of R_1 and R_2, as follows.

Remember, the covariance matrix is the multidimensional analog of variance. So, after reviewing the reasoning in (7.13), we have in the vector-valued version of that derivation that

$$
\begin{aligned}
Cov\left[\begin{pmatrix} R_1 \\ R_2 \end{pmatrix}\right] &= \frac{1}{n^2} Cov\left[\sum_{i=1}^{n} \begin{pmatrix} W_i \\ W_i^2 \end{pmatrix}\right] && (8.116)\\
&= \frac{1}{n^2} \sum_{i=1}^{n} Cov\left[\begin{pmatrix} W_i \\ W_i^2 \end{pmatrix}\right] && (8.117)\\
&= \frac{1}{n^2} \sum_{i=1}^{n} Cov\left[\begin{pmatrix} W \\ W^2 \end{pmatrix}\right] && (8.118)\\
&= \frac{1}{n} Cov\left[\begin{pmatrix} W \\ W^2 \end{pmatrix}\right] && (8.119)
\end{aligned}
$$

So

$$\Sigma = Cov\left[\begin{pmatrix} W \\ W^2 \end{pmatrix}\right] \tag{8.120}$$

Now we must estimate Σ. Taking sample analogs of (5.109), we set

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} W_i \\ W_i^2 \end{pmatrix} (W_i, W_i^2) - R\,R' = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} W_i^2 & W_i^3 \\ W_i^3 & W_i^4 \end{pmatrix} - R\,R' \tag{8.121}$$

where $R = (R_1, R_2)'$.

Also, $h'(u,v) = (-2u, 1)'$, so

$$h'(R_1, R_2) = (-2R_1, 1)' \tag{8.122}$$

Whew! We're done. We can now plug everything into (8.101).

Note that all these quantities are expressions in $E(W^k)$ for various k. It should be noted that estimating means of higher powers of a random variable requires larger samples in order to achieve comparable accuracy. Our confidence interval here may need a rather large sample to be accurate, as opposed to the situation with (7.22), in which even n = 20 should work well.

8.6.4 Example: Confidence Interval for a Measurement of Prediction Ability

Suppose we have a random sample $X_1, ..., X_n$ from some population. In other words, the X_i are independent and each is distributed as in the population. Let X represent a generic random variable having that distribution. Here we are allowing the X_i and X to be random vectors, though they won't play much explicit role anyway.

Let A and B be events associated with X. If for example X is a random vector (U,V), we might have A and B being the events U > 12 and U-V < 5. The question of interest here will be to what extent we can predict A from B.

One measure of that might be the quantity $\nu = P(A|B) - P(A)$. The larger ν is (in absolute value), the stronger the ability of B to predict A. (We could look at variations of this, such as the quotient of those two probabilities, but will not do so here.)

Let's use the delta method to derive an approximate 95% confidence interval for ν. To that end, think of four categories—A and B; A and not B; not A and B; and not A and not B. Each X_i falls into one of those categories, so the four-component vector Y consisting of counts of the numbers of X_i falling into the four categories has a multinomial distribution with r = 4.

To use the theorem, set R = Y/n, so that R is the vector of the sample proportions. For instance, R_1 will be the number of X_i satisfying both events A and B, divided by n. The vector η will then be the corresponding population proportion, so that for instance

$$\eta_2 = P(A \text{ and not } B) \tag{8.123}$$

We are interested in

$$\begin{aligned} \nu &= P(A|B) - P(A) && (8.124)\\ &= \frac{P(A \text{ and } B)}{P(A \text{ and } B) + P(\text{not } A \text{ and } B)} - [P(A \text{ and } B) + P(A \text{ and not } B)] && (8.125)\\ &= \frac{\eta_1}{\eta_1 + \eta_3} - (\eta_1 + \eta_2) && (8.126) \end{aligned}$$

By the way, since η_4 is not involved, let's shorten R to $(R_1, R_2, R_3)'$.

What about Σ? Since Y is multinomial, Equation (5.153) provides us Σ:

$$\Sigma = \frac{1}{n} \begin{pmatrix} \eta_1(1-\eta_1) & -\eta_1\eta_2 & -\eta_1\eta_3 \\ -\eta_2\eta_1 & \eta_2(1-\eta_2) & -\eta_2\eta_3 \\ -\eta_3\eta_1 & -\eta_3\eta_2 & \eta_3(1-\eta_3) \end{pmatrix} \tag{8.127}$$

We then get $\widehat{\Sigma}$ by substituting R_i for η_i. After deriving the $\widehat{\nu}_i$ from (8.124), we make the same substitution there, and then compute (8.101).

8.7 Simultaneous Confidence Intervals

Suppose in our study of heights, weights and so on of people in Davis, we are interested in estimating a number of different quantities, with our forming a confidence interval for each one. Though our confidence level for each one of them will be 95%, our *overall* confidence level will be less than that. In other words, we cannot say we are 95% confident that all the intervals contain their respective population values.

In some cases we may wish to construct confidence intervals in such a way that we can say we are 95% confident that all the intervals are correct. This branch of statistics is known as **simultaneous inference** or **multiple inference**.

Usually this kind of methodology is used in the comparison of several **treatments**. This term originated in the life sciences, e.g. comparing the effectiveness of several different medications for controlling hypertension, it can be applied in any context. For instance, we might be interested in comparing how well programmers do in several different programming languages, say Python, Ruby and Perl. We'd form three groups of programmers, one for each language, with say 20 programmers per group. Then we would have them write code for a given application. Our measurement could be the length of time T that it takes for them to develop the program to the point at which it runs correctly on a suite of test cases.

Let T_{ij} be the value of T for the jth programmer in the ith group, i = 1,2,3, j = 1,2,...,20. We would then wish to compare the three "treatments," i.e. programming languages, by estimating $\mu_i = ET_{i1}$, i = 1,2,3. Our estimators would be $U_i = \sum_{j=1}^{20} T_{ij}/20$, i = 1,2,3. Since we are comparing the three population means, we may not be satisfied with simply forming ordinary 95% confidence intervals for each mean. We may wish to form confidence intervals which *jointly* have confidence level 95%.[7]

Note very, very carefully what this means. As usual, think of our notebook idea. Each line of the notebook would contain the 60 observations; different lines would involve different sets of 60 people. So, there would be 60 columns for the raw data, three columns for the U_i. We would also have six more columns for the confidence intervals (lower and upper bounds) for the μ_i. Finally, imagine three more columns, one for each confidence interval, with the entry for each being either Right or Wrong. A confidence interval is labeled Right if it really does contain its target population value, and otherwise is labeled Wrong.

Now, if we construct individual 95% confidence intervals, that means that in a given Right/Wrong column, in the long run 95% of the entries will say Right. But for simultaneous intervals, we hope that within a line we see three Rights, and 95% of all lines will have that property.

In our context here, if we set up our three intervals to have individual confidence levels of 95%, their

[7] The word *may* is important here. It really is a matter of philosophy as to whether one uses simultaneous inference procedures.

simultaneous level will be $0.95^3 = 0.86$, since the three confidence intervals are independent. Conversely, if we want a simultaneous level of 0.95, we could take each one at a 98.3% level, since $0.95^{\frac{1}{3}} \approx 0.983$.

However, in general the intervals we wish to form will not be independent, so the above "cube root method" would not work. Here we will give a short introduction to more general procedures.

Note that "nothing in life is free." If we want simultaneous confidence intervals, they will be wider.

Another reason to form simultaneous confidence intervals is that it gives you "license to browse," i.e. to rummage through the data looking for interesting nuggets.

8.7.1 The Bonferonni Method

One simple approach is **Bonferonni's Inequality**:

Lemma 32 *Suppose $A_1, ..., A_g$ are events. Then*

$$P(A_1 \text{ or } ... \text{ or} A_g) \leq \sum_{i=1}^{g} P(A_i) \tag{8.128}$$

You can easily see this for g = 2:

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2) - P(A_1 \text{ and } A_2) \leq P(A_1) + P(A_2) \tag{8.129}$$

One can then prove the general case by mathematical induction.

Now to apply this to forming simultaneous confidence intervals, take A_i to be the event that the i^{th} confidence interval is incorrect, i.e. fails to include the population quantity being estimated. Then (8.128) says that if, say, we form two confidence intervals, each having individual confidence level (100-5/2)%, i.e. 97.5%, then the overall collective confidence level for those two intervals is at least 95%. Here's why: Let A_1 be the event that the first interval is wrong, and A_2 is the corresponding event for the second interval. Then

$$\begin{aligned}
\text{overall conf. level} &= P(\text{not } A_1 \text{ and not } A_2) && (8.130)\\
&= 1 - P(A_1 \text{ or } A_2) && (8.131)\\
&\geq 1 - P(A_1) - P(A_2) && (8.132)\\
&= 1 - 0.025 - 0.025 && (8.133)\\
&= 0.95 && (8.134)
\end{aligned}$$

8.7.2 Scheffe's Method

The Bonferonni method is unsuitable for more than a few intervals; each one would have to have such a high individual confidence level that the intervals would be very wide. Many alternatives exist, a famous one being **Scheffe's method**.[8]

Theorem 33 *Suppose $R_1, ..., R_k$ have an approximately multivariate normal distribution, with mean vector $\mu = (\mu_i)$ and covariance matrix $\Sigma = (\sigma_{ij})$. Let $\widehat{\Sigma}$ be a* **consistent** *estimator of Σ, meaning that it converges in probability to Σ as the sample size goes to infinity.*

For any constants $c_1, ..., c_k$, consider linear combinations of the R_i,

$$\sum_{i=1}^{k} c_i R_i \tag{8.135}$$

which estimate

$$\sum_{i=1}^{k} c_i \mu_i \tag{8.136}$$

Form the confidence intervals

$$\sum_{i=1}^{k} c_i R_i \pm \sqrt{k \chi^2_{\alpha;k}} s(c_1, ..., c_k) \tag{8.137}$$

where

$$[s(c_1, ..., c_k)]^2 = (c_1, ..., c_k)^T \widehat{\Sigma} (c_1, ..., c_k) \tag{8.138}$$

and where $\chi^2_{\alpha;k}$ is the upper-α percentile of a chi-square distribution with k degrees of freedom.[9]

Then all of these intervals (for infinitely many values of the c_i!) have simultaneous confidence level $1 - \alpha$.

By the way, if we are interested in only constructing confidence intervals for **contrasts**, i.e. c_i having the property that $\Sigma_i c_i = 0$, we the number of degrees of freedom reduces to k-1, thus producing narrower intervals.

[8] The name is pronounced "sheh-FAY."

[9] Recall that the distribution of the sum of squares of g independent N(0,1) random variables is called **chi-square with g degrees of freedom**. It is tabulated in the R statistical package's function **qchisq()**.

Just as in Section 8.2.2 we avoided the t-distribution, here we have avoided the F distribution, which is used instead of ch-square in the "exact" form of Scheffe's method.

8.7.3 Example

For example, again consider the Davis heights example in Section 7.2.6. Suppose we want to find approximate 95% confidence intervals for two population quantities, μ_1 and μ_2. These correspond to values of c_1, c_2 of (1,0) and (0,1). Since the two samples are independent, $\sigma_{12} = 0$. The chi-square value is 5.99,[10] so the square root in (8.137) is 3.46. So, we would compute (7.22) for $\overline{X}$ and then for $\overline{Y}$, but would use 3.46 instead of 1.96.

This actually is not as good as Bonferonni in this case. For Bonferonni, we would find two 97.5% confidence intervals, which would use 2.24 instead of 1.96.

Scheffe's method is too conservative if we just are forming a small number of intervals, but it is great if we form a lot of them. Moreover, it is very general, usable whenever we have a set of approximately normal estimators.

8.7.4 Other Methods for Simultaneous Inference

There are many other methods for simultaneous inference. It should be noted, though, that many of them are limited in scope, in contrast to Scheffe's method, which is usable whenever one has multivariate normal estimators, and Bonferonni's method, which is universally usable.

8.8 The Bootstrap Method for Forming Confidence Intervals

Many statistical applications can be quite complex, which makes them very difficult to analyze mathematically. Fortunately, there is a fairly general method for finding confidence intervals called the **bootstrap**. Here is a brief overview of the type of bootstrap confidence interval construction called **Efron's percentile method**.

8.8.1 Basic Methodology

Say we are estimating some population value θ based on i.i.d. random variables Q_i, i = 1,...,n. Note that θ and the Q_i could be vector-valued.

[10] Obtained from R via **qchisq(0.95,2)**.

Our estimator of θ is of course some function of the Q_i, $h(Q_1, ..., Q_n)$. For example, if we are estimating a population mean by a sample mean, then the function h() is defined by

$$h(u_1, ..., u_n) = \frac{u_1 + ..., +u_n}{n} \tag{8.139}$$

Our procedure is as follows:

- Estimate θ based on the original sample, i.e. set

$$\widehat{\theta} = h(Q_1, ..., Q_n) \tag{8.140}$$

- For j = 1,2,...,k:
  - Resample, i.e. create a new "sample," $\widetilde{Q}_1, .., \widetilde{Q}_n$, by drawing n times with replacement from $Q_1, .., Q_n$.
  - Calculate the value of $\widehat{\theta}$ based on the $\widetilde{Q}_i$ instead of the Q_i, i.e. set

$$\widetilde{\theta}_j = h(\widetilde{Q}_1, ..., \widetilde{Q}_n) \tag{8.141}$$

- Sort the values $\widetilde{\theta}_j$, j = 1,...,k, and let $\widetilde{\theta}_{(k)}$ be the kth-smallest value.
- Let A and B denote the 0.025 and 0.975 quantiles of the $\widetilde{\theta}_j - \widehat{\theta}$, i.e.

$$A = \widehat{\theta}_{(0.025n)} - \widehat{\theta} \text{ and } B = \widehat{\theta}_{(0.975n)} - \widehat{\theta} \tag{8.142}$$

  (The quantities 0.025n and 0.975n must be rounded, say to the nearest integer in the range 1,...,n.)
- Then your approximate 95% confidence interval for θ is

$$(\widehat{\theta} - B, \widehat{\theta} - A) \tag{8.143}$$

8.8.2 Example: Confidence Intervals for a Population Variance

As noted in Section 8.6.3, the classical chi-square method for finding a confidence interval for a population variance σ^2 is not robust to the assumption of a normally distributed parent population. In that section, we showed how to find the desired confidence interval using the delta method.

That was a solution, but the derivation was complex. An alternative would be to use the bootstrap. We resample many times, calculate the sample variance on each of the new samples, and then form a confidence interval for σ^2 as in (8.142). We show the details using R in Section 8.8.3

8.8.3 Computation in R

R includes the **boot()** function to do the mechanics of this for us. To illustrate its usage, let's consider finding a confidence interval for the population variance σ^2, based on the sample variance, s^2. Here is the code:

```
# R base doesn't include the boot package, so must load it
library(boot)

# finds the sample variance on x[c(inds)]
s2 <- function(x,inds) {
   return(var(x[inds]))
}

bt <- boot(x,s2,R=200)
cilow[rep] <- quantile(bt$t,alp)
cihi[rep] <- quantile(bt$t,1-alp)

print(mean(cilow <= 1.0 & 1.0 <= cihi))
```

How does this work? The line

```
bt <- boot(x,s2,R=200)
```

instructs R to apply the bootstrap to the data set **x**, with the statistic of interest being specified by the user in the function **s2()**. The argument **R** here is what we called k in Section 8.8.1 above, i.e. the number of times we resample n items from **x**.

Our argument **inds** in **s2()** is less obvious. Here's what happens: As noted, the **boot()** function merely shortens our work. Without it, we could simply call **sample()** to do our resampling. Say for simplicity that n is 4. We might make the call

```
j <- sample(1:4,replace=T)
```

and **j** might turn out to be, say, c(4,1,3,3). We would then apply the statistic to be bootstrapped, in our case here the sample variance, to the data $x[4], x[1], x[3], x[3]$—more compactly and efficiently expressed as $x[c(4, 1, 3, 3)]$. That's what **boot()** does for us. So, in our example above, the argument **inds** would be c(4,1,3,3) here.

In the example here, our statistic to be bootstrapped was a very common one, and thus there was already an R function for it, **var()**. In more complex settings, we'd write our own function.

8.8.4 General Applicability

Much theoretical work has been done on the bootstrap, and it is amazingly general. It has become the statistician's "Swiss army knife." However, there are certain types of estimators on which the bootstrap

fails. How can one tell in general?

One approach would be to consult the excellent book *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley, Cambridge University Press, 1997.

But a simpler method would be to test the bootstrap in the proposed setting by simulation: Write R code to generate many samples; get a bootstrap confidence interval on each one; and then see whether the number of intervals containing the true population value is approximately 95%.

In the sample variance example above, the code could be:

```
sim <- function(n,nreps,alp) {
   cilow <- vector(length=nreps)
   cihi <- vector(length=nreps)
   for (rep in 1:nreps) {
      x <- rnorm(n)
      bt <- boot(x,s2,R=200)
      cilow[rep] <- quantile(bt$t,alp)
      cihi[rep] <- quantile(bt$t,1-alp)
   }
   print(mean(cilow <= 1.0 & 1.0 <= cihi))
}
```

8.8.5 Why It Works

The mathematical theory of the bootstrap can get extremely involved, but we can at least get a glimpse of why it works here.

First review notation:

- Our random sample data is $Q_1, ..., Q_n$.
- Our estimator of θ is $\widehat{\theta} = h(Q_1, ..., Q_n)$.
- Our resampled estimators of θ are $\widehat{\theta}_1, ..., \widehat{\theta}_k$.

Remember, to get any confidence interval from an estimator, we need the distribution of that estimator. Here in our bootstrap context, our goal is to find the approximate distribution of $\widehat{\theta}$. The bootstrap achieves that goal very simply.

In essence, we are performing a simulation, drawing samples from the empirical distribution function for our Q_i data. Since the empirical cdf is an estimate of the population cdf F_Q, then the $\widehat{\theta}_j$ act like a random sample from the resulting distribution of $\widehat{\theta}$.

Indeed, if we calculate the sample standard deviation ("s") of the $\widetilde{\theta}_j$, that is an estimate of the standard error of $\widehat{\theta}$. If due to the delta method or other considerations, we know that the asymptotic distribution of $\widehat{\theta}$ is normal, then an approximate 95% confidence interval for θ would be

$$\widehat{\theta} \pm 1.96 \times \text{ standard deviation of the } \widehat{\theta}_j \tag{8.144}$$

Efron's percentile method is more general, and works better for small samples. The idea is that the above discussion implies that the values

$$\widetilde{\theta}_j - \widehat{\theta} \tag{8.145}$$

have approximately the same distribution as the values

$$\widetilde{\theta} - \theta \tag{8.146}$$

Accordingly, the probability that (8.146) is between A and B is approximately 0.95, thus giving us (8.143).

8.9 Bayesian Methods

Everyone is entitled to his own opinion, but not his own facts—Daniel Patrick Moynihan, senator from New York, 1976-2000

Whiskey's for drinkin' and water's for fightin' over—Mark Twain, on California water jurisdiction battles

Black cat, white cat, it doesn't matter as long as it catches mice—Deng Xiaoping, when asked about his plans to give private industry a greater role in China's economy

The most controversial topic in statistics by far is that of **Bayesian** methods. In fact, it is so controversial that a strident Bayesian colleague of mine even took issue with my calling it "controversial"!

The name stems from Bayes' Rule (Section 2.6),

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)} \tag{8.147}$$

No one questions the validity of Bayes' Rule, and thus there is no controversy regarding statistical procedures that make use of probability calculations based on that rule. But the key word is *probability*. As long as the various terms in (8.147) are real probabilities, there is no controversy. But instead, the debate

stems from the cases in which Bayesians replace some of the probabilities in the theorem with "feelings," i.e. NON-probabilities, arising from what they call **subjective prior distributions**. The key word is then *subjective*. Our section here will concern the controversy over the use of subjective priors.

Say we wish to estimate a population mean. Here the Bayesian analyst, before even collecting data, says, "Well, I think the population mean could be 1.2, with probability, oh, let's say 0.28, but on the other hand, it might also be 0.88, with probability, well, I'll put it at 0.49..." etc. This is the analyst's subjective prior distribution for the population mean. The analyst does this before even collecting any data. Note carefully that he is NOT claiming these are real probabilities; he's just trying to quantify his hunches. The analyst then collects the data, and uses some mathematical procedure that combines these "feelings" with the actual data, and which then outputs an estimate of the population mean or other quantity of interest.

The Bayesians justify this by saying one should use all available information, even if it is just a hunch. "The analyst is typically an expert in the field under study. You wouldn't want to throw away his/her expertise, would you?"

The non-Bayesians, known as **frequentists**, on the other hand dismiss this as unscientific and lacking in impartiality. "In research on a controversial health issue, say, you wouldn't want the researcher to incorporate his/her personal political biases into the number crunching, would you?"

8.9.1 How It Works

To introduce the idea, consider again the example of estimating p, the probability of heads for a certain penny. Suppose we were to say—before tossing the penny even once—"I think p could be any number, but more likely near 0.5, something like a normal distribution with mean 0.5 and standard deviation, oh, let's say 0.1."[11] The prior distribution is then $N(0.5, 0.1^2)$. But again, note that the Bayesians do not consider it to be a distribution in the sense of probability. We are just using our "gut feeling" here, our "hunch." This is an absolutely central point.

So, Bayesians would not regard p as random here. They would simply be using the normal "distribution" for p to describe a degree of belief, rather than a probability distribution. (I will continue to use quotation marks below for this reason.)

Nevertheless, in terms of the mathematics involved, it's as if the Bayesians are treating p as random, with p's distribution being whatever the analyst specifies as the prior. Under this "random p" assumption, the Maximum Likelihood Estimate (MLE), for instance, would change. Our data here is X, the number of heads we get from n tosses of the penny. In contrast to the frequentist approach, in which the likelihood

[11] Of course, the true value of p is between 0 and 1, while the normal distribution extends from $-\infty$ to ∞. This, as noted in Section 4.4.3.9, the use of normal distributions is common for modeling many bounded quantities.

Nevertheless, many Bayesians prefer to use a beta distribution for the prior in this kind of setting.

would be

$$L = \left(\begin{array}{c} n \\ X \end{array} \right) p^X (1-p)^{n-X} \tag{8.148}$$

it now becomes

$$L = \frac{1}{\sqrt{2\pi}\ 0.1}\ \exp -0.5[(p-0.5)/0.1]^2 \left(\begin{array}{c} n \\ X \end{array} \right) p^X (1-p)^{n-X} \tag{8.149}$$

This is basically P(A and B) = P(A) P(B|A), though using a density rather than probability mass functions. We would then find the value of p which maximizes L, and take that as our estimate.

Note how this procedure achieves a kind of balance between what our hunch says and what our data say. In (8.149), suppose the mean of p is 0.5 but n = 20 and X = 12 Then the frequentist estimator would be X/n = 0.6, while the Bayes estimator would be about 0.56. (Computation not shown here.) So our Bayesian approach "pulled" our estimate away from the frequentist estimate, toward our hunch that p is at or very near 0.5. This pulling effect would be stronger for smaller n or for a smaller standard deviation of the prior "distribution."

A Bayesian would use Bayes' Rule to compute the "distribution" of p given X, called the **posterior distribution**. The analog of (8.147) would be (8.149) divided by the integral of (8.149) as p ranges from 0 to 1, with the resulting quotient then being treated as a density. The MLE would then be the **mode**, i.e. the point of maximal density of the posterior distribution.

But we could use any measure of central tendency, and in fact typically the mean is used, rather than the mode. In other words:

> To estimate a population value θ, the Bayesian constructs a prior "distribution" for θ (again, the quotation marks indicate that it is just a quantified gut feeling, rather than a real probability distribution). Then she uses the prior together with the actual observed data to construct the posterior distribution. Finally, she takes her estimate $\hat{\theta}$ to be the mean of the posterior distribution.

8.9.2 Extent of Usage of Subjective Priors

Though some academics are staunch, often militantly proselytizing Bayesians, only a small minority of statisticians in practice use the Bayesian approach. It is not mainstream.

One way to see that Bayesian methodology is not mainstream is through the R programming language. For example, as of December 2010, only about 65 of the more than 3000 packages on CRAN, the R repos-

itory, involve Bayesian techniques. (See `http://cran.r-project.org/web/packages/tgp/index.html`.) There is actually a book on the topic, *Bayesian Computation with R*, by Jim Albert, Springer, 2007, and among those who use Bayesian techniques, many use R for that purpose. However, almost all general-purpose books on R do not cover Bayesian methodology at all.

Significantly, even among Bayesian academics, many use frequentist methods when they work on real, practical problems. Choose an academic statistician at random, and you'll likely find on the Web that he/she does not use Bayesian methods when working on real applications.

8.9.3 Arguments Against Use of Subjective Priors

As noted, most professional statisticians, including me, are frequentists. What are the arguments made in this regard?

First, the following must be noted carefully:

> Ultimately, the use of any statistical analysis is to make a decision about something. This could be a very formal decision, such as occurs when the Food and Drug Administration (FDA) decides whether to approve a new drug, or it could be informal, for instance when an ordinary citizen reads a newspaper article reporting on a study analyzing data on traffic accidents, and she decides what to conclude from the study.
>
> Frequentists believe that there is nothing wrong using one's gut feelings to make a final decision, but it should not be part of the mathematical analysis of the data. One's hunches can play a role in deciding the "preponderance of evidence," as discussed in Section 7.4.4, but that should be kept separate from our data analysis.
>
> If for example the FDA's data shows the new drug to be effective, but at the same time the FDA scientists still have their doubts, they may decide to delay approval of the drug pending further study. So they can certainly act on their hunch, or on non-data information they have concerning the drug. But the FDA, as a public agency, has a responsibility to the citizenry to state what the data say, i.e. to report the frequentist estimate, rather than merely reporting a number—the Bayesian estimate—that mixes fact and hunch.

Thus, the Bayesian rallying cry, "It would be wrong to ignore any information we possess to supplement our data, even if that information is just a hunch," is presenting us with a false choice. The frequentists have never advocated ignoring hunches. However, they don't plug their hunches into a mathematical formula, just as jurors in a trial don't plug their hunches into a mathematical formula either.

The Bayesians say that in some cases, a Bayesian estimator may, for instance, produce smaller mean squared estimation error (recall Section 8.2.3) than its frequentist counterpart, even if the prior distribution was just

in our imaginations. But again, this argument is incorrectly implicitly presuming that frequentists ignore their hunches, which is not the case.

Moreover, in most applications of statistics, there is a need for impartial estimates. As noted above, even if the FDA acts on a hunch to delay approval of a drug in spite of favorable data, the FDA owes the public (and the pharmaceutal firm) an impartial report of what the data say. Bayesian estimation is by definition not impartial. One Bayesian statistician friend put it very well, saying "I believe my own subjective priors, but I don't believe those of other people." His statement should be considered by any potential user or consumer of Bayesian statistics.

Furthermore, in practice we are typically interested in inference, i.e. confidence intervals and significance tests, rather than just point estimation. We are sampling from populations, and want to be able to legitimately make inferences about those populations. For instance, though one can derive a Bayesian 95% confidence interval for p for our coin, it really has very little meaning, and again is certainly not impartial.

Consider the following scenario. Steven is running for president. Leo, his campaign manager, has commissioned Lynn to conduct a poll to assess Steven's current support among the voters. Lynn takes her poll, and finds that 57% of those polled support Steven. But her own gut feeling as an expert in politics, is that Steven's support is only 48%. She then combines these two numbers in some Bayesian fashion, and comes up with 50.2% as her estimate of Steven's support.

She then reports to Steven that she estimates Steven's support to be 50.2%. Leo asks Lynn how she arrived at that number, and she explains that she combined her prior distribution with the data. **But Leo then says, "Lynn, I really respect your political expertise, but I'd like you to tell me separately—what did the data say, and what is your own gut feeling? Lynn then tells Leo the two numbers, 57% and 48%, separately, and Leo finds both of them useful, in different senses.**

8.9.3.1 What Would You Do?

In evaluating the frequentist/Bayesian debate, you might wish to ask yourself what you would do in the following situations:

- As a personal investor, you've developed a statistical model for the day-to-day price variation of Google stock prices, and will use it to decide whether to buy the stock today. You wish to predict the price of the stock tomorrow, based on its price the last few days. Here are your choices:

  - As a frequentist, you could use a classical mathematical model, say regression analysis (Chapter 10), say fitting a linear or polynomial model. You could use the data to estimate the parameters of the model. This would give you a predicted price for tomorrow. Note that you can still choose to ignore that predicted price in the end, based on a hunch, but you've kept that hunch separate from your data analysis.

  - As a Bayesian, you might use the say linear or polynomial regression model, but you would specify a subjective prior distribution for the parameters. Your predicted price would then be affected by that subjective prior.

  So, what would you deem wise here—a frequentist or Bayesian approach?

- We are in a presidential election, complete with opinion polls as to who is currently winning. As an involved citizen, would you rather that the pollsters simply report the data as is, with their reported margin of error being computed from the traditional frequentist methods we've seen so far, or would you prefer that they factor in their own feelings via subjective priors?

  So, what would you deem wise here—a frequentist or Bayesian approach?

- You are a physician reading a medical journal article about the effectiveness of a certain drug for alleviating high blood pressure. Would you rather that the authors of the article simply report a straightforward analysis of the data, or would you prefer that the author incorporate a subjective prior distribution into his/her mathematical model?

  So, what would you deem wise here—a frequentist or Bayesian approach?

Exercises

Note to instructor: See the Preface for a list of sources of real data on which exercises can be assigned to complement the theoretical exercises below.

1. Consider raffle ticket example in Section 8.1.1. Suppose 500 tickets are sold, and you have data on 8 of them. Continue to assume sampling with replacement. Consider the Maximum Likelihood and Methods of Moments estimators.

(a) Find the probability that the MLE is exactly equal to the true value of c.

(b) Find the exact probability that the MLE is within 50 of the true value.

(c) Find the approximate probability that the Method of Moments estimator is within 50 of the true value.

2. Suppose I = 1 or 0, with probability p and 1-p, respectively. Given I, X has a Poisson distribution with mean λ_I. Suppose we have $X_1, ..., X_n$, a random sample of size n from the (unconditional) distribution of X. (We do not know the associated values of I, i.e. $I_1, ..., I_n$.) This kind of situation occurs in various applications. The key point is the effect of the unseen variable. In terms of estimation, note that there are three parameters to be estimated.

(a) Set up the likelihood function, which if maximized with respect to the three parameters would yield the MLEs for them.

(b) The words *if* and *would* in that last sentence allude to the fact that MLEs cannot be derived in closed form. However, R's **mle()** function can be used to find their values numerically. Write R code to do this. In other words, write a function with a single argument **x**, representing the X_i, and returning the MLEs for the three parameters.

3. Find the Method of Moments and Maximum Likelihood estimators of the following parameters in famous distribution families:

- p in the binomial family (n known)
- p in the geometric family
- μ in the normal family (σ known)
- λ in the Poisson family

4. For each of the following quantities, state whether the given estimator is unbiased in the given context:

(a) (4.15), p. 97, as an estimator of σ^2

(b) $\hat{p}$, as an estimator of p, p.105

(c) $\hat{p}(1-\hat{p})$, as an estimator of p(1-p), p.105

(d) $\bar{X} - \bar{Y}$, as an estimator of $\mu_1 - \mu_2$, p.107

(e) $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_1)^2$ (assuming μ_1 is known), as an estimator of σ_1^2, p.107

(f) $\bar{X}$, as an estimator of μ_1, p.107, *but sampling (from the population of Davis) without replacement*

5. Consider the Method of Moments Estimator $\hat{c}$ in the raffle example, Section 8.1.1. Find the exact value of $Var(\hat{c})$. Use the facts that $1 + 2 + ... + r = r(r+1)/2$ and $1^2 + 2^+ ... + r^2 = r(r+1)(2r+1)/6$.

6. Suppose W has a uniform distribution on (-c,c), and we draw a random sample of size n, $W_1, ..., W_n$. Find the Method of Moments and Maximum Likelihood estimators. (Note that in the Method of Moments case, the first moment won't work.)

7. An urn contains ω marbles, one of which is black and the rest being white. We draw marbles from the urn one at a time, without replacement, until we draw the black one; let N denote the number of draws needed. Find the Method of Moments estimator of ω based on X.

8. Suppose $X_1, ..., X_n$ are uniformly distributed on (0,c). Find the Method of Moments and Maximum Likelihood estimators of c, and compare their mean squared error.

Hint: You will need the density of $M = \max_i X_i$. Derive this by noting that $M \leq t$ if and only if $X_i \leq t$ for all i = 1,2,...,n.

9. Add a single line to the code on page 190 that will print out the estimated value of Var(W).

10. In the raffle example, Section 8.1.1, find a $(1-\alpha)\%$ confidence interval for c based on $\check{c}$, the Maximum Likelihood Estimate of c.

11. In many applications, observations come in correlated clusters. For instance, we may sample r trees at random, then s leaves within each tree. Clearly, leaves from the same tree will be more similar to each other than leaves on different trees.

In this context, suppose we have a random sample $X_1, ..., X_n$, n even, such that there is correlation within pairs. Specifically, suppose the pair (X_{2i+1}, X_{2i+2}) has a bivariate normal distribution with mean (μ, μ) and covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{8.150}$$

i = 0,...,n/2-1, with the n/2 pairs being independent. Find the Method of Moments estimators of μ and ρ.

12. Suppose we have a random sample $X_1, ..., X_n$ from some population in which $EX = \mu$ and $Var(X) = \sigma^2$. Let $\overline{X} = (X_1 + ... + X_n)/n$ be the sample mean. Suppose the data points X_i are collected by a machine, and that due to a defect, the machine always records the last number as 0, i.e. $X_n = 0$. Each of the other X_i is distributed as the population, i.e. each has mean μ and variance σ^2. Find the mean squared error of $\overline{X}$ as an estimator of μ, separating the MSE into variance and squared bias components as in Section 8.2.

13. Suppose we have a random sample $X_1, ..., X_n$ from a population in which X is uniformly distributed on the region $(0,1) \cup (2,c)$ for some unknown c > 2. Find closed-form expressions for the Method of Moments and Maximum Likelihood Estimators, to be denoted by T_1 and T_2, respectively.

Chapter 9

Introduction to Model Building

All models are wrong, but some are useful.—George Box[1]

[Mathematical models] should be made as simple as possible, but not simpler.—Albert Einstein[2]

Beware of geeks bearing formulas.—Warrent Buffett, 2009, on the role of "quants" (Wall Street analysts who form probabilistic models for currency, bonds etc.) in the 2008 financial collapse.

The above quote by Box says it all. Consider for example the family of normal distributions. In real life, random variables are bounded—no person's height is negative or greater than 500 inches—and are inherently discrete, due to the finite precision of our measuring instruments. Thus, technically, no random variable in practice can have an exact normal distribution. Yet the assumption of normality pervades statistics, and has been enormously successful, provided one understands its approximate nature.

The situation is similar to that of physics. Paraphrasing Box, we might say that the physical models used when engineers design an airplane wing are all wrong—but they are useful. We know that in many analyses of bodies in motion, we can neglect the effect of air resistance. But we also know that in some situations one must include that factor in our model.

So, the field of probability and statistics is fundamentally about *modeling*. The field is extremely useful, provided the user understands the modeling issues well. For this reason, this book contains this separate chapter on modeling issues.

[1] George Box (1919-) is a famous statistician, with several statistical procedures named after him.

[2] The reader is undoubtedly aware of Einstein's (1879-1955) famous theories of relativity, but may not know his connections to probability theory. His work on **Brownian motion**, which describes the path of a molecule as it is bombarded by others, is probabilistic in nature, and later developed into a major branch of probability theory. Einstein was also a pioneer in quantum mechanics, which is probabilistic as well. At one point, he doubted the validity of quantum theory, and made his famous remark, "God does not play dice with the universe."

9.1 "Desperate for Data"

Suppose we have the samples of men's and women's heights, $X_1, ..., X_n$ and $Y_1, ..., Y_n$. Assume for simplicity that the variance of height is the same for each gender, σ^2. The means of the two populations are designated by μ_1 and μ_2.

Say we wish to guess the height of a new person who we know to be a man but for whom we know nothing else. We do not see him, etc.

9.1.1 Known Distribution

Suppose for just a moment that we actually know the distribution of X, i.e. the *population* distribution of male heights. What would be the best constant g to use as our guess for a person about whom we know nothing other than gender?

Well, we might borrow from Section 8.2 and use mean squared error,

$$E[(g - X)^2] \tag{9.1}$$

as our criterion of goodness of guessing. But we already know what the best g is, from Section 3.59: The best g is μ_1. Our best guess for this unseen man's height is the mean height of all men in the population.

9.1.2 Estimated Mean

Of course, we don't know μ_1, but we can do the next-best thing, i.e. use an estimate of it from our sample.

The natural choice for that estimator would be

$$T_1 = \overline{X}, \tag{9.2}$$

the mean height of men in our sample.

But what if n is really small, say n = 5? That's awfully small. We may wish to consider adding the women's heights to our estimate, in order to get a larger sample. Then we would estimate μ_1 by

$$T_2 = \frac{\overline{X} + \overline{Y}}{2}, \tag{9.3}$$

It may at first seem obvious that T_1 is the better estimator. Women tend to be shorter, after all, so pooling

the data from the two genders would induce a bias. On the other hand, we found in Section 8.2 that for any estimator,

$$\text{MSE = variance of the estimator} + \text{bias of the estimator}^2 \tag{9.4}$$

In other words, *some amount of bias may be tolerable*, if it will buy us a subtantial reduction in variance. After all, women are not that much shorter than men, so the bias might not be too bad. Meanwhile, the pooled estimate should have lower variance, as it is based on 2n observations instead of n; (7.8) indicates that.

Before continuing, note first that T_2 is based on a simpler model than is T_1, as T_2 ignores gender. We thus refer to T_1 as being based on the more complex model.

Which one is better? The answer will need a criterion for goodness of estimation, which we will take to be mean squared error, MSE. So, the question becomes, which has the smaller MSE, T_1 or T_2? In other words:

Which is smaller, $E[(T_1 - \mu_1)^2]$ or $E[(T_2 - \mu_1)^2]$?

9.1.3 The Bias/Variance Tradeoff

We could calculate MSE from scratch, but it would probably be better to make use of the work we already went through, producing (8.66). This is especially true in that we know a lot about variance of sample means, and we will take this route.

So, let's find the biases of the two estimators.

- T_1

  T_1 is unbiased, from (7.8). So,

  bias of $T_1 = 0$

- T_2

$$\begin{aligned} E(T_2) &= E(0.5\overline{X} + 0.5\overline{Y}) \text{ (definition)} &(9.5)\\ &= 0.5E\overline{X} + 0.5E\overline{Y} \text{ (linearity of E())} &(9.6)\\ &= 0.5\mu_1 + 0.5\mu_2 \text{ [from (7.8)]} &(9.7) \end{aligned}$$

  So,

bias of $T_2 = (0.5\mu_1 + 0.5\mu_2) - \mu_1$

On the other hand, T_2 has a smaller variance than T_1:

- T_1

  Recalling (7.13), we have

$$Var(T_1) = \frac{\sigma^2}{n} \tag{9.8}$$

- T_2

$$\begin{aligned} Var(T_2) &= Var(0.5\overline{X} + 0.5\overline{Y}) && (9.9)\\ &= 0.5^2 Var(\overline{X}) + 0.5^2 Var(\overline{Y}) \text{ (properties of Var())} && (9.10)\\ &= 2 \cdot 0.25 \cdot \frac{\sigma^2}{n} \text{ [from 7.13]} && (9.11)\\ &= \frac{\sigma^2}{2n} && (9.12) \end{aligned}$$

These findings are highly instructive. You might at first think that "of course" T_1 would be the better predictor than T_2. But for a small sample size, the smaller (actually 0) bias of T_1 is not enough to counteract its larger variance. T_2 is biased, yes, but it is based on double the sample size and thus has half the variance.

In light of (8.66), we see that T_1, the "true" predictor, may not necessarily be the better of the two predictors. Granted, it has no bias whereas T_2 does have a bias, but the latter has a smaller variance.

So, under what circumstances will T_1 be better than T_2? Let's answer this by using (8.65):

$$MSE(T_1) = \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n} \tag{9.13}$$

$$MSE(T_2) = \frac{\sigma^2}{2n} + \left(\frac{\mu_1 + \mu_2}{2} - \mu_1\right)^2 = \frac{\sigma^2}{2n} + \left(\frac{\mu_2 - \mu_1}{2}\right)^2 \tag{9.14}$$

T_1 is a better predictor than T_2 if (9.13) is smaller than (9.14), which is true if

$$\left(\frac{\mu_2 - \mu_1}{2}\right)^2 > \frac{\sigma^2}{2n} \tag{9.15}$$

Granted, we don't know the values of the μ_1 and σ^2, so in a real situation, we won't really know whether to use T_1 or T_2. But the above analysis makes the point that under some circumstances, it really is better to pool the data in spite of bias.

9.1.4 Implications

So you can see that T_1 is better only if either

- n is large enough, or
- the difference in population mean heights between men and women is large enough, or
- there is not much variation within each population, e.g. most men have very similar heights

Since that third item, small within-population variance, is rarely seen, let's concentrate on the first two items. The big revelation here is that:

> A more complex model is more accurate than a simpler one only if either
>
> - we have enough data to support it, or
> - the complex model is sufficiently different from the simpler one

In height/gender example above, if n is too small, we are "desperate for data," and thus make use of the female data to augment our male data. Though women tend to be shorter than men, the bias that results from that augmentation is offset by the reduction in estimator variance that we get. But if n is large enough, the variance will be small in either model, so when we go to the more complex model, the advantage gained by reducing the bias will more than compensate for the increase in variance.

THIS IS AN ABSOLUTELY FUNDAMENTAL NOTION IN STATISTICS.

This was a very simple example, but you can see that in complex settings, fitting too rich a model can result in very high MSEs for the estimates. In essence, everything becomes noise. (Some people have cleverly coined the term **noise mining**, a play on the term **data mining**.) This is the famous **overfitting** problem.

In our unit on statistical relations, Chapter 10, we will show the results of a scary experiment done at the Wharton School, the University of Pennsylvania's business school. The researchers deliberately added fake data to a prediction equation, and standard statistical software identified it as "significant"! This is partly a problem with the word itself, as we saw in Section 7.4, but also a problem of using far too complex a model, as will be seen in that future unit.

Note that of course (9.15) contains several unknown population quantities. I derived it here merely to establish a principle, namely that a more complex model may perform more poorly under some circumstances.

It would be possible, though, to make (9.15) into a practical decision tool, by estimating the unknown quantities, e.g. replacing μ_1 by $\overline{X}$. This then creates possible problems with confidence intervals, whose derivation did not include this extra decision step. Such estimators, termed **adaptive**, are beyond the scope of this book.

9.2 Assessing "Goodness of Fit" of a Model

Our example in Section 8.1.4 concerned how to estimate the parameters of a gamma distribution, given a sample from the distribution. But that assumed that we had already decided that the gamma model was reasonable in our application. Here we will be concerned with how we might come to such decisions.

Assume we have a random sample $X_1, ..., X_n$ from a distribution having density f_X.

9.2.1 The Chi-Square Goodness of Fit Test

The classic way to do this would be the **Chi-Square Goodness of Fit Test**. We would set

$$H_0 : f_X \text{ is a member of the exponential parametric family} \tag{9.16}$$

This would involve partitioning $(0, \infty)$ into k intervals (s_{i-1}, s_i) of our choice, and setting

$$N_i = \text{number of } X_i \text{ in } (s_{i-1}, s_i) \tag{9.17}$$

We would then find the Maximum Likelihood Estimate (MLE) of λ, on the assumption that the distribution of X really is exponential. The MLE turns out to be the reciprocal of the sample mean, i.e.

$$\widehat{\lambda} = 1/\overline{X} \tag{9.18}$$

This would be considered the parameter of the "best-fitting" exponential density for our data. We would then estimate the probabilities

$$p_i = P[X \epsilon (s_{i-1}, s_i)] = e^{-\lambda s_{i-1}} - e^{-\lambda s_i},\ i = 1, ..., k. \tag{9.19}$$

by

$$\widehat{p_i} = e^{-\widehat{\lambda} s_{i-1}} - e^{-\widehat{\lambda} s_i},\ i = 1, ..., k. \tag{9.20}$$

Note that N_i has a binomial distribution, with n trials and success probability p_i. Using this, the expected value of EN_i is estimated to be

$$\nu_i = n(e^{-\widehat{\lambda} s_{i-1}} - e^{-\widehat{\lambda} s_i}),\ i = 1, ..., k. \tag{9.21}$$

Our test statistic would then be

$$Q = \sum_{i=1}^{k} \frac{(N_i - v_i)^2}{v_i} \tag{9.22}$$

where v_i is the expected value of N_i under the assumption of "exponentialness." It can be shown that Q is approximately chi-square distributed with k-2 degrees of freedom.[3] Note that only large values of Q should be suspicious, i.e. should lead us to reject H_0; if Q is small, it indicates a good fit. If Q were large enough to be a "rare event," say larger than $\chi_{0.95,k-2}$, we would decide NOT to use the exponential model; otherwise, we would use it.

Hopefully the reader has immediately recognized the problem here. If we have a large sample, this procedure will pounce on tiny deviations from the exponential distribution, and we would decide not to use the exponential model—even if those deviations were quite minor. Again, no model is 100% correct, and thus a goodness of fit test will eventually tell us not to use *any* model at all.

9.2.2 Kolmogorov-Smirnov Confidence Bands

Again consider the problem above, in which we were assessing the fit of a exponential model. In line with our major point that confidence intervals are far superior to hypothesis tests, we now present **Kolmogorov-Smirnov confidence bands**, which work as follows.

Recall the concept of empirical cdfs, presented in Section 8.4.1. It turns out that the distribution of

$$M = \max_{-\infty < t \infty} |\widehat{F}_X(t) - F_X(t)| \tag{9.23}$$

[3] We have k intervals, but the N_i must sum to n, so there are only k-1 free values. We then subtract one more degree of freedom, having estimated the parameter λ.

is the same for all distributions having a density. This fact (whose proof is related to the general method for simulating random variables having a given density, in Section 4.6) tells us that, without knowing anything about the distribution of X, we can be sure that M has the same distribution. And it turns out that

$$F_M(1.358n^{-1/2}) = 0.95 \tag{9.24}$$

Define "upper" and "lower" functions

$$U(t) = \widehat{F}_X(t) + 1.358n^{-1/2}, \;\; L(t) = \widehat{F}_X(t) - 1.358n^{-1/2} \tag{9.25}$$

So, what (9.23) and (9.24) tell us is

$$0.95 = P\,(\text{the curve } F_X \text{ is entirely between U and L}) \tag{9.26}$$

So, the pair of curves, (L(t), U(t)) is called a a **95% confidence band** for F_X.

The usefulness is similar to that of confidence intervals. If the band is very wide, we know we really don't have enough data to decide much about the distribution of X. But if the band is narrow but some member of the family comes reasonably close to the band, we would probably decide that the model is a good one, even if no member of the family falls within the band. Once again, we should NOT pounce on tiny deviations from the model.

Warning: The Kolmogorov-Smirnov procedure available in the R language performs only a hypothesis test, rather than forming a confidence band. In other words, it simply checks to see whether a member of the family falls within the band. This is not what we want, because we may be perfectly happy if a member is only *near* the band.

Of course, another way, this one less formal, of assessing data for suitability for some model is to plot the data in a histogram or something of that naure.

9.3 Bias Vs. Variance—Again

In our unit on estimation, Section 8.4, we saw a classic tradeoff in histogram- and kernel-based density estimators. With histograms, for instance, the wider bin width produces a graph which is smoother, but possibly *too* smooth, i.e. with less oscillation than the true population curve has. The same problem occurs with larger values of h in the kernel case.

This is actually yet another example of the bias/variance tradeoff, discussed in above and, as mentioned, **ONE OF THE MOST RECURRING NOTIONS IN STATISTICS**. A large bin width, or a large value

of h, produces more bias. In general, the large the bin width or h, the further $E[\widehat{f}_R(t)$ is from the true value of $f_R(t)$. This occurs because we are making use of points which are not so near t, and thus at which the density height is different from that of $f_R(t)$. On the other hand, because we are making use of more points, $Var[\widehat{f}_r(t)]$ will be smaller.

THERE IS NO GOOD WAY TO CHOOSE THE BIN WIDTH OR h. Even though there is a lot of theory to suggest how to choose the bin width or h, no method is foolproof. This is made even worse by the fact that the theory generally has a goal of minimizing *integrated* mean squared error,

$$\int_{-\infty}^{\infty} E\left[\left(\widehat{f}_R(t) - f_R(t)\right)^2\right] \, dt \tag{9.27}$$

rather than, say, the mean squared error at a particular point of interest, v:

$$E\left[\left(\widehat{f}_R(t) - f_R(t)\right)^2\right] \tag{9.28}$$

9.4 Robustness

Traditionally, the term *robust* in statistics has meant resilience to violations in assumptions. For example, in Section 7.2.8, we presented Student-t, a method for finding exact confidence intervals for means, assuming normally-distributed populations. But as noted at the outset of this chapter, no population in the real world has an exact normal distribution. The question at hand (which we will address below) is, does the Student-t method still give approximately correct results if the sample population is not normal? If so, we say that Student-t is **robust** to the normality assumption.

Later, there was quite a lot of interest among statisticians in estimation procedures that do well even if there are **outliers** in the data, i.e. erroneous observations that are in the fringes of the sample. Such procedures are said to be robust to outliers.

Our interest here is on robustness to assumptions. Let us first consider the Student-t example. As discussed in Section 7.2.8, the main statistic here is

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \tag{9.29}$$

where μ is the population mean and s is the unbiased version of the sample variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \tag{9.30}$$

The distribution of T, under the assumption of a normal population, has been tabulated, and tables for it appear in virtually every textbook on statistics. But what if the population is not normal, as is inevitably the case?

The answer is that it doesn't matter. For large n, even for samples having, say, n = 20, the distribution of T is close to N(0,1) by the Central Limit Theorem regardless of whether the population is normal.

By contrast, consider the classic procedure for performing hypothesis tests and forming confidence intervals for a population variance σ^2, which relies on the statistic

$$K = \frac{(n-1)s^2}{\sigma^2} \tag{9.31}$$

where again s^2 is the unbiased version of the sample variance. If the sampled population is normal, then K can be shown to have a chi-square distribution with n-1 degrees of freedom. This then sets up the tests or intervals. However, it has been shown that these procedures are not robust to the assumption of a normal population. See *The Analysis of Variance: Fixed, Random, and Mixed Models*, by Hardeo Sahai and Mohammed I. Ageel, Springer, 2000, and the earlier references they cite, especially the pioneering work of Scheffe'.

Exercises

Note to instructor: See the Preface for a list of sources of real data on which exercises can be assigned to complement the theoretical exercises below.

1. In our example in Section 9.1, assume $\mu_1 = 70, \mu_2 = 66, \sigma = 4$ and the distribution of height is normal in the two populations. Suppose we are predicting the height of a man who, unknown to us, has height 68. We hope to guess within two inches. Find $P(|T_1 - 68|) < 2$ and $P(|T_2 - 68|) < 2$ for various values of n.

2. In Section 8.7 we discussed *simultaneous inference*, the forming of confidence intervals whose joint confidence level was 95% or some other target value. The Kolmogorov-Smirnov confidence band in Section 9.2.2 allows us to computer infinitely many confidence intervals for $F_X(t)$ at different values of t, at a "price" of only 1.358. Still, if we are just estimating $F_X(t)$ at a single value of t, an individual confidence interval using (7.29) would be narrower than that given to us by Kolmogorov-Smirnov. Compare the widths of these two intervals in a situation in which the true value of $F_X(t) = 0.4$.

3. Say we have a random sample $X_1, ..., X_n$ from a population with mean μ and variance σ^2. The usual estimator of μ is the sample mean $\bar{X}$, but here we will use what is called a *shrinkage estimator*: Our estimate of μ will be $0.9\bar{X}$. Find the mean squared error of this estimator, and give an inequality (you don't have to algebraically simplify it) that shows under what circumstances $0.9\bar{X}$ is better than $\bar{X}$. (Strong advice: Do NOT "reinvent the wheel." Make use of what we have already derived.)

.

Chapter 10

Statistical Relations Between Variables

In many senses, this chapter is the real core of statistics, especially from a computer science point of view.

In this chapter 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 covered in this chapter, we are interested in relations among several variables, symmetrically, i.e. not having one variable play a special role.

10.1 Regression Analysis

10.1.1 The Goals: Prediction and Understanding

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.

- **Prediction:** Here we are trying to predict one variable from one or more others.
- **Understanding:** Here we wish to determine which of several variables have a greater effect on (or relation to) a given variable.

[1] Yogi Berra (1925-) 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."

Denote the **predictor variables** by, $X^{(1)}, ..., X^{(r)}$. They are also called **independent variables**. The variable to be predicted, Y, is often called the **response variable**, or the **dependent variable**.

A common statistical methodology used for such analyses is called **regression analysis**. In the important special cases in which the response variable Y is an indicator variable (Section 3.6),[2] taking on just 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.

10.1.2 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 Understanding. 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 Danny Raz, Nidhan Choudhuri and Prasun Sinha, 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 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 Understanding 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

[2] Sometimes called a **dummy variable**.

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.

10.1.3 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.

Intuitively, though, we would guess that mean weight increases with height. To state this precisely, take Y to be the weight W and $X^{(1)}$ to be the height H, and define

$$m_{W;H}(t) = E(W|H = t) \tag{10.1}$$

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 $X = t$.

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*!

10.1.4 Estimating That Relationship from Sample Data

As noted, though, $m_{W;H}(t)$ 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

$$(H_1, W_1), ..., (H_{1000}, W_{1000}) \tag{10.2}$$

being their heights and weights. We again wish to use this data to estimate population values. 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.[3] 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 $\widehat{m}_{W;H}(t)$.

There are many methods like this (see Section 10.3), 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 9.1.

Typically the parametric model chosen is linear, i.e. we assume that $m_{W;H}(t)$ is a linear function of t:

$$m_{W;H}(t) = ct + d \tag{10.3}$$

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.

Equation (10.3) 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. (10.3). Again, the quantities c and d in (10.3) are population values, and as such, we must estimate them from the data.

So, how can we estimate c and d? We'll go into details in Section 10.1.9, but here is a preview:

[3] Of course, the population of Davis is finite, but there is the conceptual population of all people who *could* live in Davis.

Using the result on page 45, together with the Law of Total Expectation in Section 5.7.3, we have that the minimum value of the quantity

$$E\left[(W - g(H))^2\right] \tag{10.4}$$

overall all possible functions $g(H)$, is attained by setting

$$g(H) = m_{W;H}(H) \tag{10.5}$$

In other words, $m_{W;H}(H)$ is the best predictor of W among all possible functions of H, in the sense of minimizing mean squared prediction error.[4]

Since we are assuming the model (10.3), this in turn means that

$$E\left[(W - (rH + s))^2\right] \tag{10.6}$$

is minimized by setting r = c and s = d. Now, 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 (10.6) is

$$\frac{1}{n}\sum_{i=1}^{n}[W_i - (rH_i + s)]^2 \tag{10.7}$$

Here (10.6) is the mean squared prediction error using r and s in the population, and (10.7) is the mean squared prediction error using r and s in our sample. Since r = c and s = d minimize (10.6), it is natural to estimate them by the r and s that minimize (10.7).

These are then the classical *least-squares estimators* of c and d.

> **Major Point 3:** In statistical regression analysis, one uses a linear model as in (10.3), estimating the coefficients by minimizing (10.7).

We will elaborate on this in Section 10.1.9.

[4] But if we wish to minimize the mean absolute prediction error, $E\left(|W - g(H)|\right)$, the best function turns out to be is $g(H) =$ median$(W|H)$.

10.1.5 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

$$X = \left(X^{(1)}, X^{(2)}\right) = (H, A) = \text{(height, age)} \tag{10.8}$$

so as to study the relationship of weight with height and age. If we used a linear model, we would write for $t = (t_1, t_2)$,

$$m_{W;H,A}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 \tag{10.9}$$

In other words

$$\text{mean weight} = \beta_0 + \beta_1 \text{ height} + \beta_2 \text{ age} \tag{10.10}$$

(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.

10.1.6 Interaction Terms

Equation (10.9) 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 (10.9), with the same number for height in both, and subtract; you get $10\beta_2$, an expression that has no height term.

If we feel that the assumption is not a good one (there are also data plotting techniques to help assess this), we can add an **interaction term** to (10.9), 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

$$m_{W;H}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_1 t_2 \tag{10.11}$$

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 10.1.1. 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 $\begin{pmatrix} k \\ 2 \end{pmatrix}$ 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 10.1.10.1). And with so many interaction terms, the model would be difficult to interpret.

10.1.7 Nonrandom Predictor Variables

In our weight/height/age example above, all three variables are random. If we repeat the "experiment," i.e. we choose another sample of 1000 people, these new people will have different weights, different heights and different ages from the people in the first sample.

But we must point out that the function $m_{Y;X}$ for the regression function of Y and X makes sense even if X is nonrandom. To illustrate this, let's look at the ALOHA network example in our introductory chapter on discrete probability, Section 2.1.

```
# simulation of simple form of slotted ALOHA

# a node is active if it has a message to send (it will never have more
# than one in this model), inactive otherwise

# the inactives have a chance to go active earlier within a slot, after
# which the actives (including those newly-active) may try to send; if
# there is a collision, no message gets through

# parameters of the system:
# s = number of nodes
# b = probability an active node refrains from sending
# q = probability an inactive node becomes active

# parameters of the simulation:
# nslots = number of slots to be simulated
# nb = number of values of b to run; they will be evenly spaced in (0,1)

# will find mean message delay as a function of b;

# we will rely on the "ergodicity" of this process, which is a Markov
# chain (see http://heather.cs.ucdavis.edu/~matloff/132/PLN/Markov.tex),
# which means that we look at just one repetition of observing the chain
# through many time slots

# main loop, running the simulation for many values of b
alohamain <- function(s,q,nslots,nb) {
   deltab = 0.7 / nb  # we'll try nb values of b in (0.2,0.9)
   md <- matrix(nrow=nb,ncol=2)
   b <- 0.2
   for (i in 1:nb) {
      b <- b + deltab
      w <- alohasim(s,b,q,nslots)
      md[i,] <- alohasim(s,b,q,nslots)
   }
```

```
36     return(md)
37  }
38
39  # simulate the process for h slots
40  alohasim <- function(s,b,q,nslots) {
41     # status[i,1] = 1 or 0, for node i active or not
42     # status[i,2] = if node i active, then epoch in which msg was created
43     # (could try a list structure instead a matrix)
44     status <- matrix(nrow=s,ncol=2)
45     # start with all active with msg created at time 0
46     for (node in 1:s) status[node,] <- c(1,0)
47     nsent <- 0  # number of successful transmits so far
48     sumdelay <- 0  # total delay among successful transmits so far
49     # now simulate the nslots slots
50     for (slot in 1:nslots) {
51        # check for new actives
52        for (node in 1:s) {
53           if (!status[node,1])  # inactive
54              if (runif(1) < q) status[node,] <- c(1,slot)
55        }
56        # check for attempted transmissions
57        ntrysend <- 0
58        for (node in 1:s) {
59           if (status[node,1])  # active
60              if (runif(1) > b) {
61                 ntrysend <- ntrysend + 1
62                 whotried <- node
63              }
64        }
65        if (ntrysend == 1) {  # something gets through iff exactly one tries
66           # do our bookkeeping
67           sumdelay <- sumdelay + slot - status[whotried,2]
68           # this node now back to inactive
69           status[whotried,1] <- 0
70           nsent <- nsent + 1
71        }
72     }
73     return(c(b,sumdelay/nsent))
74  }
```

A minor change is that I replaced the probability p, the probability that an active node would send in the original example to b, the probability of *not* sending (b for "backoff"). Let A denote the time A (measured in slots) between the creation of a message and the time it is successfully transmitted.

We are interested in mean delay, i.e. the mean of A. (Note that our Y_i here are sample mean values of A, whereas we want to draw inferences about the population mean value of A.) We are particularly interested in the effect of b here on that mean. Our goal here, as described in Section 10.1.1, could be Prediction, so that we could have an idea of how much delay to expect in future settings. Or, we may wish to explore finding an optimal b, i.e. one that minimizing the mean delay, in which case our goal would be more in the direction of Understanding.

I ran the program with certain arguments, and then plotted the data:

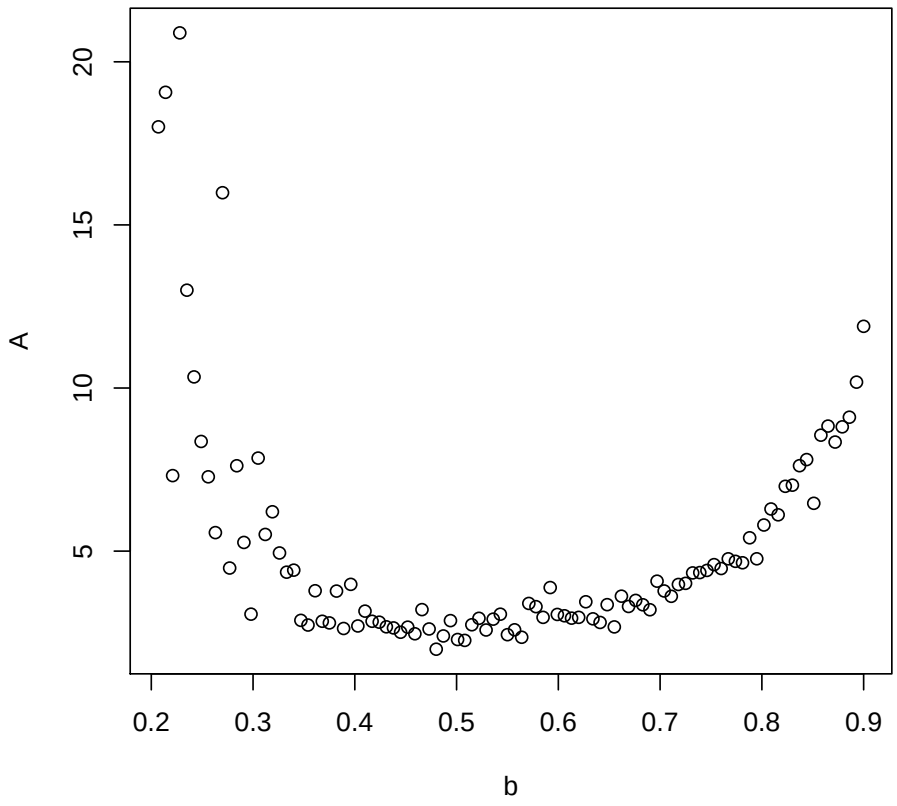


Figure 10.1: Scatter Plot

```
> md <- alohamain(4,0.1,1000,100)
> plot(md,cex=0.5,xlab="b",ylab="A")
```

The plot is shown in Figure 10.1.

Note that though our values b here are nonrandom, the A values are indeed random. To dramatize that point, I ran the program again. (Remember, unless you specify otherwise, R will use a different seed for its random number stream each time you run a program.) I've superimposed this second data set on the first, using filled circles this time to represent the points:

```
md2 <- alohamain(4,0.1,1000,100)
points(md2,cex=0.5,pch=19)
```

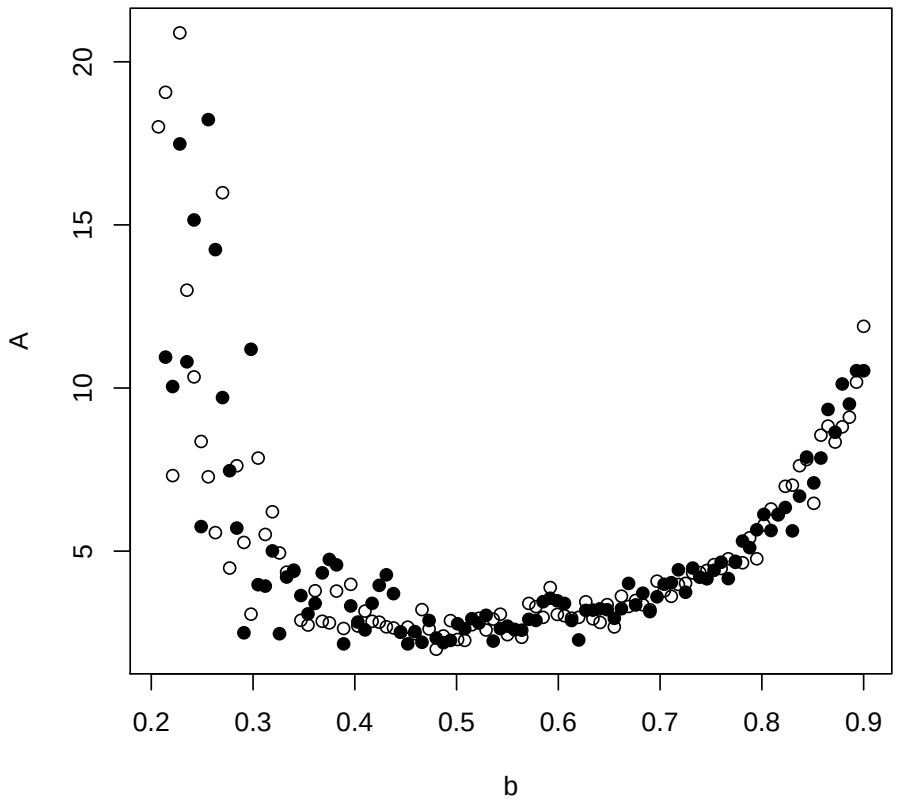


Figure 10.2: Scatter Plot, Two Data Sets

The plot is shown in Figure 10.2.

We do expect some kind of U-shaped relation, as seen here. For b too small, the nodes are clashing with each other a lot, causing long delays to message transmission. For b too large, we are needlessly backing off in many cases in which we actually would get through.

This looks like a quadratic relationship, meaning the following. Take our response variable Y to be A, take our first predictor $X^{(1)}$ to be b, and take our second predictor $X^{(2)}$ to be b^2. Then when we say A and b have a quadratic relationship, we mean

$$m_{A;b}(b) = \beta_0 + \beta_1 b + \beta_2 b^2 \tag{10.12}$$

for some constants $\beta_0, \beta_1, \beta_2$. So, we are using a three-parameter family for our model of $m_{A;b}$. No model is exact, but our data seem to indicate that this one is reasonably good, and if further investigation confirms that, it provides for a nice compact summary of the situation.

Again, we'll see how to estimate the β_i in Section 10.1.9.

We could also try adding two more predictor variables, consisting of $X^{(3)} = q$ and $X^{(4)} = s$, the node activation probability and number of nodes, respectively. We would collect more data, in which we varied the values of q and s, and then could entertain the model

$$m_{A;b,q}(u,v,w) = \beta_0 + \beta_1 u + \beta_2 u^2 + \beta_3 v + \beta_4 w \tag{10.13}$$

10.1.8 Prediction

Let's return to our weight/height/age example. We are informed of a certain person, of height 70.4 and age 24.8, but weight unknown. What should we predict his weight to be?

The intuitive answer (justified formally by Section 10.4.1) is that we predict his weight to be the mean weight for his height/age group,

$$m_{W;H,A}(70.4, 24.8) \tag{10.14}$$

But that is a population value. Say we estimate the function $m_{W;H}$ using that data, yielding $\widehat{m}_{W;H}$. Then we could take as our prediction for the new person's weight

$$\widehat{m}_{W;H,A}(70.4, 24.8) \tag{10.15}$$

If our model is (10.9), then (10.15) is

$$\widehat{m}_{W;H}(t) = \widehat{\beta}_0 + \widehat{\beta}_1 70.4 + \widehat{\beta}_2 24.8 \tag{10.16}$$

where the $\widehat{\beta}_i$ are estimated from our data by least-squares.

10.1.9 Parametric Estimation of Linear Regression Functions

10.1.9.1 Meaning of "Linear"

Here we model $m_{Y;X}$ as a linear function of $X^{(1)}, ..., X^{(r)}$:

$$m_{Y;X}(t) = \beta_0 + \beta_1 t^{(1)} + ... + \beta_r t^{(r)} \tag{10.17}$$

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 (10.12).

Instead, the word *linear* refers to the regression function being linear in the parameters. So, for instance, (10.12) 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 (10.22) below.

10.1.9.2 Point Estimates and Matrix Formulation

So, how do we estimate the β_i? Look for instance at (10.12). Keep in mind that in (10.12), the β_i are population values. We need to estimate them from our data. How do we do that? As previewed in Section 10.1.4, the usual method is least-squares. Here we will go into the details.

Let's define (b_i, A_i) to be the i^{th} pair from the simulation. In the program, this is **md[i,]**. Our estimated parameters will be denoted by $\hat{\beta}_i$. As in (10.7), the estimation methodology involves finding the values of $\hat{\beta}_i$ which minimize the sum of squared differences between the actual A values and their predicted values:

$$\sum_{i=1}^{100} [A_i - (\hat{\beta}_0 + \hat{\beta}_1 b_i + \hat{\beta}_2 b_i^2)]^2 \tag{10.18}$$

Obviously, this is a calculus problem. We set the partial derivatives of (10.18) with respect to the $\hat{\beta}_i$ to 0, giving use three linear equations in three unknowns, and then solve.

For the general case (10.17), we have r+1 equations in r+1 unknowns. This is most conveniently expressed in matrix terms. Let $X_i^{(j)}$ be the value of $X^{(j)}$ for the i^{th} observation in our sample, and let Y_i be the corresponding Y value. Plugging this data into (10.1.9.1), we have

$$E(Y_i | X_i^{(1)}, ..., X_i^{(r)}) = \beta_0 + \beta_1 X_i^{(1)} + ... + \beta_r X_i^{(r)}, \ i = 1, ..., n \tag{10.19}$$

That's a system of n linear equations, which from your linear algebra class you know can be represented more compactly by a matrix, as follows.

Let Q be the n x (r+1) matrix whose (i,j) element is $X_i^{(j)}$, with $X_i^{(0)}$ taken to be 1. For instance, if we are predicting weight from height and age based on a sample of 100 people, then Q would look like this:

$$\begin{pmatrix} 1 & H_1 & A_1 \\ 1 & H_2 & A_2 \\ ... & & \\ 1 & H_{100} & A_{100} \end{pmatrix} \tag{10.20}$$

For example, row 5 of Q would consist of a 1, then the height and age of the fifth person in our sample.

Also, let

$$V = (Y_1, ..., Y_n)', \tag{10.21}$$

Then the system (10.19) in matrix form is

$$E(V|Q) = Q\beta \tag{10.22}$$

where

$$\beta = (\beta_0, \beta_1, ..., \beta_r)' \tag{10.23}$$

Keep in mind that the derivation below is conditional on the $X_j^{(i)}$, i.e. conditional on Q, as shown above. 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).

Now to estimate the β_i, let

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_r)' \tag{10.24}$$

with our goal now being to find $\hat{\beta}$. The matrix form of (10.18) (now for the general case, not just ALOHA) is

$$(V - Q\hat{\beta})'(V - Q\hat{\beta}) \tag{10.25}$$

Then it can be shown that, after all the partial derivatives are taken and set to 0, the solution is

$$\hat{\beta} = (Q'Q)^{-1}Q'V \tag{10.26}$$

Note by the way that this implies that $\hat{\beta}$ is an unbiased estimate of β:

$$\begin{aligned} E\hat{\beta} &= E[(Q'Q)^{-1}Q'V] \text{ (10.26)} && (10.27)\\ &= (Q'Q)^{-1}Q'EV \text{ (linearity of E())} && (10.28)\\ &= (Q'Q)^{-1}Q' \cdot Q\beta \text{ (10.22)} && (10.29)\\ &= \beta && (10.30) \end{aligned}$$

In some applications, we assume there is no constant term β_0 in (10.17). This means that our Q matrix no longer has the column of 1s on the left end, but everything else above is valid.

10.1.9.3 Back to Our ALOHA Example

R or any other statistical package does the work for us. In R, we can use the **lm()** ("linear model") function:

```
> md <- cbind(md,md[,1]^2)
> lmout <- lm(md[,2] ~ md[,1] + md[,3])
```

First I added a new column to the data matrix, consisting of b^2. I then called **lm()**, with the argument

```
md[,2] ~ md[,1] + md[,3]
```

R documentation calls this model specification argument the **formula**. It states that I wish to use the first and third columns of **md**, i.e. b and b^2, as predictors, and use A, i.e. second column, as the response variable.[5]

The return value from this call, which I've stored in **lmout**, is an object of class **lm**. One of the member variables of that class, **coefficients**, is the vector $\widehat{\beta}$:

```
> lmout$coefficients
(Intercept)      md[, 1]      md[, 3]
   27.56852   -90.72585    79.98616
```

[5]Unfortunately, R did not allow me to put the squared column directly into the formula, forcing me to use **cbind()** to make a new matrix.

So, $\widehat{\beta}_0 = 27.57$ and so on.

The result is

$$\widehat{m}_{A,b}(t) = 27.57 - 90.73t + 79.99t^2 \tag{10.31}$$

Another member variable in the **lm** class is **fitted.values**. This is the "fitted curve," meaning the values of (10.31) at $b_1, ..., b_{100}$. In other words, this is (10.31). I plotted this curve on the same graph,

```
> lines(cbind(md[,1],lmout$fitted.values))
```

See Figure 10.3. As you can see, the fit looks fairly good. What should we look for?

Remember, we don't expect the curve to go through the points—we are estimating the mean of A for each b, not the A values themselves. There is always variation around the mean. If for instance we are looking at the relationship between people heights and weights, the mean weight for people of height 70 inches might be, say, 160 pounds, but we know that some 70-inch-tall people weigh more than this and some weigh less.

However, there seems to be a tendency for our estimates of $\widehat{m}_{A,b}(t)$ to be too low for values in the middle range of t, and possible too high for t around 0.3 or 0.4. **However, with a sample size of only 100, it's difficult to tell.** It's always important to keep in mind that the data are random; a different sample may show somewhat different patterns. Nevertheless, we should consider a more complex model.

So I tried a quartic, i.e. fourth-degree, polynomial model. I added third- and fourth-power columns to **md**, calling the result **md4**, and invoked the call

```
lm(md4[,2] ~ md4[,1] + md4[,3] + md4[,4] + md4[,5])
```

The result was

```
> lmout$coefficients
(Intercept)    md4[, 1]    md4[, 3]    md4[, 4]    md4[, 5]
   95.98882  -664.02780  1731.90848 -1973.00660   835.89714
```

In other words, we have an estimated regression function of

$$\widehat{m}_{A,b}(t) = 95.98882 - 664.02780\, t + 1731.90848\, t^2 - 1973.00660\, t^3 + 835.89714\, t^4 \tag{10.32}$$

The fit is shown in Figure 10.4. It looks much better. On the other hand, we have to worry about overfitting. We return to this issue in Section 10.1.10.1).

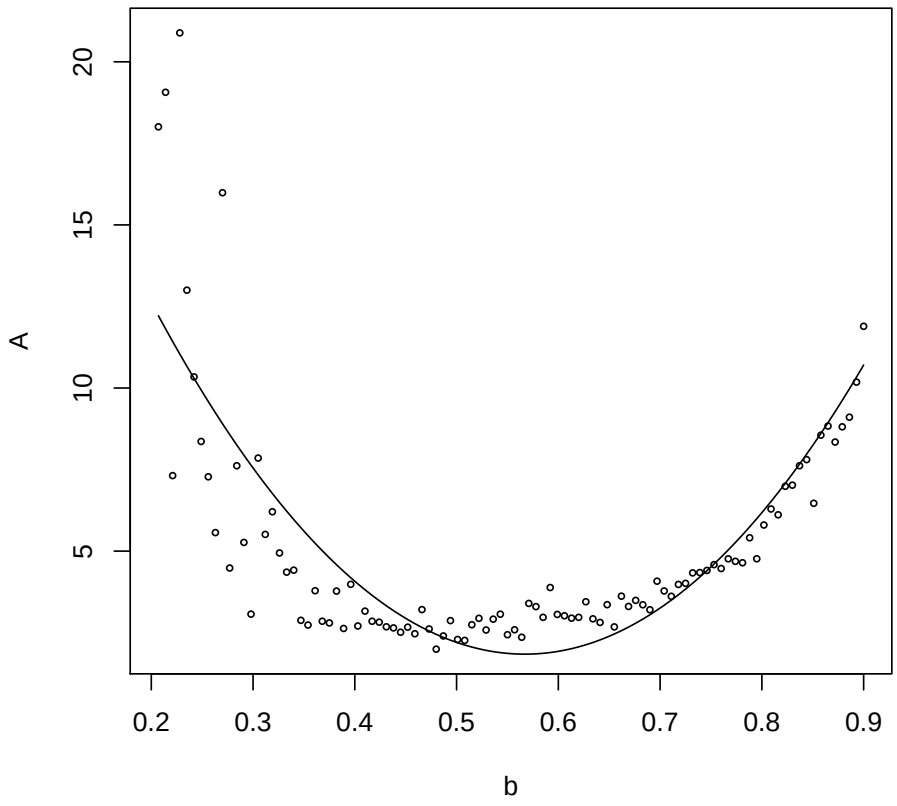


Figure 10.3: Quadratic Fit Superimposed

10.1.9.4 Approximate Confidence Intervals

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.

For instance, recall the study on object-oriented programming in Section 10.1.1. The goal there was primarily Understanding, 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 (10.26) shows that the $\widehat{\beta}_i$ are sums of the components of V, i.e. the Y_j. So, the Central Limit

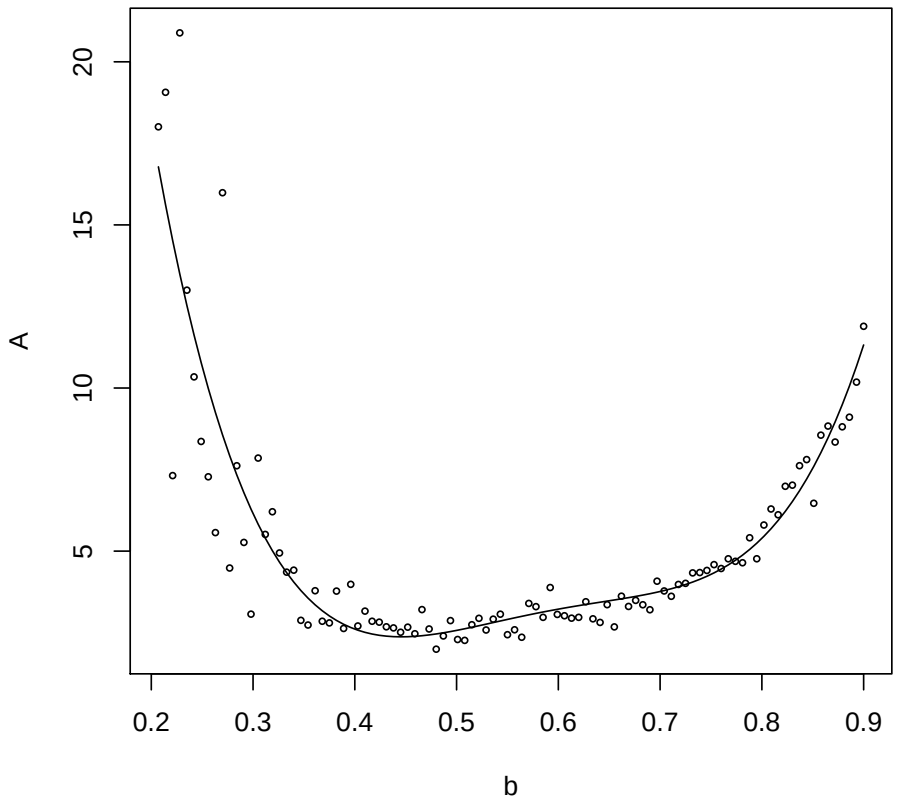


Figure 10.4: Fourth Degree Fit Superimposed

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 (10.17). Now, though, we need to add an assumption:[6]

$$Var(Y|X = t) = \sigma^2 \tag{10.33}$$

for all t. Note that this and the independence of the sample observations (e.g. the various people sampled in

[6] Actually, we could derive some usable, though messy,standard errors without this assumption.

the Davis height/weight example are independent of each other) implies that

$$Cov(V|Q) = \sigma^2 I \tag{10.34}$$

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 it's probably accurate enough for our purposes here.

We can derive the covariance matrix of $\hat{\beta}$ as follows. Again to avoid clutter, let $B = (Q'Q)^{-1}$. Theorem from linear algebra say 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:

$$\begin{aligned}
Cov(\hat{\beta}) &= Cov(BQ'V) \text{ ((10.26))} && (10.35)\\
&= BQ'Cov(V)(BQ')' \text{ (5.107)} && (10.36)\\
&= BQ'\sigma^2 I(BQ')' \text{ (10.34)} && (10.37)\\
&= \sigma^2 BQ'QB \text{ (lin. alg.)} && (10.38)\\
&= \sigma^2 (Q'Q)^{-1} \text{ (def. of B)} && (10.39)
\end{aligned}$$

Whew! That's a lot of work for you, if your linear algebra is rusty. But it's worth it, because (10.39) 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

$$\begin{aligned}
\sigma^2 &= Var(Y|X = t) && (10.40)\\
&= Var(Y|X^{(1)} = t_1, ..., X^{(r)} = t_r) && (10.41)\\
&= E\left[\{Y - m_{Y;X}(t)\}^2\right] && (10.42)\\
&= E\left[(Y - \beta_0 - \beta_1 t_1 - ... - \beta_r t_r)^2\right] && (10.43)
\end{aligned}$$

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:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i^{(1)} - ... - \hat{\beta}_r X_i^{(r)})^2 \tag{10.44}$$

As in Chapter 8, 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

$$\widehat{Cov}(\hat{\beta}) = s^2(Q'Q)^{-1} \tag{10.45}$$

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 7.2.4, included factors like $1/\sqrt{n}$, you might wonder why you don't see such a factor in (10.45).

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.

10.1.9.5 Once Again, Our ALOHA Example

In R we can obtain (10.45) via the generic function **vcov()**:

```
> vcov(lmout)
             (Intercept)    md4[, 1]   md4[, 3]   md4[, 4]   md4[, 5]
(Intercept)    92.73734   -794.4755   2358.860  -2915.238   1279.981
md4[, 1]     -794.47553   6896.8443 -20705.705  25822.832 -11422.355
md4[, 3]     2358.86046 -20705.7047  62804.912 -79026.086  35220.412
md4[, 4]    -2915.23828  25822.8320 -79026.086 100239.652 -44990.271
md4[, 5]     1279.98125 -11422.3550  35220.412 -44990.271  20320.809
```

What is this telling us? For instance, it says that the (4,4) position (starting at (0,0) in the matrix (10.45) is equal to 20320.809, so the standard error of $\widehat{\beta}_4$ is the square root of this, 142.6. Thus an approximate 95% confidence interval for the true population β_4 is

$$835.89714 \pm 1.96 \cdot 142.6 = (556.4, 1115.4) \tag{10.46}$$

That interval is quite wide. The margin of error, $1.96 \cdot 142.6 = 279.5$ is more than half of the left endpoint of the interval, 556.4. Remember what this tells us—that our sample of size 100 is not very large. On the other hand, the interval is quite far from 0, which indicates that our fourth-degree model is substantially better than our quadratic one.

Applying the R function **summary()** to a linear model object such as **lmout** here gives standard errors for the $\widehat{\beta}_i$ and lots of other information.

10.1.9.6 Estimation Vs. Prediction

In statistical parlance, there is a keen distinction made between the words *estimation* and *prediction*. To explain this, let's again consider the example of predicting Y = weight from X = (height,age). Say we have someone of height 67 inches and age 27, and want to guess—i.e. *predict*—her weight.

From Section 10.4.1, we know that the best prediction is m[(67,27)]. However, we do not know the value of that quantity, so we must *estimate* it from our data. So, our *predicted value* for this person's weight will be $\hat{m}[(67, 27)]$, i.e. our *estimate* for the value of the regression function at the point (67,27).

10.1.9.7 Exact Confidence Intervals

Note carefully that we have not assumed that Y, given X, is normally distributed. In the height/weight context, for example, such an assumption would mean that weights in a specific height subpopulation, say all people of height 70 inches, have a normal distribution.

If we do make such an assumption, then we can get exact confidence intervals (which of course, only hold if we really do have an exact normal distribution in the population). This again uses Student-t distributions. In that analysis, s^2 has n-(r+1) in its denominator instead of our n, just as there was n-1 in the denominator for s^2 when we estimated a single population variance. The number of degrees of freedom in the Student-t distribution is likewise n-(r+1). But as before, for even moderately large n, it doesn't matter.

10.1.10 Model Selection

The issues raised in Chapter 9 become crucial in regression and classification problems. In this chapter, we will typically deal with models having large numbers of parameters. A central principle will be that simpler models are preferable, provided of course they are accurate. Hence the Einstein quote in Chapter 9. Simpler models are often called **parsimonious**.

Here I use the term *model selection* to mean which predictor variables we will use. If we have data on many predictors, we almost certainly will not be able to use them all, for the following reason:

10.1.10.1 The Overfitting Problem in Regression

Recall (10.12). There we assumed a second-degree polynomial for $m_{A;b}$. Why not a third-degree, or fourth, and so on?

You can see that if we carry this notion to its extreme, we get absurd results. If we fit a polynomial of degree 99 to our 100 points, we can make our fitted curve exactly pass through every point! This clearly would give us a meaningless, useless curve. We are simply fitting the noise.

Recall that we analyzed this problem in Section 9.1.4 in our chapter on modeling. There we noted an absolutely fundamental principle in statistics:

> In choosing between a simpler model and a more complex one, the latter is more accurate only if either
>
> - we have enough data to support it, or
> - the complex model is sufficiently different from the simpler one

This is extremely important in regression analysis. For example, look at our regression model for A against b in the ALOHA simulation in earlier sections. We did analyses for a simpler model, a quadratic polynomial, and a more complex model, a quartic (polynomial of degree 4). Rephrasing the above points in this context, we would say,

> In choosing between the quadratic and quartic models, the latter is more accurate only if either
>
> - we have enough data to support it, or
> - at least one of the coefficients β_3 and β_4 is quite different from 0

In the weight/height/age example in Section 10.1.3, this would be phrased as

> In deciding whether to predict from height only, versus from both height and age, the latter is more accurate only if either
>
> - we have enough data to support it, or
> - the coefficient β_2 is quite different from 0

If we use too many predictor variables,[7], our data is "diluted," by being "shared" by so many β_i. As a result, $Var(\beta_i)$ will be large, with big implications: Whether our goal is Prediction or Understanding, our estimates will be so poor that neither goal is achieved.

The questions raised in turn by the above considerations, i.e. **How much** data is enough data?, and **How different** from 0 is "quite different"?, are addressed below in Section 10.1.10.2.

A detailed mathematical example of overfitting in regression is presented in my paper A Careful Look at the Use of Statistical Methodology in Data Mining (book chapter), by N. Matloff, in *Foundations of Data Mining and Granular Computing*, edited by T.Y. Lin, Wesley Chu and L. Matzlack, Springer-Verlag Lecture Notes in Computer Science, 2005.

[7] In the ALOHA example above, b, b^2, b^3 and b^4 are separate predictors, even though they are of course correlated.

10.1.10.2 Methods for Predictor Variable Selection

So, we typically must discard some, maybe many, of our predictor variables. In the weight/height/age example, we may need to discard the age variable. In the ALOHA example, we might need to discard b^4 and even b^3. How do we make these decisions?

Note carefully that **this is an unsolved problem.** If anyone ever claims they have a foolproof way to do this, they do not understand the problem in the first place. Entire books have been written on this subject (e.g. *Subset Selection in Regression*, by Alan Miller, pub. by Chapman and Hall, 2002), discussing myriad different methods, but again, none of them is foolproof.

Most of the methods for variable selection use hypothesis testing in one form or another. Typically this takes the form

$$H_0 : \beta_i = 0 \tag{10.47}$$

In the context of (10.10), for instance, a decision as to whether to include age as one of our predictor variables would mean testing

$$H_0 : \beta_2 = 0 \tag{10.48}$$

If we reject H_0, then we use the age variable; otherwise we discard it.

I hope I've convinced the reader, through Sections 7.4 and 9.2.1, that this is not a good idea. As usual, the hypothesis test is asking the wrong question. For instance, in the weight/height/age example, the test is asking whether β_2 is zero or not—yet we know it is not zero, before even looking at our data. *What we want to know* is whether β_2 is far enough from 0 for age to give us better predictions of weight. Those are two very, very different questions.

A very interesting example of overfitting using real data may be found in the paper, Honest Confidence Intervals for the Error Variance in Stepwise Regression, by Foster and Stine, `www-stat.wharton.upenn.edu/~stine/research/honests2.pdf`. The authors, of the University of Pennsylvania Wharton School, took real financial data and deliberately added a number of extra "predictors" that were in fact random noise, independent of the real data. They then tested the hypothesis (10.47). They found that each of the fake predictors was "significantly" related to Y! This illustrates both the dangers of hypothesis testing and the possible need for multiple inference procedures.[8] This problem has always been known by thinking statisticians, but the Wharton study certainly dramatized it.

Well, then, what can be done instead? First, there is the same alternative to hypothesis testing that we discussed before—confidence intervals. We saw an example of that in (10.46). Granted, the interval was

[8] They added so many predictors that r became greater than n. However, the problems they found would have been there to a large degree even if r were less than n but r/n was substantial.

very wide, telling us that it would be nice to have more data. But even the lower bound of that interval is far from zero, so it looks pretty safe to use b^4 as a predictor.

Moreover, a confidence interval for β_i tells us whether the variable $X^{(i)}$ would have much value as a predictor. Once again, consider the weight/height/age example. Suppose our confidence interval for β_2 is (0.04,0.06). In other words, we estimate β_2 to be 0.05, with a margin of error of 0.01. The 0.01 is telling us that our sample size is good enough for an accurate assessment of the situation, but the interval's location—centered at 0.05–says, for instance, a 10-year difference in age only makes about half a pound difference in mean weight. In that situation age would be of almost no value in predicting weight.

An example of this using real data is given in Section 10.2.3.2.

A method that enjoys some popularity in certain circles is the **Akaike Information Criterion** (AIC). It uses a formula, backed by some theoretical analysis, which creates a tradeoff between richness of the model and size of the standard errors of the $\hat{\beta}_i$. The R statistical package includes a function **AIC()** for this, which is used by **step()** in the regression case.

The most popular alternative to hypothesis testing for variable selection today is probably **cross validation**. Here we split our data into a **training set**, which we use to estimate the β_i, and a **validation set**, in which we see how well our fitted model predicts new data, say in terms of average squared prediction error. We do this for several models, i.e. several sets of predictors, and choose the one which does best in the validation set. I like this method very much, though I often simply stick with confidence intervals.

A rough rule of thumb is that one should have $r < \sqrt{n}$.[9]

10.1.11 Nonlinear Parametric Regression Models

We pointed out in Section 10.1.9.1 that the word *linear* in *linear regression model* means linear in β, not in t. This is the most popular approach, as it is computationally easy, but nonlinear models are often used.

The most famous of these is the **logistic** model, for the case in which Y takes on only the values 0 and 1. As we have seen before (Section 3.6), in this case the expected value becomes a probability. The logistic model for a nonvector X is then

$$m_{Y;X}(t) = P(Y = 1|X = t) = \frac{1}{1 + e^{-(\beta_0+\beta_1 t)}} \tag{10.49}$$

It extends to the case of vector-valued X in the obvious way.

The logistic model is quite widely used in computer science, in medicine, economics, psychology and so on.

[9] Asymptotic Behavior of Likelihood Methods for Exponential Families When the Number of Parameters Tends to Infinity, Stephen Portnoy, *Annals of Statistics*, 1968.

Here is an example of a nonlinear model used in kinetics of chemical reactions, with r = 3: [10]

$$m_{Y;X}(t) = \frac{\beta_1 t^{(2)} - t^{(3)}/\beta_5}{1 + \beta_2 t^{(1)} + \beta_3 t^{(2)} + \beta_4 t^{(3)}} \tag{10.50}$$

Here the X vector is (hydrogen, n-pentane, isopentane) and Y is the reaction rate.

Unfortunately, in most cases, the least-squares estimates of the parameters in nonlinear regression do not have closed-form solutions, and numerical methods must be used. But R does that for you, via the **nls()** function in general, and via **glm()** for the logistic and related models in particular.

10.1.12 Regression Diagnostics

Researchers in regression analysis have devised some **diagnostic** methods, meaning methods to check the fit of a model, the validity of assumptions [e.g. (10.33)], search for data points that may have an undue influence (and may actually be in error), and so on.

The R package has tons of diagnostic methods. See for example Chapter 4 of *Linear Models with R*, Julian Faraway, Chapman and Hall, 2005.

10.1.13 Nominal Variables

Recall our example in Section 10.1.2 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. 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 $X^{(2)} = X^{(3)} = 0$.

Here we are dealing with a **nominal** variable, Language, which has three values, C++, Java and C, and representing it by the two indicator variables $X^{(2)}$ and $X^{(3)}$. 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.

You can see that if a nominal variable takes on q values, we need q-1 indicator variables to represent it.

[10] See `http://www.mathworks.com/access/helpdesk/help/toolbox/stats/rsmdemo.html`.

We say that the variable has q **levels**. Note carefully that although we speak of this as one variable, it is implemented as q-1 variables.

10.1.14 The Case in Which All Predictors Are Nominal Variables: Analysis of "Variance"

(Note to readers: The material in this section is arguably of lesser value to computer science. As such, it can easily be skipped. However, it does provide motivation for our treatment of the log-linear model in Section 10.5.4.)

Continuing the ideas in Section 10.1.13, suppose in the software engineering study they had kept the project size constant, and instead of $X^{(1)}$ being project size, this variable recorded whether the programmer uses an integrated development environment (IDE). Say $X^{(1)}$ is 1 or 0, depending on whether the programmer uses the Eclipse IDE or no IDE, respectively. Continue to assume the study included the nominal Language variable, i.e. assume the study included the indicator variables $X^{(2)}$ (C++) and $X^{(3)}$ (Java). Now all of our predictors would be nominal/indicator variables. Regression analysis in such settings is called **analysis of variance** (ANOVA).

Each nominal variable is called a **factor**. So, in our software engineering example, the factors are IDE and Language. Note again that in terms of the actual predictor variables, each factor is represented by one or more indicator variables; here IDE has one indicator variables and Language has two.

Analysis of variance is a classic statistical procedure, used heavily in agriculture, for example. We will not go into details here, but mention it briefly both for the sake of completeness and for its relevance to Sections 10.1.6 and 10.5.4. (The reader is strongly advised to review Sections 10.1.6 before continuing.)

10.1.14.1 It's a Regression!

The term *analyisis of variance* is a misnomer. A more appropriate name would be **analysis of means**, as it is in fact a regression analysis, as follows.

First, note in our software engineering example we basically are talking about six groups, because there are six different combinations of values for the triple $(X^{(1)}, X^{(2)}, X^{(3)})$. For instance, the triple (1,0,1) means that the programmer is using an IDE and programming in Java. Note that triples of the form (w,1,1) are impossible.

So, all that is happening here is that we have six groups with six means. But that is a regression! Remember, for variables U and V, $m_{V;U}(t)$ is the mean of all values of V in the subpopulation group of people (or cars or whatever) defined by U = s. If U is a continuous variable, then we have infinitely many such groups, thus infinitely many means. In our software engineering example, we only have six groups, but the principle is

the same. We can thus cast the problem in regression terms:

$$m_{Y;X}(i,j,k) = E(Y|X^{(1)} = i, X^{(2)} = j, X^{(3)} = k),\ i,j,k = 0,1, j+k \leq 1 \tag{10.51}$$

Note the restriction $j + k \leq 1$, which reflects the fact that j and k can't both be 1.

Again, keep in mind that we are working with means. For instance, $m_{Y;X}(0,1,0)$ is the population mean project completion time for the programmers who do not use Eclipse and who program in C++.

Since the triple (i,j,k) can take on only six values, m can be modeled fully generally in the following six-parameter linear form:

$$m_{Y;X}(i,j,k) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 ij + \beta_5 ik \tag{10.52}$$

where β_4 and β_5 are the coefficients of two interaction terms, as in Section 10.1.6.

10.1.14.2 Interaction Terms

It is crucial to understand the interaction terms. Without the ij and ik terms, for instance, our model would be

$$m_{Y;X}(i,j,k) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k \tag{10.53}$$

which would mean (as in Section 10.1.6) that the difference between using Eclipse and no IDE is the same for all three programming languages, C++, Java and C. That common difference would be β_1. If this condition—the impact of using an IDE is the same across languages—doesn't hold, at least approximately, then we would use the full model, (10.52). More on this below.

Note carefully that there is no interaction term corresponding to jk, since that quantity is 0, and thus there is no three-way interaction term corresponding to ijk either.

But suppose we add a third factor, Education, represented by the indicator $X^{(4)}$, having the value 1 if the programmer has a least a Master's degree, 0 otherwise. Then m would take on 12 values, and the full model would have 12 parameters:

$$m_{Y;X}(i,j,k,l) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 l + \beta_5 ij + \beta_6 ik + \beta_7 il + \beta_8 jl + \beta_9 kl + \beta_{10} ijl + \beta_{11} ikl \tag{10.54}$$

Again, there would be no ijkl term, as jk = 0.

Here β_1, β_2, β_3 and β_4 are called the **main effects**, as opposed to the coefficients of the interaction terms, called of course the **interaction effects**.

The no-interaction version would be

$$m_{Y;X}(i,j,k,l) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 l \tag{10.55}$$

10.1.14.3 Now Consider Parsimony

In the three-factor example above, we have 12 groups and 12 means. Why not just treat it that way, instead of applying the powerful tool of regression analysis? The answer lies in our desire for parsimony, as noted in Section 10.1.10.1.

If for example (10.55) were to hold, at least approximately, we would have a far more satisfying model. We could for instance then talk of "the" effect of using an IDE, rather than qualifying such a statement by stating what the effect would be for each different language and education level. Moreover, if our sample size is not very large, we would get more accurate estimates of the various subpopulation means, once again due to bias/variance tradeoff.

Or it could be that, while (10.55) doesn't hold, a model with only two-way interactions,

$$m_{Y;X}(i,j,k,l) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 l + \beta_5 ij + \beta_6 ik + \beta_7 il + \beta_8 jl + \beta_9 kl \tag{10.56}$$

does work well. This would not be as nice as (10.55), but it still would be more parsimonious than (10.54).

Accordingly, the major thrust of ANOVA is to decide how rich a model is needed to do a good job of describing the situation under study. There is an implied hierarchy of models of interest here:

- the full model, including two- and three-way interactions, (10.54)
- the model with two-factor interactions only, (10.56)
- the no-interaction model, (10.55)

Traditionally these are determined via hypothesis testing, which involves certain partitionings of sums of squares similar to (10.18). (This is where the name *analysis of variance* stems from.) The null distribution of the test statistic often turns out to be an F-distribution. Of course, in this book, we consider hypothesis testing inappropriate, preferring to give some careful thought to the estimated parameters, but it is standard. Further testing can be done on individual β_1 and so on. Often people use simultaneous inference procedures, discussed briefly in Section 8.7 of our chapter on estimation and testing, since many tests are performed.

10.1.14.4 Reparameterization

Classical ANOVA uses a somewhat different parameterization than that we've considered here. For instance, consider a single-factor setting (called **one-way ANOVA**) with three levels. Our predictors are then $X^{(1)}$ and $X^{(2)}$. Taking our approach here, we would write

$$m_{Y;X}(i,j) = \beta_0 + \beta_1 i + \beta_2 j \tag{10.57}$$

The traditional formulation would be

$$\mu_i = \mu + \alpha_i,\ i = 1,2,3 \tag{10.58}$$

where

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3}{3} \tag{10.59}$$

and

$$\alpha_i = \mu_i - \mu \tag{10.60}$$

Of course, the two formulations are equivalent. It is left to the reader to check that, for instance,

$$\mu = \beta_0 + \frac{\beta_1 + \beta_2}{2} \tag{10.61}$$

There are similar formulations for ANOVA designs with more than one factor.

Note that the classical formulation overparameterizes the problem. In the one-way example above, for instance, there are four parameters (μ, α_1, α_2, α_3) but only three groups. This would make the system indeterminate, but we add the constraint

$$\sum_{i=1}^{3} \alpha_i = 0 \tag{10.62}$$

Equation (10.26) then must make use of **generalized matrix inverses**.

10.1.15 The Famous "Error Term"

Books on linear regression analysis—and there are hundreds, if not thousands of these—generally introduce the subject as follows. They consider the linear case with r = 1, and write

$$Y = \beta_0 + \beta_1 X + \epsilon,\ E\epsilon = 0 \tag{10.63}$$

with ϵ being independent of X. They also assume that ϵ has a normal distribution with variance σ^2.

Let's see how this compares to what we have been assuming here so far. In the linear case with r = 1, we would write

$$m_{Y;X}(t) = E(Y|X = t) = \beta_0 + \beta_1 t \tag{10.64}$$

Note that in our context, we could define ϵ as

$$\epsilon = Y - m_{Y;X}(X) \tag{10.65}$$

Equation (10.63) is consistent with (10.64): The former has $E\epsilon = 0$, and so does the latter, since

$$E\epsilon = EY - E[m_{Y;X}(X)] = EY - E[E(Y|X)] = EY - EY = 0 \tag{10.66}$$

In order to produce confidence intervals, we later added the assumption (10.33), which you can see is consistent with (10.63) since the latter assumes that $Var(\epsilon) = \sigma^2$ no matter what value X has.

Now, what about the normality assumption in (10.63)? That would be equivalent to saying that in our context, the conditional distribution of Y given X is normal, which is an assumption we did not make. Note that in the weight/height example, this assumption would say that, for instance, the distribution of weights among people of height 68.2 inches is normal.

No matter what the context is, the variable ϵ is called the **error term**. Originally this was an allusion to measurement error, e.g. in chemistry experiments, but the modern interpretation would be prediction error, i.e. how much error we make when we us $m_{Y;X}(t)$ to predict Y.

10.2 The Classification Problem

As mentioned earlier, in the special case in which Y is an indicator variable, with the value 1 if the object is in a class and 0 if not, the regression problem is called the **classification problem**. In electrical engineering it

is called **pattern recognition**, and the predictors are called **features**. In computer science the term **machine learning** usually refers to classification problems. Different terms, same concept.

If there are c classes, we need c (or c-1) Y variables, which I will denote by $Y^{(i)}$, i = 1,...,c.

Here are some examples:

- A forest fire is now in progress. Will the fire reach a certain populated neighborhood? Here Y would be 1 if the fire reaches the neighborhood, 0 otherwise. The predictors might be wind direction, distance of the fire from the neighborhood, air temperature and humidity, and so on.
- Is a patient likely to develop diabetes? This problem has been studied by many researchers, e.g. Using Neural Networks To Predict the Onset of Diabetes Mellitus, Murali S. Shanker *J. Chem. Inf. Comput. Sci.*, 1996, 36 (1), pp 3541. A famous data set involves Pima Indian women, with Y being 1 or 0, depending on whether the patient does ultimately develop diabetes, and the predictors being the number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, serum insulin level, body mass index, diabetes pedigree function and age.
- Is a disk drive likely to fail sure? This has been studied for example in Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application, by Joseph F. Murray, Gordon F. Hughes, and Kenneth Kreutz-Delgado, *Journal of Machine Learning Research* 6 (2005) 783-816. Y was 1 or 0, depending on whether the drive failed, and the predictors were temperature, number of read errors, and so on.

10.2.1 The Mean Here Is a Probability

Now, here is a key point: As we have frequently noted the mean of any indicator random variable is the probability that the variable is equal to 1 (Section 3.6). Thus in the case in which our response variable Y takes on only the values 0 and 1, i.e. classification problems, the regression function reduces to

$$m_{Y;X}(t) = P(Y = 1|X = t) \tag{10.67}$$

(Remember that X and t are vector-valued.)

As a simple but handy example, suppose Y is gender (1 for male, 0 for female), $X^{(1)}$ is height and $X^{(2)}$ is weight, i.e. we are predicting a person's gender from the person's height and weight. Then for example, $m_{Y;X}(70, 150)$ is the probability that a person of height 70 inches and weight 150 pounds is a man. Note again that this probability is a population fraction, the fraction of men among all people of height 70 and weight 150 in our population.

Make a mental note of the optimal prediction rule, if we know the population regression function:

Given X = t, the optimal prediction rule is to predict that Y = 1 if and only if $m_{Y;X}(t) > 0.5$.

So, if we known a certain person is of height 70 and weight 150, our best guess for the person's gender is to predict the person is male if and only if $m_{Y;X}(70, 150) > 0.5$.

The optimality makes intuitive sense, and is proved in Section 10.4.2.

10.2.2 Logistic Regression: a Common Parametric Model for the Regression Function in Classification Problems

Remember, we often try a parametric model for our regression function first, as it means we are estimating a finite number of quantities, instead of an infinite number. Probably the most commonly-used model is that of the logistic function (often called "logit"), introduced in Section 10.1.11. Its r-predictor form is

$$m_{Y;X}(t) = P(Y = 1|X = t) = \frac{1}{1 + e^{-(\beta_0+\beta_1 t_1+...+\beta_r t_r)}} \tag{10.68}$$

For instance, consider the patent example in Section 10.1.2. Under the logistic model, the population proportion of all patents that are publicly funded, among those that contain the word "NSF," do not contain "NIH," and make five claims would have the value

$$\frac{1}{1 + e^{-(\beta_0+\beta_1+5\beta_3)}} \tag{10.69}$$

10.2.2.1 The Logistic Model: Intuitive Motivation

The logistic function itself,

$$\frac{1}{1 + e^{-u}} \tag{10.70}$$

has values between 0 and 1, and is thus a candidate for modeling a probability. Also, it is monotonic in u, making it further attractive, as in many classification problems we believe that $m_{Y;X}(t)$ should be monotonic in the predictor variables.

10.2.2.2 The Logistic Model: Theoretical Motivation

But there are much stronger reasons to use the logit model, as it includes many common parametric models for X. To see this, note that we can write, for vector-valued discrete X and t,

$$
\begin{aligned}
P(Y=1|X=t) &= \frac{P(Y=1 \text{ and } X=t)}{P(X=t)} && (10.71)\\
&= \frac{P(Y=1)P(X=t|Y=1)}{P(X=t)} && (10.72)\\
&= \frac{P(Y=1)P(X=t|Y=1)}{P(Y=1)P(X=t|Y=1)+P(Y=0)P(X=t|Y=0)} && (10.73)\\
&= \frac{1}{1+\frac{(1-q)P(X=t|Y=0)}{qP(X=t|Y=1)}} && (10.74)
\end{aligned}
$$

where $q = P(Y = 1)$ is the proportion of members of the population which have $Y = 1$. (Keep in mind that this probability is unconditional!!!! In the patent example, for instance, if say $q = 0.12$, then 12% of all patents in the patent population—without regard to words used, numbers of claims, etc.—are publicly funded.)

If X is a continuous random vector, then the analog of (10.74) is

$$P(Y=1|X=t) = \frac{1}{1+\frac{(1-q)f_{X|Y=0}(t)}{qf_{X|Y=1}(t)}} \tag{10.75}$$

Now suppose X, given Y, has a normal distribution. In other words, within each class, Y is normally distributed. Consider the case of just one predictor variable, i.e. r = 1. Suppose that given $Y = i$, X has the distribution $N(\mu_i, \sigma^2)$, i = 0,1. Then

$$f_{X|Y=i}(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-0.5\left(\frac{t-\mu_i}{\sigma}\right)^2\right] \tag{10.76}$$

After doing some elementary but rather tedious algebra, (10.75) reduces to the logistic form

$$\frac{1}{1+e^{-(\beta_0+\beta_1 t)}} \tag{10.77}$$

where β_0 and β_1 are functions of μ_0, μ_0 and σ.

In other words, if X is normally distributed in both classes, with the same variance but different means, then $m_{Y;X}$ has the logistic form! And the same is true if X is multivariate normal in each class, with different mean vectors but equal covariance matrices. (The algebra is even more tedious here, but it does work out.)

So, not only does the logistic model have an intuitively appealing form, it is also implied by one of the most famous distributions X can have within each class—the multivariate normal.

If you reread the derivation above, you will see that the logit model will hold for any within-class distributions for which

$$\ln\left(\frac{f_{X|Y=0}(t)}{f_{X|Y=1}(t)}\right) \tag{10.78}$$

(or its discrete analog) is linear in t. Well guess what—this condition is true for exponential distributions too! Work it out for yourself.

In fact, a number of famous distributions imply the logit model.

10.2.3 Variable Selection in Classification Problems

10.2.3.1 Problems Inherited from the Regression Context

In Section 10.1.10.2, it was pointed out that the problem of predictor variable selection in regression is unsolved. Since the classification problem is a special case of regression, there is no surefire way to select predictor variables there either.

10.2.3.2 Example: Forest Cover Data

And again, using hypothesis testing to choose predictors is not the answer. To illustrate this, let's look again at the forest cover data we saw in Section 7.2.7.

There were seven classes of forest cover there. Let's restrict attention to classes 1 and 2. In my R analysis I had the class 1 and 2 data in objects **cov1** and **cov2**, respectively. I combined them,

```
> cov1and2 <- rbind(cov1,cov2)
```

and created a new variable to serve as Y:

```
cov1and2[,56] <- ifelse(cov1and2[,55] == 1,1,0)
```

Let's see how well we can predict a site's class from the variable HS12 (hillside shade at noon) that we investigated in that past chapter, using a logistic model.

In R we fit logistic models via the **glm()** function, for generalized linear models. The word *generalized* here refers to models in which some function of $m_{Y;X}(t)$ is linear in parameters β_i. For the classification model,

$$\ln\left(m_{Y;X}(t)/[1-m_{Y;X}(t)]\right) = \beta_0 + \beta_1 t^{(1)} + ... + \beta_r t^{(r)} \tag{10.79}$$

This kind of generalized linear model is specified in R by setting the named argument **family** to **binomial**. Here is the call:

```
> g <- glm(cov1and2[,56] ~ cov1and2[,8],family=binomial)
```

The result was:

```
> summary(g)

Call:
glm(formula = cov1and2[, 56] ~ cov1and2[, 8], family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.165  -0.820  -0.775   1.504   1.741

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.515820   1.148665   1.320   0.1870
cov1and2[, 8] -0.010960   0.005103  -2.148   0.0317 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 959.72  on 810  degrees of freedom
Residual deviance: 955.14  on 809  degrees of freedom
AIC: 959.14

Number of Fisher Scoring iterations: 4
```

So, $\widehat{\beta}_1 = -0.01$. This is tiny, as can be seen from our data in the last chapter. There we found that the estimated mean values of HS12 for cover types 1 and 2 were 223.8 and 226.3, a difference of only 2.5. That difference in essence gets multiplied by 0.01. More concretely, in (10.49), plug in our estimates 1.52 and -0.01 from our R output above, first taking t to be 223.8 and then 226.3. The results are 0.328 and 0.322, respectively. In other words, HS12 isn't having much effect on the probability of cover type 1, and so it cannot be a good predictor of cover type.

Yet the R output says that β_1 is "significantly" different from 0, with a p-value of 0.03. Thus, we see once again that hypothesis testing does not achieve our goal. Again, cross validation is a better method for choosing predictors.

10.2.4 Y Must Have a Marginal Distribution!

In our material here, we have tacitly assumed that the vector (Y,X) has a distribution. That may seem like an odd and puzzling remark to make here, but **it is absolutely crucial**. Let's see what it means.

Consider the study on object-oriented programming in Section 10.1.1, but turned around. (This example will be somewhat contrived, but it will illustrate the principle.) Suppose we know how many lines of code are in a project, which we will still call $X^{(1)}$, and we know how long it took to complete, which we will now take as $X^{(2)}$, and from this we want to guess whether object-oriented or procedural programming was used (without being able to look at the code, of course), which is now our new Y.

Here is our huge problem: Given our sample data, there is no way to estimate q in (10.74). That's because the authors of the study simply took two groups of programmers and had one group use object-oriented programming and had the other group use procedural programming. If we had sampled programmers at random from actual projects done at this company, that would enable us to estimate q, the population proportion of projects done with OOP. But we can't do that with the data that we do have. Indeed, in this setting, it may not even make sense to speak of q in the first place.

Mathematically speaking, if you think about the process under which the data was collected in this study, there does exist some conditional distribution of X given Y, but Y itself has no distribution. So, we can NOT estimate P(Y=1|X). About the best we can do is try to guess Y on the basis of whichever value of i makes $f_{X|Y=i}(X)$ larger.

10.3 Nonparametric Estimation of Regression and Classification Functions

In some applications, there may be no good parametric model, say linear or logistic, for $m_{Y;X}$. Or, we may have a parametric model that we are considering, but we would like to have some kind of nonparametric estimation method available as a means of checking the validity of our parametric model. So, how do we estimate a regression function nonparametrically?

Many, many methods have been developed. We introduce a few here.

10.3.1 Methods Based on Estimating $m_{Y;X}(t)$

To guide our intuition on this, let's turn again to the example of estimating the relationship between height and weight. Consider estimation of the quantity $m_{W;H}(68.2)$, the *population* mean weight of all people of height 68.2.

10.3.1.1 Kernel-Based Methods

We could take our estimate of $m_{W;H}(68.2)$, $\widehat{m}_{W;H}(68.2)$, to be the average weight of all the people in our sample who have that height. But we may have very few people of that height (or even none), so that our estimate may have a high variance, i.e. may not be very accurate.

What we could do instead is to take the mean weight of all the people in our sample whose heights are *near* 68.2, say between 67.7 and 68.7. That would bias things a bit, but we'd get a lower variance. This is again an illustration of the variance/bias tradeoff introduced in Section 9.1.3.

All nonparametric regression/classification methods work like this, though with many variations. (As noted earlier, the classification problem is a special case of regression, so in the following material we will usually not distinguish between the two.)

As our definition of "near," we could take all people in our sample whose heights are within h amount of 68.2. This should remind you of our density estimators in Section 8.4 of our chapter on estimation and testing. As we saw there, a generalization would be to use a **kernel** method. For instance, for univariate X and t:

$$\widehat{m}_{Y;X}(t) = \frac{\sum_{i=1}^{n} Y_i k\left(\frac{t-X_i}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{t-X_i}{h}\right)} \tag{10.80}$$

This looks imposing, but it is simply a weighted average of the Y values in our sample, with the larger weights being placed on observations for which X is close to t.

As before, the choice of h here involves a bias/variance tradeoff. We might try choosing h via cross validation, as discussed in Section 10.1.10.2.

There is an R package that includes a function **nkreg()** for kernel regression. The R base has a similar method, called **LOESS**. Note: That is the class name, but the R function is called **lowess()**.

10.3.1.2 Nearest-Neighbor Methods

Similarly, we could take a **nearest-neighbor** approoach, for instance estimating $m_{Y;X}(68.2)$ to be the mean weight of the k people in our sample with heights nearest 68.2. Here k controls bias/variance tradeoff.

10.3.1.3 The Naive Bayes Method

The NB method is not "Bayesian" in the sense of Section 8.9. Instead, its name comes simply from its usage of Bayes' Rule for conditional probability. It basically makes the same computations as in Section 10.2.2.2,

for the case in which the predictors are indicator variables and are independent of each other, given the class.

Under that assumption, the numerator in (10.74) becomes

$$P(Y = 1) \ P[X^{(1)} = t_1 | Y = 1] \ ... \ P[X^{(r)} = t_r | Y = 1] \tag{10.81}$$

All of those quantities (and similarly, those in the denominator of (10.74) can be estimated directly as sample proportions. For example, $\widehat{P}[X^{(1)} = t_1 | Y = 1]$ would be the fraction of $X_j^{(1)}$ that are equal to t_1, among those observations for which $Y_j = 1$.

A common example of the use of Naive Bayes is text mining, as in Section 5.8.1.3. Our independence assumption in this case means that the probability that, for instance, a document of a certain class contains both of the words *baseball* and *strike* is the product of the individual probabilities of those words.

Clearly the independence assumption is not justified in this application. But if our vocabulary is large, that assumption limits the complexity of our model, which may be necessary from a bias/variance tradeoff point of view (Section 9.1.3).

10.3.2 Methods Based on Estimating Classification Boundaries

In the methods presented above, we are estimating the function $m_{Y;X}(t)$. But with support vector machines and CART below, we are in a way working backwards. In the classification case (which is what we will focus on), for instance, our goal is to estimate the values of t for which the regression function equals 0.5:

$$B = \{t : m_{Y;X}(t) = 0.5\} \tag{10.82}$$

Recall that r is the number of predictor variables we have. Then note the geometric form that the set B in (10.82) will take on: discrete points if r = 1; a curve if r = 2; a surface if r = 3; and a hypersurface if r $>$ 3.

The motivation for using (10.82) stems from the fact, noted in Section 10.2.1, that if we know $m_{Y;X}(t)$, we will predict Y to be 1 if and only if $m_{Y;X}(t) > 0.5$. Since (10.82) represents the boundary between the portions of the X space for which $m_{Y;X}(t)$ is either larger or smaller than 0.5, it is the boundary for our prediction rule, i.e. the boundary separating the regions in X space in which we predict Y to be 1 or 0.

Lest this become too abstract, again consider the simple example of predicting gender from height and weight. Consider the (u,v) plane, with u and v representing height and weight, respectively. Then (10.82) is some curve in that plane. If a person's (height,weight) pair is on one side of the curve, we guess that the person is male, and otherwise guess female.

If the logistic model (10.68) holds, then that curve is actually a straight line. To see this, note that in (10.68),

the equation (10.82) boils down to

$$\beta_0 + \beta_1 u + \beta_2 v = 0 \tag{10.83}$$

whose geometric form is a straight line.

10.3.2.1 Support Vector Machines (SVMs)

This method has been getting a lot of publicity in computer science circles (maybe too much; see below). It is better explained for the classification case.

In the form of dot product (or inner product) from linear algebra, (10.83) is

$$(\beta_1, \beta_2)'(u, v) = -\beta_0 \tag{10.84}$$

What SVM does is to generalize this, for instance changing the criterion to, say

$$\beta_0 u^2 + \beta_1 uv + \beta_2 v^2 + \beta_3 u + \beta_4 v = 1 \tag{10.85}$$

Now our (u, v) plane is divided by a curve instead of by a straight line (though it includes straight lines as special cases), thus providing more flexibility and thus potentially better accuracy.

In SVM terminology, (10.85) uses a different **kernel** than regular dot product. (This of course should not be confused our the term *kernel* in kernel-based regression above.) The actual method is more complicated than this, involving transforming the original predictor variables and then using an ordinary inner product in the transformed space. In the above example, the transformation consists of squaring and multiplying our variables. That takes us from two-dimensional space (just u and v) to five dimensions (u, v, u^2, v^2 and uv).

There are various other details that we've omitted here, but the essence of the method is as shown above.

Of course, a good choice of the kernel is crucial to the successful usage of this method. It is the analog of h and k in the nearness-based methods above.

10.3.2.2 CART

Another nonparametric method is that of **Classification and Regression Trees** (CART). It's again easiest explained in the classification context, say the diabetes example above.

In the diabetes example, we might try to use glucose variable as our first predictor. The data may show that a high glucose value implies a high likelihood of developing diabetes, while a low value does the opposite. We

would then find a **split** on this variable, meaning a a cutoff value that defines "high" and "low." Pictorially, we draw this as the root of a tree, with the left branch indicating a tentative guess of no diabetes and the right branch corresponding to a guess of diabetes.

Actually, we could do this for all our predictor variables, and find which one produces the best split at the root stage. But let's assume that we find that glucose is that variable.

Now we repeat the process. For the left branch—all the subset of our data corresponding to "low" glucose—we find the variable that best splits that branch, say body mass index. We do the same for the right branch, say finding that age gives the best split. We keep going until the resulting cells are too small for a reasonable split.

it is either really high or really low, we predict diabetes from this information alone and stop. If not, we then look at body mass index, and so on.

An example with real data is given in a tutorial on the use of **rpart**, an R package that does analysis of the CART type, *An Introduction to Recursive Partitioning Using the RPART Routines*, by Terry Therneau and Elizabeth Atkinson. The data was on treatment of cardiac arrest patients by emergency medical technicians.

The response variable here is whether the technicians were able to revive the patient, with predictors $X^{(1)}$ = initial heart rhythm, $X^{(2)}$ = initial response to defrillabration, and $X^{(3)}$ = initial response to drugs. The resulting tree was

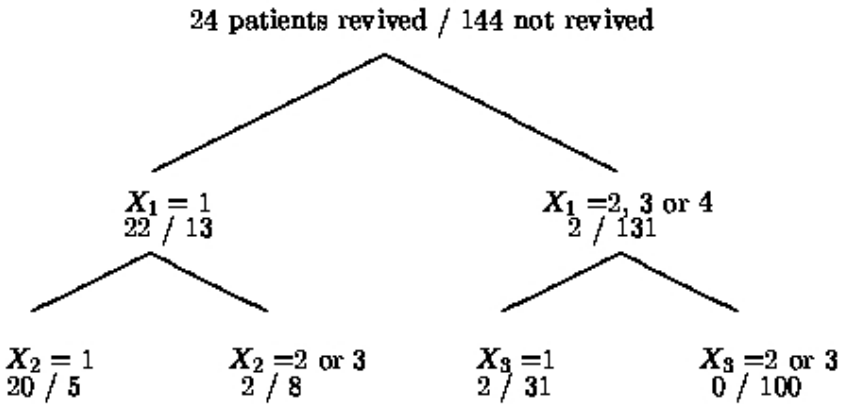
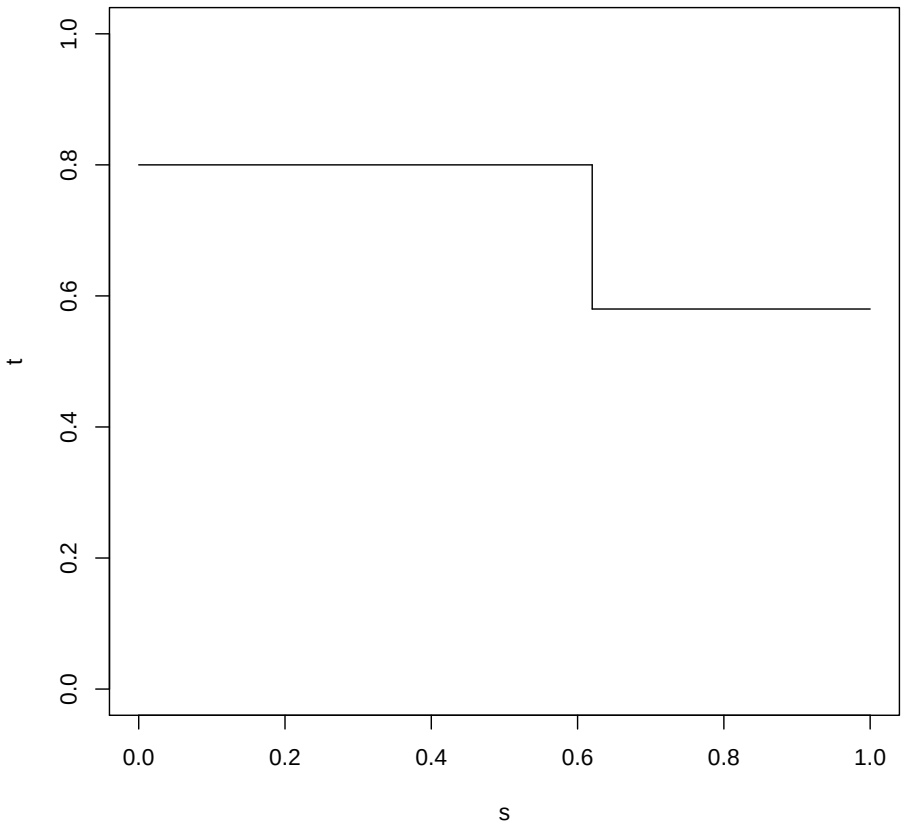


Figure 1: Revival data

So, if for example a patient has $X^{(1)} = 1$ and $X^{(2)} = 3$, we would guess him to be revivable.

CART is a boundary method, as SVM is. Say for instance we have two variables, represented graphically by s and t, and our root node rule is $s > 0.62$. In the left branch, the rule is $t > 0.8$ and in the right branch

it's $t > 0.58$. This boils down to a boundary line as follows:



CART obviously has an intuitive appeal, easily explained to nonstatisticians, and easy quite easy to implement. It also has the virtue of working equally well with discrete or continuous predictor variables.

The analogs here of the h in the kernel method and k in nearest-neighbor regression are the choice of where to define the splits, and when to stop splitting. Cross validation is often used for making such decisions.

10.3.3 Comparison of Methods

Beware! There are no "magic" solutions to statistical problems. The statements one sees by some computer science researchers to the effect that SVMs are generally superior to other prediction methods are **unfounded**.

First, note that every one of the above methods involves some choice of tuning parameter, such as h in the kernel method, k in the nearest-neighbor method, the split points in CART, and in the case of SVM, the form of which kernel to use. For SVM the choice of kernel is crucial, yet difficult.

Second, the comparisons are often unfair, notably comparisons of the logit model to SVM. Such comparisons usually limit the logit experiments to first-degree terms without interactions. But in (10.68) we could throw in second-degree terms, etc., thus producing a curved partitioning line just like SVM does.

I highly recommend the site `www.dtreg.com/benchmarks.htm`, which compares six different types of classification function estimators—including logistic regression and SVM—on several dozen real data sets. The overall percent misclassification rates, averaged over all the data sets, was fairly close, ranging from a high of 25.3% to a low of 19.2%. The much-vaunted SVM came in at 20.3%. That's nice, but it was only a tad better than logit's 20.9%—and remember, that's with logit running under the handicap of having only first-degree terms.

Or consider the annual KDDCup competition, in which teams from around the world compete to solve a given classification problem with the lowest misclassification rate. In KDDCup2009, for instance, none of the top teams used SVM. See *SIGKDD Explorations*, December 2009 issue.

Considering that logit has a big advantage in that one gets an actual equation for the classification function, complete with parameters which we can estimate and make confidence intervals for, it is not clear just what role SVM and the other nonparametric estimators should play, in general, though in specific applications they may be appropriate.

10.4 Optimality Issues

Being "optimal" is highly dependent on models being correct and appropriate, but optimality does give us further confidence in a model. In this section, we present two optimality results.

10.4.1 Optimality of the Regression Function for General Y

In predicting Y from X (with X random), we might assess our predictive ability by the **mean squared prediction error** (MSPE):

$$\text{MSPE} = E\left[(Y - w(X))^2\right] \tag{10.86}$$

where w is some function we will use to form our prediction for Y based on X. What w is best, i.e. which w minimizes MSPE?

To answer this question, condition on X in (10.86):

$$\text{MSPE} = E\left[E\{(Y - w(X))^2|X\}\right] \tag{10.87}$$

Theorem 34 *The best w is m, i.e. the best way to predict Y from X is to "plug in" X in the regression function.*

Recall from Section 9.1.1:

Lemma 35 *For any random variable Z, the constant c which minimizes*

$$E[(Z - c)^2] \tag{10.88}$$

is

$$c = EZ \tag{10.89}$$

Apply the lemma to the inner expectation in (10.87), with Z being Y and c being some function of X. The minimizing value is EZ, i.e. $E(Y|X)$ since our expectation here is conditional on X.

All of this tells us that the best function w in (10.86) is $m_{Y;X}$. This proves the theorem.

Note carefully that all of this was predicated on the use of a quadratic loss function, i.e. on minimizing mean squared error. If instead we wished to minimize mean absolute error, the solution would turn out to be to use the conditional median of Y given X, not the mean.

10.4.2 Optimality of the Regression Function for 0-1-Valued Y

Again, our context is that we want to guess Y, knowing X. Since Y is 0-1 valued, our guess for Y based on X, g(X), should be 0-1 valued too. What is the best g?

Again, since Y and g are 0-1 valued, our criterion should be what will I call Probability of Correct Classification (PCC):[11]

$$\text{PCC} = P[Y = g(X)] \tag{10.90}$$

[11] This assumes equal costs for the two kinds of classification errors, i.e. that guessing Y = 1 when Y = 0 is no more or no less serious than the opposite error.

Now proceed as in (10.87):

$$\text{PCC} = E\left[P\{Y = g(X)|X\}\right] \tag{10.91}$$

The analog of Lemma 35 is

Lemma 36 *Suppose W takes on values in the set A = {0,1}, and consider the problem of maximizing*

$$P(W = c),\ c \epsilon A \tag{10.92}$$

The solution is

$$\begin{cases} 1, & \text{if } P(W = 1) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{10.93}$$

Proof

Again recalling that c is either 1 or 0, we have

$$\begin{aligned} P(W = c) &= P(W = 1)c + [1 - P(W = 1)]\,(1 - c) && (10.94) \\ &= [2P(W = 1) - 1]\,c + 1 - P(W = 1) && (10.95) \end{aligned}$$

The result follows.

■

Applying this to (10.91), we see that the best g is given by

$$g(t) = \begin{cases} 1, & \text{if } m_{Y;X}(t) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{10.96}$$

So we find that the regression function is again optimal, in this new context.

10.5 Symmetric Relations Among Several 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.

10.5.1 Principal Components Analysis

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,

$$X_2 \approx c + dX_1 \tag{10.97}$$

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 (10.97). We saw a graph illustrating this in our chapter on multivariate distributions, page 141.

In general, consider a k-component random vector

$$X = (X_1, ..., X_k)' \tag{10.98}$$

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 (10.97) 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.

10.5.2 How to Calculate Them

Here's how it works. 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

$$Q'\Sigma Q = D \tag{10.99}$$

where D is a diagonal matrix. (This is a special case of **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

$$W_i = C_i'X,\ i = 1, ..., k \tag{10.100}$$

so that the W_i are scalar random variables, and set

$$W = (W_1, ..., W_k)' \tag{10.101}$$

Then

$$W = Q'X \tag{10.102}$$

Now, use the material on covariance matrices from our chapter on multivariate analysis, page 129,

$$Cov(W) = Cov(Q'X) = Q'Cov(X)Q = D \ \text{ (from (10.99))} \tag{10.103}$$

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.

All this will become clearer in the example below.

10.5.3 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 (also many discrete ones, but there is another tool for that case, the log-linear model, discussed in Section 10.5.4).

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:

```
> 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.

The coefficients of the linear combinations which produce W from X, i.e. the Q matrix, are available via **prc$rotation**.

10.5.4 Log-Linear Models

Here we discuss a procedure which is something of an analog of principal components for discrete variables. Our material on ANOVA will also come into play. It is recommended that the reader review Sections 10.1.14 and 10.5.1 before continuing.

10.5.4.1 The Setting

Let's consider a variation on the software engineering example in Sections 10.1.2 and 10.1.14. Assume we have the factors, IDE, Language and Education. Our change—**of extreme importance**—is that we will now assume that these factors are **RANDOM**. What does this mean?

In the original example described in Section 10.1.2, programmers were *assigned* to languages, and in our extensions of that example, we continued to assume this. Thus for example the number of programmers who use an IDE and program in Java was fixed; if we repeated the experiment, that number would stay the same. If we were sampling from some programmer population, our new sample would have new programmers, but the number using and IDE and Java would be the same as before, as our study procedure specifies this.

By contrast, let's now assume that we simply sample programmers at random, and ask them whether they prefer to use an IDE or not, and which language they prefer.[12] Then for example the number of programmers who prefer to use an IDE and program in Java will be random, not fixed; if we repeat the experiment, we will get a different count.

Suppose we now wish to investigate relations between the factors. Are choice of platform and language related to education, for instance?

10.5.4.2 The Data

Denote our three factors by $X^{(s)}$, s = 1,2,3. Here $X^{(1)}$, IDE, will take on the values 1 and 2 instead of 1 and 0 as before, 1 meaning that the programmer prefers to use an IDE, and 2 meaning not so. $X^{(3)}$ changes this way too, and $X^{(2)}$ will take on the values 1 for C++, 2 for Java and 3 for C. Note that we no longer use indicator variables.

Let $X_r^{(s)}$ denote the value of $X^{(s)}$ for the rth programmer in our sample, r = 1,2,...,n. Our data are the counts

$$N_{ijk} = \text{number of r such that } X_r^{(1)} = i, X_r^{(2)} = j \text{ and } X_r^{(3)} = k \tag{10.104}$$

For instance, if we sample 100 programmers, our data might look like this:

[12] Other sampling schemes are possible too.

```
prefers to use IDE:

                Bachelor's or less     Master's or more
          C++                   18                   15
         Java                   22                   10
            C                    6                    4


prefers not to use IDE:

                Bachelor's or less     Master's or more
          C++                    7                    4
         Java                    6                    2
            C                    3                    3
```

So for example $N_{122} = 10$ and $N_{212} = 4$.

Here we have a three-dimensional **contingency table**. Each N_{ijk} value is a **cell** in the table.

10.5.4.3 The Models

Let p_{ijk} be the population probability of a randomly-chosen programmer falling into cell ijk, i.e.

$$p_{ijk} = P\left(X^{(1)} = i \text{ and } X^{(2)} = j \text{ and } X^{(3)} = k\right) = E(N_{ijk})/n \tag{10.105}$$

As mentioned, we are interested in relations between the factors, in the form of independence, full and partial. Consider first the case of full independence:

$$\begin{aligned} p_{ijk} &= P\left(X^{(1)} = i \text{ and } X^{(2)} = j \text{ and } X^{(3)} = k\right) && (10.106) \\ &= P\left(X^{(1)} = i\right) \cdot P\left(X^{(2)} = j\right) \cdot P\left(X^{(3)} = k\right) && (10.107) \end{aligned}$$

Taking logs of both sides in (10.107), we see that independence of the three factors is equivalent to saying

$$\log(p_{ijk}) = a_i + b_j + c_k \tag{10.108}$$

for some numbers a_i, b_j and c_j. The numbers must be nonpositive, and since

$$\sum_m P(X^{(s)} = m) = 1 \tag{10.109}$$

we must have, for instance,

$$\sum_{g=1}^{2} \exp(c_g) = 1 \tag{10.110}$$

The point is that (10.108) looks like our no-interaction ANOVA models, e.g. (10.53). On the other hand, if we assume instead that Education is independent of IDE and Language but that IDE and Language are not independent of each other, our model would be

$$\begin{aligned} \log(p_{ijk}) &= P\left(X^{(1)} = i \text{ and } X^{(2)} = j\right) \cdot P\left(X^{(3)} = k\right) && (10.111) \\ &= a_i + b_j + d_{ij} + c_k && (10.112) \end{aligned}$$

Here we have written $P\left(X^{(1)} = i \text{ and } X^{(2)} = j\right)$ as a sum of "main effects" a_i and b_j, and "interaction effects," d_{ij}, analogous to ANOVA.

Another possible model would have IDE and Language conditionally independent, given Education, meaning that at any level of education, a programmer's preference to use IDE or not, and his choice of programming language, are not related. We'd write the model this way:

$$\begin{aligned} \log(p_{ijk}) &= P\left(X^{(1)} = i \text{ and } X^{(2)} = j\right) \cdot P\left(X^{(3)} = k\right) && (10.113) \\ &= a_i + b_j + f_{ik} + h_{jk} + c_k && (10.114) \end{aligned}$$

Note carefully that the type of independence in (10.114) has a quite different interpretation than that in (10.112).

The full model, with no independence assumptions at all, would have three two-way interaction terms, as well as a three-way interaction term.

10.5.4.4 Parameter Estimation

Remember, whenever we have parametric models, the statistician's "Swiss army knife" is maximum likelihood estimation. That is what is most often used in the case of log-linear models.

How, then, do we compute the likelihood of our data, the N_{ijk}? It's actually quite straightforward, because the N_{ijk} have a multinomial distribution. Then

$$L = \frac{n!}{\Pi_{i,j,k} N_{ijk}!} p_{ijk}^{N_{ijk}} \tag{10.115}$$

We then write the p_{ijk} in terms of our model parameters. Take for example (10.112), where we write

$$p_{ijk} = e^{a_i+b_j+d_{ij}+c_k} \tag{10.116}$$

We then substitute (10.116) in (10.115), and maximize the latter with respect to the a_i, b_j, d_{ij} and c_k, subject to constraints such as (10.110).

The maximization may be messy. But certain cases have been worked out in closed form, and in any case today one would typically do the computation by computer. In R, for example, there is the **loglin()** function for this purpose.

10.5.4.5 The Goal: Parsimony Again

Again, we'd like "the simplest model possible, but not simpler." This means a model with as much independence between factors as possible, subject to the model being accurate.

Classical log-linear model procedures do model selection by hypothesis testing, testing whether various interaction terms are 0. The tests often parallel ANOVA testing, with chi-square distributions arising instead of F-distributions.

10.6 Simpson's (Non-)Paradox

Suppose each individual in a population either possesses or does not possess traits *A*, *B* and *C*, and that we wish to predict trait *A*. Let $\bar{A}$, $\bar{B}$ and $\bar{C}$ denote the situations in which the individual does not possess the given trait. Simpson's Paradox then describes a situation in which

$$P(A|B) > P(A|\bar{B}) \tag{10.117}$$

and yet

$$P(A|B,C) < P(A|\bar{B},C) \tag{10.118}$$

In other words, the possession of trait B seems to have a positive predictive power for A by itself, but when in addition trait C is held constant, the relation between B and A turns negative.

An example is given by Fabris and Freitas,[13] concerning a classic study of tuberculosis mortality in 1910.

[13] C.C. Fabris and A.A. Freitas. Discovering Surprising Patterns by Detecting Occurrences of Simpson's Paradox. In *Research and Development in Intelligent Systems XVI (Proc. ES99, The 19th SGES Int. Conf. on Knowledge-Based Systems and Applied*

Here the attribute A is mortality, B is city (Richmond, with $\bar{B}$ being New York), and C is race (African-American, with $\bar{C}$ being Caucasian). In probability terms, the data show that (these of course are sample estimates)

- P(mortality | Richmond) = 0.0022
- P(mortality | New York) = 0.0019
- P(mortality | Richmond, black) = 0.0033
- P(mortality | New York, black) = 0.0056
- P(mortality | Richmond, white) = 0.0016
- P(mortality | New York, white) = 0.0018

The data also show that

- P(black | Richmond) = 0.37
- P(black | New York) = 0.002

a point which will become relevant below.

At first, New York looks like it did a better job than Richmond. However, once one accounts for race, we find that New York is actually worse than Richmond. Why the reversal? The answer stems from the fact that racial inequities being what they were at the time, blacks with the disease fared much worse than whites. Richmond's population was 37% black, proportionally far more than New York's 0.2%. So, Richmond's heavy concentration of blacks made its overall mortality rate look worse than New York's, even though things were actually much worse in New York.

But is this really a "paradox"? Closer consideration of this example reveals that the only reason this example (and others like it) is surprising is that the predictors were used in the wrong order. One normally looks for predictors one at a time, first finding the best single predictor, then the best pair of predictors, and so on. If this were done on the above data set, the first predictor variable chosen would be race, not city. In other words, the sequence of analysis would look something like this:

- P(mortality | Richmond) = 0.0022
- P(mortality | New York) = 0.0019

Artificial Intelligence), 148-160. Springer-Verlag, 1999

- P(mortality | black) = 0.0048
- P(mortality | white) = 0.0018
- P(mortality | black, Richmond) = 0.0033
- P(mortality | black, New York) = 0.0056
- P(mortality | white, Richmond) = 0.0016
- P(mortality | white, New York) = 0.0018

The analyst would have seen that race is a better predictor than city, and thus would have chosen race as the best single predictor. The analyst would then investigate the race/city predictor pair, and would never reach a point in which city alone were in the selected predictor set. Thus no anomalies would arise.

Exercises

Note to instructor: See the Preface for a list of sources of real data on which exercises can be assigned to complement the theoretical exercises below.

1. Suppose we are interested in documents of a certain type, which we'll call Type 1. Everything that is not Type 1 we'll call Type 2, with a proportion q of all documents being Type 1. Our goal will be to try to guess document type by the presence of absence of a certain word; we will guess Type 1 if the word is present, and otherwise will guess Type 2.

Let T denote document type, and let W denote the event that the word is in the document. Also, let p_i be the proportion of documents that contain the word, among all documents of Type i, i = 1,2. The event C will denote our guessing correctly.

Find the overall probability of correct classification, $P(C)$, and also $P(C|W)$.

Hint: Be careful of your conditional and unconditional probabilities here.

2. In the quartic model in ALOHA simulation example, find an approximate 95% confidence interval for the true population mean wait if our backoff parameter b is set to 0.6.

Hint: You will need to use the fact that a linear combination of the components of a multivariate normal random vector has a univariate normal distributions as discussed in Section 5.8.2.1.

3. Consider the linear regression model with one predictor, i.e. r = 1. Let Y_i and X_i represent the values of the response and predictor variables for the ith observation in our sample.

(a) Assume as in Section 10.1.9.4 that $Var(Y|X = t)$ is a constant in t, σ^2. Find the exact value of $Cov(\hat{\beta}_0, \hat{\beta}_1)$, as a function of the X_i and σ^2. Your final answer should be in scalar, i.e. non-matrix form.

(b) Suppose we wish to fit the model $m_{Y;X}(t) = \beta_1 t$, i.e. the usual linear model but without the constant term, β_0. Derive a formula for the least-squares estimate of β_1.

4. Suppose the random pair (X, Y) has density $8st$ on $0 < t < s < 1$. Find $m_{Y;X}(s)$ and $Var(Y|X = t),\ 0 < s < 1$.

5. We showed that (10.75) reduces to the logistic model in the case in which the distribution of X given Y is normal. Show that this is also true in the case in which that distribution is exponential, i.e.

$$f_{X|Y}(t, i) = \lambda_i e^{-\lambda_i t},\ t > 0 \tag{10.119}$$

6. The code below reads in a file, **data.txt**, with the header record

```
"age", "weight", "systolic blood pressure", "height"
```

and then does the regression analysis.

Suppose we wish to estimate β in the model

$$\text{mean weight} = \beta_0 + \beta_1 \text{height} + \beta_2 \text{age}$$

Fill in the blanks in the code:

```
dt <- _____________(_____________________________________)
regr <- lm(_____________________________________________________)
cvmat <- _________________(regr)
print("the estimated value of beta2-beta0 is",
   ____________________________________________________)
print("the estimated variance of beta2 - beta0 is",
   ____________________________ %*% cvmat %*%  ______________________________)
# calculate the matrix Q
q <- cbind(_____________________________________________________)
```

7. In this problem, you will conduct an R simulation experiment similar to that of Foster and Stine on overfitting, discussed in Section 10.1.10.2.

Generate data $X_i^{(j)},\ i = 1, ..., n,\ j = 1, ..., r$ from a N(0,1) distribution, and $\epsilon_i,\ i = 1, ..., n$ from N(0,4). Set $Y_i = X_i^{(1)} + \epsilon_i,\ i = 1, ..., n$. This simulates drawing a random sample of n observations from an (r+1)-variate population.

Now suppose the analyst, unaware that Y is related to only $X^{(1)}$, fits the model

$$m_{Y;X^{(1)},...,X^{(r)}}(t_1, ..., t_r) = \beta_0 + \beta_1 t^{(1)} + ... + \beta_r t^{(r)} \tag{10.120}$$

In actuality, $\beta_j = 0$ for $j > 1$ (and for $i = 0$). But the analyst wouldn't know this. Suppose the analyst selects predictors by testing the hypotheses $H_0 : \beta_i = 0$, as in Section 10.1.10.2, with $\alpha = 0.05$.

Do this for various values of r and n. You should find that, for fixed n and increasing r. You begin to find that some of the predictors are declared to be "significantly" related to Y (complete with asterisks) when in fact they are not (while $X^{(1)}$, which really is related to Y, may be declared NOT "significant." This illustrates the folly of using hypothesis testing to do variable selection.

8. Suppose given X = t, the distribution of Y has mean γt and variance σ^2, for all t in (0,1). This is a fixed-X regression setting, i.e. X is nonrandom: For each i = 1,...,n we observe Yi drawn at random from the distribution of Y given X = i/n. The quantities γ and σ^2 are unknown.

Our goal is to estimate $m_{Y;X}(0.75)$. We have two choices for our estimator:

- We can estimate in the usual least-squares manner, denoting our estimate by G, and then use as our estimator $T_1 = 0.75G$.
- We can take our estimator T_2 to be $(Y_1 + ... + Y_n)/n$,

Perform a tradeoff analysis similar to that of Section 8.2, determining under what conditions T_1 is superior to T_2 and vice versa. Our criterion is mean squared error (MSE), $E[(T_i - m_{Y;X}(0.75)^]$. Make your expressions as closed-form as possible.

Advice: This is a linear model, albeit one without an intercept term. The quantity G here is simply $\hat{\sigma}$. G will turn out to be a linear combination of the Xs (which are constants), so its variance is easy to find.

9. Suppose X has an $N(\mu, \mu^2)$ distribution, i.e. with the standard deviation equal to the mean. (A common assumption in regression contexts.) Show that $h(X) = \ln(X)$ will be a variance-stabilizing transformation, a concept discussed in Section 8.6.2.

10. Consider a random pair (X, Y) for which the linear model $E(Y|X) = \beta_0 + \beta_1 X$ holds, and think about predicting Y, first without X and then with X, minimizing mean squared prediction error (MSPE) in each case. From Section 10.4.1, we know that without X, the best predictor is EY, while with X it is $E(Y|X)$, which under our assumption here is $\beta_0 + \beta_1 X$. Show that the reduction in MSPE accrued by using X, i.e.

$$\frac{E\left[(Y - EY)^2\right] - E\left[\{Y - E(Y|X)\}^2\right]}{E\left[(Y - EY)^2\right]} \tag{10.121}$$

is equal to $\rho^2(X, Y)$.

11. In an analysis published on the Web (Sparks *et al*, Disease Progress over Time, *The Plant Health Instructor*, 2008, the following R output is presented:

```
> severity.lm <- lm(diseasesev~temperature,data=severity)
> summary(severity.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.66233    1.10082   2.418  0.04195 *
temperature  0.24168    0.06346   3.808  0.00518 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Fill in the blanks:

(a) The model here is

mean _____ = β_0 + β_1 _____

(b) The two null hypotheses being tested here are H_0 : _____ and H_0 : _____.

12. In the notation of this chapter, give matrix and/or vector expressions for each of the following in the linear regression model:

(a) s^2, our estimator of σ^2

(b) the standard error of the estimated value of the regression function $m_{Y;X}(t)$ at $t = c$, where $c = (c_0, c_1, ..., c_r)$

Chapter 11

Markov Chains

One of the most famous stochastic models is that of a **Markov chain**. This type of model is widely used in computer science, biology, physics, business and so on.

11.1 Discrete-Time Markov Chains

11.1.1 Example: Finite Random Walk

To motivate this discussion, let us start with a simple example: Consider a **random walk** on the set of integers between 1 and 5, moving randomly through that set, say one move per second, according to the following scheme. If we are currently at position i, then one time period later we will be at either i-1, i or i+1, according to the outcome of rolling a fair die—we move to i-1 if the die comes up 1 or 2, stay at i if the die comes up 3 or 4, and move to i+1 in the case of a 5 or 6. For the special cases i = 1 and i = 5, we simply move back to 2 or 4, respectively. (In random walk terminology, these are called **reflecting barriers**.)

The integers 1 through 5 form the **state space** for this process; if we are currently at 4, for instance, we say we are in state 4. Let X_t represent the position of the particle at time t, t = 0, 1,2,....

The random walk is a **Markov process**. The process is "memoryless," meaning that we can "forget the past"; given the present and the past, the future depends only on the present:

$$P(X_{t+1} = s_{t+1}|X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) = P(X_{t+1} = s_{t+1}|X_t = s_t) \tag{11.1}$$

The term *Markov process* is the general one. If the state space is discrete, i.e. finite or countably infinite, then we usually use the more specialized term, *Markov chain*.

Although this equation has a very complex look, it has a very simple meaning: The distribution of our next position, given our current position and all our past positions, is dependent only on the current position. In other words, the system is "memoryless," somewhat in analogy to the properties of the exponential distribution discussed in Section 6.1. (In fact exponential distributions will play a key role when we get to continuous-time Markov chains in Section 11.4. It is clear that the random walk process above does have this memoryless property; for instance, if we are now at position 4, the probability that our next state will be 3 is 1/3—no matter where we were in the past.

Continuing this example, let p_{ij} denote the probability of going from position i to position j in one step. For example, $p_{21} = p_{23} = \frac{1}{3}$ while $p_{24} = 0$ (we can reach position 4 from position 2 in two steps, but not in one step). The numbers p_{ij} are called the **one-step transition probabilities** of the process. Denote by P the matrix whose entries are the p_{ij}:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{11.2}$$

By the way, it turns out that the matrix P^k gives the k-step transition probabilities. In other words, the element (i,j) of this matrix gives the probability of going from i to j in k steps.

11.1.2 Long-Run Distribution

In typical applications we are interested in the long-run distribution of the process, for example the long-run proportion of the time that we are at position 4. For each state i, define

$$\pi_i = \lim_{t \to \infty} \frac{N_{it}}{t} \tag{11.3}$$

where N_{it} is the number of visits the process makes to state i among times 1, 2,..., t. In most practical cases, this proportion will exist and be independent of our initial position X_0. The π_i are called the **steady-state probabilities**, or the **stationary distribution** of the Markov chain.

Intuitively, the existence of π_i implies that as t approaches infinity, the system approaches steady-state, in the sense that

$$\lim_{t \to \infty} P(X_t = i) = \pi_i \tag{11.4}$$

Actually, the limit (11.4) may not exist in some cases. We'll return to that point later, but for typical cases it does exist, and we will usually assume this.

11.1.2.1 Derivation of the Balance Equations

Equation (11.4) suggests a way to calculate the values π_i, as follows.

First note that

$$P(X_{t+1} = i) = \sum_k P(X_t = k \text{ and } X_{t+1} = i) = \sum_k P(X_t = k) P(X_{t+1} = i | X_t = k) = \sum_k P(X_t = k) p_{ki} \tag{11.5}$$

where the sum goes over all states k. For example, in our random walk example above, we would have

$$P(X_{t+1} = 3) = \sum_{k=1}^{5} P(X_t = k \text{ and } X_{t+1} = 3) = \sum_{k=1}^{5} P(X_t = k) P(X_{t+1} = 3 | X_t = k) = \sum_{k=1}^{5} P(X_t = k) p_{k3} \tag{11.6}$$

Then as $t \to \infty$ in Equation (11.5), intuitively we would have

$$\pi_i = \sum_k \pi_k p_{ki} \tag{11.7}$$

Remember, here we know the p_{ki} and want to find the π_i. Solving these **balance equations** equations (one for each i), gives us the π_i.

For the random walk problem above, for instance, the solution is $\pi = (\frac{1}{11}, \frac{3}{11}, \frac{3}{11}, \frac{3}{11}, \frac{1}{11})$. Thus in the long run we will spend 1/11 of our time at position 1, 3/11 of our time at position 2, and so on.

11.1.2.2 Solving the Balance Equations

A matrix formulation is also useful. Letting π denote the row vector of the elements π_i, i.e. $\pi = (\pi_1, \pi_2, ...)$, these equations (one for each i) then have the matrix form

$$\pi = \pi P \tag{11.8}$$

or

$$(I - P')\pi = 0 \tag{11.9}$$

where as usual ' denotes matrix transpose.

Note that there is also the constraint

$$\sum_i \pi_i = 1 \tag{11.10}$$

One of the equations in the system is redundant. We thus eliminate one of them, say by removing the last row of I-P in (11.9). This can be used to calculate the π_i.

To reflect (11.10), which in matrix form is

$$1_n'\pi = 1 \tag{11.11}$$

where 1_n is a column vector of all 1s, n is the number of states, and we replace the removed row in I-P by a row of all 1s, and in the right-hand side of (11.9) we replace the last 0 by a 1. We can then solve the system.

All this can be done with R's **solve()** function:

```
1  findpi1 <- function(p) {
2     n <- nrow(p)
3     imp <- diag(n) - t(p)  # I-P
4     imp[n,] <- rep(1,n)
5     rhs <- c(rep(0,n-1),1)
6     pivec <- solve(imp,rhs)
7     return(pivec)
8  }
```

Or one can note from (11.8) that π is a left eigenvector of P with eigenvalue 1, so one can use R's **eigen()** function. It can be proven that if P is irreducible and aperiodic (defined later in this chapter), every eigenvalue other than 1 is smaller than 1 (so we can speak of *the* eigenvalue 1), and the eigenvector corresponding to 1 has all components real.

Since π is a left eigenvector, the argument in the call must be P' rather than P. In addition, since an eigenvector is only unique up to scalar multiplication, we must deal with the fact that the return value of **eigen()** may have negative components, and will likely not satisfy (11.10). Here is the code:

```
1  findpi2 <- function(p) {
2     n <- nrow(p)
```

```
3     # find first eigenvector of P'
4     pivec <- eigen(t(p))$vectors[,1]
5     # guaranteed to be real, but could be negative
6     if (pivec[1] < 0) pivec <- -pivec
7     # normalize
8     pivec <- pivec / sum(pivec)
9     return(pivec)
10 }
```

But Equation (11.9) may not be easy to solve. For instance, if the state space is infinite, then this matrix equation represents infinitely many scalar equations. In such cases, you may need to try to find some clever trick which will allow you to solve the system, or in many cases a clever trick to analyze the process in some way other than explicit solution of the system of equations.

And even for finite state spaces, the matrix may be extremely large. In some cases, you may need to resort to numerical methods.

11.1.2.3 Periodic Chains

Note again that even if Equation (11.9) has a solution, this does not imply that (11.4) holds. For instance, suppose we alter the random walk example above so that

$$p_{i,i-1} = p_{i,i+1} = \frac{1}{2} \tag{11.12}$$

for i = 2, 3, 4, with transitions out of states 1 and 5 remaining as before. In this case, the solution to Equation (11.9) is $(\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8})$. This solution is still valid, in the sense that Equation (11.3) will hold. For example, we will spend 1/4 of our time at Position 4 in the long run. But the limit of $P(X_i = 4)$ will not be 1/4, and in fact the limit will not even exist. If say X_0 is even, then X_i can be even only for even values of i. We say that this Markov chain is **periodic** with period 2, meaning that returns to a given state can only occur after amounts of time which are multiples of 2.

11.1.2.4 The Meaning of the Term "Stationary Distribution"

Though we have informally defined the term *stationary distribution* in terms of long-run proportions, the technical definition is this:

Definition 37 *Consider a Markov chain. Suppose we have a vector π of nonnegative numbers that sum to 1. Let X_0 have the distribution π. If that results in X_1 having that distribution too (and thus also all X_n), we say that π is the* **stationary distribution** *of this Markov chain.*

Note that this definition stems from (11.5).

For instance, in our (first) random walk example above, this would mean that if we have X_0 distributed on the integers 1 through 5 with probabilities $(\frac{1}{11}, \frac{3}{11}, \frac{3}{11}, \frac{3}{11}, \frac{1}{11})$, then for example $P(X_1 = 1) = \frac{1}{11}$, $P(X_1 = 4) = \frac{3}{11}$ etc. This is indeed the case, as you can verify using (11.5) with t = 0.

In our "notebook" view, here is what we would do. Imagine that we generate a random integer between 1 and 5 according to the probabilities $(\frac{1}{11}, \frac{3}{11}, \frac{3}{11}, \frac{3}{11}, \frac{1}{11})$,[1] and set X_0 to that number. We would then generate another random number, by rolling an ordinary die, and going left, right or staying put, with probability 1/3 each. We would then write down X_1 and X_2 on the first line of our notebook. We would then do this experiment again, recording the results on the second line, then again and again. In the long run, 3/11 of the lines would have, for instance, $X_0 = 4$, and 3/11 of the lines would have $X_1 = 4$. In other words, X_1 would have the same distribution as X_0.

11.1.3 Example: Stuck-At 0 Fault

11.1.3.1 Description

In the above example, the labels for the states consisted of single integers i. In some other examples, convenient labels may be r-tuples, for example 2-tuples (i,j).

Consider a serial communication line. Let $B_1, B_2, B_3, ...$ denote the sequence of bits transmitted on this line. It is reasonable to assume the B_i to be independent, and that $P(B_i = 0)$ and $P(B_i = 1)$ are both equal to 0.5.

Suppose that the receiver will eventually fail, with the type of failure being **stuck at 0**, meaning that after failure it will report all future received bits to be 0, regardless of their true value. Once failed, the receiver stays failed, and should be replaced. Eventually the new receiver will also fail, and we will replace it; we continue this process indefinitely.

Let ρ denote the probability that the receiver fails on any given bit, with independence between bits in terms of receiver failure. Then the lifetime of the receiver, that is, the time to failure, is geometrically distributed with "success" probability ρ i.e. the probability of failing on receipt of the i-th bit after the receiver is installed is $(1 - \rho)^{i-1}\rho$ for i = l,2,3,...

However, the problem is that we will not know whether a receiver has failed (unless we test it once in a while, which we are not including in this example). If the receiver reports a long string of 0s, we should suspect that the receiver has failed, but of course we cannot be sure that it has; it is still possible that the message being transmitted just happened to contain a long string of 0s.

Suppose we adopt the policy that, if we receive k consecutive 0s, we will replace the receiver with a new

[1] Say by rolling an 11-sided die.

unit. Here k is a design parameter; what value should we choose for it? If we use a very small value, then we will incur great expense, due to the fact that we will be replacing receiver units at an unnecessarily high rate. On the other hand, if we make k too large, then we will often wait too long to replace the receiver, and the resulting error rate in received bits will be sizable. Resolution of this tradeoff between expense and accuracy depends on the relative importance of the two. (There are also other possibilities, involving the addition of redundant bits for error detection, such as parity bits. For simplicity, we will not consider such refinements here. However, the analysis of more complex systems would be similar to the one below.)

11.1.3.2 Initial Analysis

A natural state space in this example would be

$$\{(i,j) : i = 0, 1, ..., k-1; j = 0, 1; i + j \neq 0\} \tag{11.13}$$

where i represents the number of consecutive 0s that we have received so far, and j represents the state of the receiver (0 for failed, 1 for nonfailed). Note that when we are in a state of the form (k-1,j), if we receive a 0 on the next bit (whether it is a true 0 or the receiver has failed), our new state will be (0,1), as we will install a new receiver. Note too that there is no state (0,0), since if the receiver is down it must have received at least one bit.

The calculation of the transition matrix P is straightforward, though it requires careful thought. For example, suppose the current state is (2,1), and that we are investigating the expense and bit accuracy corresponding to a policy having k = 5. What can happen upon receipt of the next bit? The next bit will have a true value of either 0 or 1, with probability 0.5 each. The receiver will change from working to failed status with probability ρ. Thus our next state could be:

- (3,1), if a 0 arrives, and the receiver does not fail;
- (0,1), if a 1 arrives, and the receiver does not fail; or
- (3,0), if the receiver fails

The probabilities of these three transitions out of state (2,1) are:

$$p_{(2,1),(3,1)} = 0.5(1-\rho) \tag{11.14}$$

$$p_{(2,1),(0,1)} = 0.5(1-\rho) \tag{11.15}$$

$$p_{(2,1),(3,0)} = \rho \tag{11.16}$$

Other entries of the matrix P can be computed similarly. Note by the way that from state (4,1) we will go to (0,1), no matter what happens.

Formally specifying the matrix P using the 2-tuple notation as above would be very cumbersome. In this case, it would be much easier to map to a one-dimensional labeling. For example, if k = 5, the nine states (1,0),...,(4,0),(0,1),(1,1),...,(4,1) could be renamed states 1,2,...,9. Then we could form P under this labeling, and the transition probabilities above would appear as

$$p_{78} = 0.5(1-\rho) \tag{11.17}$$
$$p_{75} = 0.5(1-\rho) \tag{11.18}$$
$$p_{73} = \rho \tag{11.19}$$

11.1.3.3 Going Beyond Finding π

Finding the π_i should be just the first step. We then want to use them to calculate various quantities of interest.[2] For instance, in this example, it would also be useful to find the error rate ϵ, and the mean time (i.e., the mean number of bit receptions) between receiver replacements, μ. We can find both ϵ and μ in terms of the π_i, in the following manner.

The quantity ϵ is the proportion of the time during which the true value of the received bit is 1 but the receiver is down, which is 0.5 times the proportion of the time spent in states of the form (i,0):

$$\epsilon = 0.5(\pi_1 + \pi_2 + \pi_3 + \pi_4) \tag{11.20}$$

This should be clear intuitively, but it would also be instructive to present a more formal derivation of the same thing. Let E_n be the event that the n-th bit is received in error, with D_n denoting the event that the receiver is down. Then

$$\epsilon = \lim_{n\to\infty} P(E_n) \tag{11.21}$$
$$= \lim_{n\to\infty} P(B_n = 1 \text{ and } D_n) \tag{11.22}$$
$$= \lim_{n\to\infty} P(B_n = 1)P(D_n) \tag{11.23}$$
$$= 0.5(\pi_1 + \pi_2 + \pi_3 + \pi_4) \tag{11.24}$$

Here we used the fact that B_n and the receiver state are independent.

[2] Note that unlike a classroom setting, where those quantities would be listed for the students to calculate, in research we must decide on our own which quantities are of interest.

Note that with the interpretation of π as the stationary distribution of the process, in Equations (11.21) above, we do not even need to take limits.

Equations (11.21) follow a pattern we'll use repeatedly in this chapter. In subsequent examples we will not show the steps with the limits, but the limits are indeed there. Make sure to mentally go through these steps yourself.[3]

Now to get μ in terms of the π_i note that since μ is the long-run average number of bits between receiver replacements, it is then the reciprocal of η, the long-run fraction of bits that result in replacements. For example, say we replace the receiver on average every 20 bits. Over a period of 1000 bits, then (speaking on an intuitive level) that would mean about 50 replacements. Thus approximately 0.05 (50 out of 1000) of all bits results in replacements.

$$\mu = \frac{1}{\eta} \tag{11.25}$$

Again suppose k = 5. A replacement will occur only from states of the form (4,j), and even then only under the condition that the next reported bit is a 0. In other words, there are three possible ways in which replacement can occur:

(a) We are in state (4,0). Here, since the receiver has failed, the next reported bit will definitely be a 0, regardless of that bit's true value. We will then have a total of k = 5 consecutive received 0s, and therefore will replace the receiver.

(b) We are in the state (4,1), and the next bit to arrive is a true 0. It then will be reported as a 0, our fifth consecutive 0, and we will replace the receiver, as in (a).

(c) We are in the state (4,1), and the next bit to arrive is a true 1, but the receiver fails at that time, resulting in the reported value being a 0. Again we have five consecutive reported 0s, so we replace the receiver.

Therefore,

$$\eta = \pi_4 + \pi_9(0.5 + 0.5\rho) \tag{11.26}$$

Again, make sure you work through the full version of (11.26), using the pattern in (11.21).

Thus

$$\mu = \frac{1}{\eta} = \frac{1}{\pi_4 + 0.5\pi_9(1+\rho)} \tag{11.27}$$

[3]The other way to work this out rigorously is to assume that X_0 has the distribution π, as in Section 11.1.2.4. Then no limits are needed in (11.21. But this may be more difficult to understand.

This kind of analysis could be used as the core of a cost-benefit tradeoff investigation to determine a good value of k. (Note that the π_i are functions of k, and that the above equations for the case k = 5 must be modified for other values of k.)

11.1.4 Example: Shared-Memory Multiprocessor

(Adapted from *Probabiility and Statistics, with Reliability, Queuing and Computer Science Applicatiions*, by K.S. Trivedi, Prentice-Hall, 1982 and 2002, but similar to many models in the research literature.)

11.1.4.1 The Model

Consider a shared-memory multiprocessor system with m memory modules and m CPUs. The address space is partitioned into m chunks, based on either the most-significant or least-significant $\log_2 m$ bits in the address.[4]

The CPUs will need to access the memory modules in some random way, depending on the programs they are running. To make this idea concrete, consider the Intel assembly language instruction

```
add %eax, (%ebx)
```

which adds the contents of the EAX register to the word in memory pointed to by the EBX register. Execution of that instruction will (absent cache and other similar effects, as we will assume here and below) involve two accesses to memory—one to fetch the old value of the word pointed to by EBX, and another to store the new value. Moreover, the instruction itself must be fetched from memory. So, altogether the processing of this instruction involves three memory accesses.

Since different programs are made up of different instructions, use different register values and so on, the sequence of addresses in memory that are generated by CPUs are modeled as random variables. In our model here, the CPUs are assumed to act independently of each other, and successive requests from a given CPU are independent of each other too. A CPU will choose the ith module with probability q_i. A memory request takes one unit of time to process, though the wait may be longer due to queuing. In this very simplistic model, as soon as a CPU's memory request is fulfilled, it generates another one. On the other hand, while a CPU has one memory request pending, it does not generate another.

Let's assume a crossbar interconnect, which means there are m^2 separate paths from CPUs to memory modules, so that if the m CPUs have memory requests to m different memory modules, then all the requests can be fulfilled simultaneously. Also, assume as an approximation that we can ignore communication delays.

[4] You may recognize this as high-order and low-order interleaving, respectively.

How good are these assumptions? One weakness, for instance, is that many instructions, for example, do not use memory at all, except for the instruction fetch, and as mentioned, even the latter may be suppressed due to cache effects.

Another example of potential problems with the assumptions involves the fact that many programs will have code like

```
for (i = 0; i < 10000; i++) sum += x[i];
```

Since the elements of the array x will be stored in consecutive addresses, successive memory requests from the CPU while executing this code will not be independent. The assumption would be more justified if we were including cache effects, or (noticed by Earl Barr) if we are studying a timesharing system with a small quantum size.

Thus, many models of systems like this have been quite complex, in order to capture the effects of various things like caching, nonindependence and so on in the model. Nevertheless, one can often get some insight from even very simple models too. In any case, for our purposes here it is best to stick to simple models, so as to understand more easily.

Our state will be an m-tuple $(N_1, ..., N_m)$, where N_i is the number of requests currently pending at memory module i. Recalling our assumption that a CPU generates another memory request immediately after the previous one is fulfilled, we always have that $N_1 + ... + N_m = m$.

It is straightforward to find the transition probabilities p_{ij}. Here are a couple of examples, with m = 2:

- $p_{(2,0),(1,1)}$: Recall that state (2,0) means that currently there are two requests pending at Module 1, one being served and one in the queue, and no requests at Module 2. For the transition $(2,0) \to (1,1)$ to occur, when the request being served at Module 1 is done, it will make a new request, this time for Module 2. This will occur with probability q_2. Meanwhile, the request which had been queued at Module 1 will now start service. So, $p_{(2,0),(1,1)} = q_2$.

- $p_{(1,1),(1,1)}$: In state (1,1), both pending requests will finish in this cycle. To go to (1,1) again, that would mean that the two CPUs request different modules from each other—CPUs 1 and 2 choose Modules 1 and 2 or 2 and 1. Each of those two possibilities has probability $q_1 q_2$, so $p_{(1,1),(1,1)} = 2q_1 q_2$.

We then solve for the π, using (11.7). It turns out, for example, that

$$\pi_{(1,1)} = \frac{q_1 q_2}{1 - 2q_1 q_2} \tag{11.28}$$

11.1.4.2 Going Beyond Finding π

Let B denote the number of memory requests completed in a given memory cycle. Then we may be interested in E(B), the number of requests completed per unit time, i.e. per cycle. We can find E(B) as follows. Let S denote the current state. Then, continuing the case m = 2, we have from the Law of Total Expectation,[5]

$$\begin{aligned} E(B) &= E[E(B|S)] && (11.29)\\ &= P(S=(2,0))E(B|S=(2,0)) + P(S=(1,1))E(B|S=(1,1)) + P(S=(0,2))E(B|S=(0,2)) && (11.30)\\ &= \pi_{(2,0)}E(B|S=(2,0)) + \pi_{(1,1)}E(B|S=(1,1)) + \pi_{(0,2)}E(B|S=(0,2)) && (11.31) \end{aligned}$$

All this equation is doing is finding the overall mean of B by breaking down into the cases for the different states.

Now if we are in state (2,0), only one request will be completed this cycle, and B will be 1. Thus $E(B|S=(2,0))=1$. Similarly, $E(B|S=(1,1))=2$ and so on. After doing all the algebra, we find that

$$EB = \frac{1-q_1q_2}{1-2q_1q_2} \tag{11.32}$$

The maximum value of E(B) occurs when $q_1 = q_2 = \frac{1}{2}$, in which case E(B)=1.5. This is a lot less than the maximum capacity of the memory system, which is m = 2 requests per cycle.

So, we can learn a lot even from this simple model, in this case learning that there may be a substantial underutilization of the system. This is a common theme in probabilistic modeling: Simple models may be worthwhile in terms of insight provided, even if their numerical predictions may not be too accurate.

11.1.5 Example: Slotted ALOHA

Recall the slotted ALOHA model from Chapter 2:

- Time is divided into slots or epochs.
- There are n nodes, each of which is either idle or has a **single** message transmission pending. So, a node doesn't generate a new message until the old one is successfully transmitted (a very unrealistic assumption, but we're keeping things simple here).
- In the middle of each time slot, each of the idle nodes generates a message with probability q.

[5] Actually, we could take a more direct route in this case, noting that B can only take on the values 1 and 2. Then $EB = P(B=1) + 2P(B=2) = \pi_{(2,0} + \pi_{s(0,2)} + 2\pi_{(1,1)}$. But the analysis below extends better to the case of general m.

- Just before the end of each time slot, each active node attempts to send its message with probability p.
- If more than one node attempts to send within a given time slot, there is a **collision**, and each of the transmissions involved will fail.
- So, we include a **backoff** mechanism: At the middle of each time slot, each node with a message will with probability q attempt to send the message, with the transmission time occupying the remainder of the slot.

So, q is a design parameter, which must be chosen carefully. If q is too large, we will have too mnay collisions, thus increasing the average time to send a message. If q is too small, a node will often refrain from sending even if no other node is there to collide with.

Define our state for any given time slot to be the number of nodes currently having a message to send at the very beginning of the time slot (before new messages are generated). Then for $0 < i < n$ and $0 < j < n - i$ (there will be a few special boundary cases to consider too), we have

$$p_{i,i-1} = \underbrace{(1-q)^{n-i}}_{\text{no new msgs}} \cdot \underbrace{i(1-p)^{i-1}p}_{\text{one xmit}} \tag{11.33}$$

$$p_{ii} = \underbrace{(1-q)^{n-i} \cdot [1 - i(1-p)^{i-1}p]}_{\text{no new msgs and no succ xmits}} + \underbrace{(n-i)(1-q)^{n-i-1}q \cdot (i+1)(1-p)^i p}_{\text{one new msg and one xmit}} \tag{11.34}$$

$$\begin{aligned} p_{i,i+j} &= \underbrace{\binom{n-i}{j} q^j (1-q)^{n-i-j} \cdot [1 - (i+j)(1-p)^{i+j-1}p]}_{\text{j new msgs and no succ xmit}} \\ &+ \underbrace{\binom{n-i}{j+1} q^{j+1} (1-q)^{n-i-j-1} \cdot (i+j+1)(1-p)^{i+j}p}_{\text{j+1 new msgs and succ xmit}} \end{aligned} \tag{11.35}$$

Note that in (11.34) and (11.35), we must take into account the fact that a node with a newly-created messages might try to send it. In (11.35), for instance, in the first term we have j new messages, on top of the i we already had, so i+j messages might try to send. The probability that there is no successful transmission is then $1 - (i+j)(1-p)^{i+j-1}p$.

The matrix P is then quite complex. We always hope to find a closed-form solution, but that is unlikely in this case. Solving it on a computer is easy, though, say by using the **solve()** function in the R statistical language.

11.1.5.1 Going Beyond Finding π

Once again various interesting quantities can be derived as functions of the π, such as the system throughput τ, i.e. the number of successful transmissions in the network per unit time. Here's how to get τ:

First, suppose for concreteness that in steady-state the probability of there being a successful transmission in a given slot is 20%. Then after, say, 100,000 slots, about 20,000 will have successful transmissions—a throughput of 0.2. So, the long-run probability of successful transmission is the same as the long-run fraction of slots in which there are successful transmissions! That in turn can be broken down in terms of the various states:

$$\begin{aligned} \tau &= P(\text{success xmit}) \\ &= \sum_s P(\text{success xmit} \mid \text{in state s}) P(\text{in state s}) \end{aligned} \tag{11.36}$$

Now, to calculate $P(\text{success xmit} \mid \text{in state s})$, recall that in state s we start the slot with s nonidle nodes, but that we may acquire some new ones; each of the n-s idle nodes will create a new message, with probability q. So,

$$P(\text{success xmit} \mid \text{in state s}) = \sum_{j=0}^{n-s} \binom{n-s}{j} q^j (1-q)^{n-s-j} \cdot (s+j)(1-p)^{s+j-1} p \tag{11.37}$$

Substituting into (11.36), we have

$$\tau = \sum_{s=0}^{n} \sum_{j=0}^{n-s} \binom{n-s}{j} q^j (1-q)^{n-s-j} \cdot (s+j)(1-p)^{s+j-1} p \cdot \pi_s \tag{11.38}$$

With some more subtle reasoning, one can derive the mean time a message waits before being successfully transmitted, as follows:

Focus attention on one particular node, say Node 0. It will repeatedly cycle through idle and busy periods, I and B. We wish to find E(B). I has a geometric distribution with parameter q,[6] so

$$E(I) = \frac{1}{q} \tag{11.39}$$

[6] If a message is sent in the same slot in which it is created, we will count B as 1. If it is sent in the following slot, B = 2, etc. B will have a modified geometric distribution starting at 0 instead of 1, but we will ignore this here for the sake of simplicity.

Then if we can find E(I+B), we will get E(B) by subtraction.

To find E(I+B), note that there is a one-to-one correspondence between I+B cycles and successful transmissions; each I+B period ends with a successful transmission at Node 0. Imagine again observing this node for, say, 100000 time slots, and say E(I+B) is 2000. That would mean we'd have about 50 cycles, thus 50 successful transmissions from this node. In other words, the throughput would be approximately 50/100000 = 0.02 = 1/E(I+B). So, a fraction

$$\frac{1}{E(I+B)} \tag{11.40}$$

of the time slots have successful transmissions from this node.

But that quantity is the throughput for this node (number of successful transmissions per unit time), and due to the symmetry of the system, that throughput is 1/n of the total throughput of the n nodes in the network, which we denoted above by τ.

So,

$$E(I+B) = \frac{n}{\tau} \tag{11.41}$$

Thus from (11.39) we have

$$E(B) = \frac{n}{\tau} - \frac{1}{q} \tag{11.42}$$

where of course τ is the function of the π_i in (11.36).

Now let's find the proportion of attempted transmissions which are successful. This will be

$$\frac{E(\text{number of successful transmissions in a slot})}{E(\text{number of attempted transmissions in a slot})} \tag{11.43}$$

(To see why this is the case, again think of watching the network for 100,000 slots.) Then the proportion of successful transmissions during that period of time is the number of successful transmissions divided by the number of attempted transmissions. Those two numbers are approximately the numerator and denominator of 11.43.

Now, how do we evaluate (11.43)? Well, the numerator is easy, since it is τ, which we found before. The denominator will be

$$\sum_s \pi_s [sp + (n-s)pq] \tag{11.44}$$

The factor sp+spq comes from the following reasoning. If we are in state s, the s nodes which already have something to send will each transmit with probability p, so there will be an expected number sp of them that try to send. Also, of the n-s which are idle at the beginning of the slot, an expected sq of them will generate new messages, and of those sq, and estimated sqp will try to send.

11.2 Simulation of Markov Chains

Simulation of Markov chains is identical to the patterns we've seen in earlier chapters, except for one somewhat subtle difference. To see this, consider the first simulation code presented in this book, in Section 2.12.1.

There we were simulating X_1 and X_2, the state of the system during the first two time slots. A rough outline of the code is

```
do nreps times
   simulate X1 and X2
   record X1, X1 and update counts
calculate probabilities as counts/nreps
```

We "played the movie" **nreps** times, calculating the behavior of X_1 and X_2 over many plays.

But suppose instead that we had been interested in finding

$$\lim_{n \to \infty} \frac{X_1 + ... + X_n}{n} \tag{11.45}$$

i.e. the long-run average number of active nodes over infinitely many time slots. **In that case, we would need to play the movie only once.**

Here's an example, simulating the stuck-at 0 example from Section 11.1.3:

```
# simulates the stuck-at 0 fault example, finding mean time between
# replacements; we'll keep simulating until we have nreplace replacements
# of the receiver, then divide that into the number of bits received, to
# get the mean time between replacements
sasim <- function(nreplace,rho,k) {
   replace <- 0  # number of receivers replaced so far
   up <- TRUE  # receiver is up
   nbits <- 0  # number of bits received so far
   ncsec0 <- 0  # current number of consecutive 0s
   while (TRUE) {
      bit <- sample(0:1,1)
      nbits <- nbits + 1
      if (runif(1) < rho) {
         up <- FALSE
```

```
15            bit <- 0
16         }
17         if (bit == 0) {
18            ncsec0 <- ncsec0 + 1
19            if (ncsec0 == k) {
20               replace <- replace + 1
21               ncsec0 <- 0
22               up <- TRUE
23            }
24         }
25         if (replace == nreplace) break
26      }
27      return(nbits/nreplace)
28   }
```

This follows from the fact that the limit in (11.3) occurs even in "one play."

11.3 Hidden Markov Models

The word *hidden* in the term *Hidden Markov Model* (HMM) refers to the fact that the state of the process is hidden, i.e. unobservable.

Actually, we've already seen an example of this, back in Section 11.1.3. There the state, actually just part of it, was unobservable, namely the status of the receiver being up or down. But here we are not trying to guess X_n from Y_n (see below), so it probably would not be considered an HMM.

Note too the connection to mixture models, Section 5.10.

An HMM consists of a Markov chain X_n which is unobservable, together with observable values Y_n. The X_n are governed by the transition probabilities p_{ij}, and the Y_n are generated from the X_n according to

$$r_{km} = P(Y_n = m | X_n = k) \tag{11.46}$$

Typically the idea is to guess the X_n from the Y_n and our knowledge of the p_{ij} and r_{km}. The details are too complex to give here, but you can at least understand that Bayes' Rule comes into play.

A good example of HMMs would be in text mining applications. Here the Y_n might be words in the text, and X_n would be their parts of speech (POS)—nouns, verbs, adjectives and so on. Consider the word *round*, for instance. Your first thought might be that it is an adjective, but it could be a noun (e.g. an elimination round in a tournament) or a verb (e.g. to round off a number or round a corner). The HMM would help us to guess which, and therefore guess the true meaning of the word.

HMMs are also used in speech process, DNA modeling and many other applications.

11.4 Continuous-Time Markov Chains

In the Markov chains we analyzed above, events occur only at integer times. However, many Markov chain models are of the **continuous-time** type, in which events can occur at any times. Here the **holding time**, i.e. the time the system spends in one state before changing to another state, is a continuous random variable.

The state of a Markov chain at any time now has a continuous subscript. Instead of the chain consisting of the random variables $X_n,\ n = 1, 2, 3, ...$ (you can also start n at 0 in the sense of Section 11.1.2.4), it now consists of $\{X_t : t \in [0, \infty)\}$. The Markov property is now

$$P(X_{t+u} = k|X_s \text{ for all } 0s \leq t) = P(X_{t+u} = k|X_t) \text{ for all } t, u \geq 0 \tag{11.47}$$

11.4.1 Holding-Time Distribution

In order for the Markov property to hold, the distribution of holding time at a given state needs to be "memoryless." You may recall that exponentially distributed random variables have this property. In other words, if a random variable W has density

$$f(t) = \lambda e^{-\lambda t} \tag{11.48}$$

for some λ then

$$P(W > r + s|W > r) = P(W > s) \tag{11.49}$$

for all positive r and s. Actually, one can show that exponential distributions are the only continuous distributions which have this property. Therefore, *holding times in Markov chains must be exponentially distributed.*

It is difficult for the beginning modeler to fully appreciate the memoryless property. You are urged to read the material on exponential distributions in Section 4.4.5.1 before continuing.

Because it is central to the Markov property, the exponential distribution is assumed for all basic activities in Markov models. In queuing models, for instance, both the interarrival time and service time are assumed to be exponentially distributed (though of course with different values of λ). In reliability modeling, the lifetime of a component is assumed to have an exponential distribution.

Such assumptions have in many cases been verified empirically. If you go to a bank, for example, and record data on when customers arrive at the door, you will find the exponential model to work well (though you may have to restrict yourself to a given time of day, to account for nonrandom effects such as heavy traffic at the noon hour). In a study of time to failure for airplane air conditioners, the distribution was also found

to be well fitted by an exponential density. On the other hand, in many cases the distribution is not close to exponential, and purely Markovian models cannot be used for anything more than a rough approximation.

11.4.2 The Notion of "Rates"

A key point is that the parameter λ in (11.48) has the interpretation of a rate, in the sense we will now discuss. First, recall that $1/\lambda$ is the mean. Say light bulb lifetimes have an exponential distribution with mean 100 hours, so $\lambda = 0.01$. In our lamp, whenever its bulb burns out, we immediately replace it with a new on. Imagine watching this lamp for, say, 100,000 hours. During that time, we will have done approximately 100000/100 = 1000 replacements. That would be using 1000 light bulbs in 100000 hours, so we are using bulbs at the rate of 0.01 bulb per hour. For a general λ, we would use light bulbs at the rate of λ bulbs per hour. This concept is crucial to what follows.

11.4.3 Stationary Distribution

We again define π_i to be the long-run proportion of time the system is in state i, and we again will derive a system of linear equations to solve for these proportions.

11.4.3.1 Intuitive Derivation

To this end, let λ_i denote the parameter in the holding-time distribution at state i, and define the quantities

$$\rho_{rs} = \lambda_r p_{rs} \tag{11.50}$$

with the following interpretation. In the context of the ideas in our example of the rate of light bulb replacements in Section 11.4.2, one can view (11.50) as the rate of transitions from r to s, *during the time we are in state r*.

Then, equating the rate of transitions into i and the rate out of i, we have

$$\pi_i \lambda_i = \sum_{j \neq i} \pi_j \lambda_j p_{ji} \tag{11.51}$$

These equations can then be solved for the π_i.

11.4.3.2 Computation

Motivated by (11.51), define the matrix Q by

$$q_{ij} = \begin{cases} \lambda_j p_{ji}, & \text{if } i \neq j \\ -\lambda_i, & \text{if } i = j \end{cases} \tag{11.52}$$

Q is called the **infinitesimal generator** of the system, so named because it is the basis of the system of differential equations that can be used to find the finite-time probabilistic behavior of X_t.

The name also reflects the rates notion we've been discussing, due to the fact that, say in our light bulb example in Section 11.4.2,

$$P(\text{bulb fails in next h time}) = \lambda h + o(h) \tag{11.53}$$

Then (11.51) is stated in matrix form as

$$Q'\pi = 0 \tag{11.54}$$

Here is R code to solve the system:

```
findpicontin <- function(q) {
   n <- nrow(q)
   newq <- t(q)
   newq[n,] <- rep(1,n)
   rhs <- c(rep(0,n-1),1)
   pivec <- solve(newq,rhs)
   return(pivec)
}
```

To solve the equations (11.51), we'll need a property of exponential distributions derived previously in Section 5.5.7, copied here for convenience:

Theorem 38 *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*

(a) Z is exponentially distributed with parameter $\lambda_1 + ... + \lambda_k$

(b) $P(Z = W_i) = \frac{\lambda_i}{\lambda_1 + ... + \lambda_k}$

11.4.4 Example: Machine Repair

Suppose the operations in a factory require the use of a certain kind of machine. The manager has installed two of these machines. This is known as a **gracefully degrading system**: When both machines are working, the fact that there are two of them, instead of one, leads to a shorter wait time for access to a machine. When one machine has failed, the wait is longer, but at least the factory operations may continue. Of course, if both machines fail, the factory must shut down until at least one machine is repaired.

Suppose the time until failure of a single machine, carrying the full load of the factory, has an exponential distribution with mean 20.0, but the mean is 25.0 when the other machine is working, since it is not so loaded. Repair time is exponentially distributed with mean 8.0.

We can take as our state space $\{0,1,2\}$, where the state is the number of working machines. Now, let us find the parameters λ_i and p_{ji} for this system. For example, what about λ_2? The holding time in state 2 is the minimum of the two lifetimes of the machines, and thus from the results of Section 5.5.7, has parameter $\frac{1}{25.0} + \frac{1}{25.0} = 0.08$.

For λ_1, a transition out of state 1 will be either to state 2 (the down machine is repaired) or to state 0 (the up machine fails). The time until transition will be the minimum of the lifetime of the up machine and the repair time of the down machine, and thus will have parameter $\frac{1}{20.0} + \frac{1}{8.0} = 0.175$. Similarly, $\lambda_0 = \frac{1}{8.0} + \frac{1}{8.0} = 0.25$.

It is important to understand how the Markov property is being used here. Suppose we are in state 1, and the down machine is repaired, sending us into state 2. Remember, the machine which had already been up has "lived" for some time now. But the memoryless property of the exponential distribution implies that this machine is now "born again."

What about the parameters p_{ji}? Well, p_{21} is certainly easy to find; since the transition $2 \rightarrow 1$ is the *only* transition possible out of state 2, $p_{21} = 1$.

For p_{12}, recall that transitions out of state 1 are to states 0 and 2, with rates 1/20.0 and 1/8.0, respectively. So,

$$p_{12} = \frac{1/8.0}{1/20.0 + 1/8.0} = 0.72 \tag{11.55}$$

Working in this manner, we finally arrive at the complete system of equations (11.51):

$$\pi_2(0.08) = \pi_1(0.125) \tag{11.56}$$

$$\pi_1(0.175) = \pi_2(0.08) + \pi_0(0.25) \tag{11.57}$$

$$\pi_0(0.25) = \pi_1(0.05) \tag{11.58}$$

Of course, we also have the constraint $\pi_2 + \pi_1 + \pi_0 = 1$. The solution turns out to be

$$\pi = (0.072, 0.362, 0.566) \tag{11.59}$$

Thus for example, during 7.2% of the time, there will be no machine available at all.

Several variations of this problem could be analyzed. We could compare the two-machine system with a one-machine version. It turns out that the proportion of down time (i.e. time when no machine is available) increases to 28.6%. Or we could analyze the case in which only one repair person is employed by this factory, so that only one machine can be repaired at a time, compared to the situation above, in which we (tacitly) assumed that if both machines are down, they can be repaired in parallel. We leave these variations as exercises for the reader.

11.4.5 Example: Migration in a Social Network

The following is a simplified version of research in online social networks.

There is a town with two social groups, with each person being in exactly one group. People arrive from outside town, with exponentially distributed interarrival times at rate α, and join one of the groups with probability 0.5 each. Each person will occasionally switch groups, with one possible "switch" being to leave town entirely. A person's time before switching is exponentially distributed with rate σ; the switch will either be to the other group or to the outside world, with probabilities q and 1-q, respectively. Let the state of the system be (i,j), where i and j are the number of current members in groups 1 and 2, respectively.

Let's find a typical balance equation, say for the state (8,8):

$$\pi_{(8,8)}(\alpha + 16 \cdot \sigma) = (\pi_{(9,8)} + \pi_{(8,9)}) \cdot 9\sigma(1-q) + (\pi_{(9,7)} + \pi_{(7,9)}) \cdot 9\sigma q + (\pi_{(8,7)} + \pi_{(7,8)}) \cdot 0.5\alpha \tag{11.60}$$

The reasoning is straightforward. How can we move out of state (8,8)? Well, there could be an arrival (rate α), or any one of the 16 people could switch groups (rate 16σ), etc.

Now, in a "going beyond finding the π" vein, let's find the long-run fraction of transfers into group 1 that come from group 2, as opposed to from the outside.

The rate of transitions into that group from outside is 0.5α. When the system is in state (i,j), the rate of transitions into group 1 from group 2 is $j\sigma q$, so the overall rate is $\sum_{i,j} \pi_{(i,j)} j\sigma q$. Thus the fraction of new members coming in to group 1 from transfers is

$$\frac{\sum_{i,j} \pi_{(i,j)} j\sigma q}{\alpha + \sum_{i,j} \pi_{(i,j)} j\sigma q} \tag{11.61}$$

The above reasoning is very common, quite applicable in many situations. By the way, note that $\sum_{i,j} \pi_{(i,j)} j \sigma q = \sigma q E N$, where N is the number of members of group 1.

11.4.6 Continuous-Time Birth/Death Processes

We noted earlier that the system of equations for the π_i may not be easy to solve. In many cases, for instance, the state space is infinite and thus the system of equations is infinite too. However, there is a rich class of Markov chains for which closed-form solutions have been found, called **birth/death processes**.[7]

Here the state space consists of (or has been mapped to) the set of nonnegative integers, and p_{ji} is nonzero only in cases in which $|i - j| = 1$. (The name "birth/death" has its origin in Markov models of biological populations, in which the state is the current population size.) Note for instance that the example of the gracefully degrading system above has this form. An M/M/1 queue—one server, "Markov" (i.e. exponential) interarrival times and Markov service times—is also a birth/death process, with the state being the number of jobs in the system.

Because the p_{ji} have such a simple structure, there is hope that we can find a closed-form solution to (11.51), and it turns out we can. Let $u_i = \rho_{i,i+1}$ and $d_i = \rho_{i,i-1}$ ('u' for "up," 'd' for "down"). Then (11.51) is

$$\pi_{i+1} d_{i+1} + \pi_{i-1} u_{i-1} = \pi_i \lambda_i = \pi_i (u_i + d_i),\ i \geq 1 \tag{11.62}$$

$$\pi_1 d_1 = \pi_0 \lambda_0 = \pi_0 u_0 \tag{11.63}$$

In other words,

$$\pi_{i+1} d_{i+1} - \pi_i u_i = \pi_i d_i - \pi_{i-1} u_{i-1},\ i \geq 1 \tag{11.64}$$

$$\pi_1 d_1 - \pi_0 u_0 = 0 \tag{11.65}$$

Applying (11.64) recursively to the base (11.65), we see that

$$\pi_i d_i - \pi_{i-1} u_{i-1} = 0,\ i \geq 1 \tag{11.66}$$

so that

$$\pi_i = \pi_{i-1} \frac{u_{i-1}}{d_i}\ i \geq 1 \tag{11.67}$$

[7] Though we treat the continuous-time case here, there is also a discrete-time analog.

and thus

$$\pi_i = \pi_0 r_i \tag{11.68}$$

where

$$r_i = \Pi_{k=1}^{i} \frac{u_{k-1}}{d_k} \tag{11.69}$$

where $r_i = 0$ for $i > m$ if the chain has no states past m.

Then since the π_i must sum to 1, we have that

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} r_i} \tag{11.70}$$

and the other π_i are then found via (11.68).

Note that the chain might be finite, i.e. have $u_i = 0$ for some i. In that case it is still a birth/death chain, and the formulas above for π still apply.

11.5 Hitting Times Etc.

In this section we're interested in the amount of time it takes to get from one state to another, including cases in which this might be infinite.

11.5.1 Some Mathematical Conditions

There is a rich mathematical theory regarding the asymptotic behavior of Markov chains. We will not present such material here in this brief introduction, but we will give an example of the implications the theory can have.

A state in a Markov chain is called **recurrent** if it is guaranteed that, if we start at that state, we will return to the state infinitely many times. A nonrecurrent state is called **transient.**

Let T_{ii} denote the time needed to return to state i if we start there. Keep in mind that T_{ii} is the time from one entry to state i to the next entry to state i. So, it includes time spent in i, which is 1 unit of time for a discrete-time chain and a random exponential amount of time in the continuous-time case, and then time spent away from i, up to the time of next entry to i. Note that an equivalent definition of recurrence is that

$P(T_{ii} < \infty) = 1$, i.e. we are sure to return to i at least once. By the Markov property, if we are sure to return once, then we are sure to return again once after that, and so on, so this implies infinitely many visits.

A recurrent state i is called **positive recurrent** if $E(T_{ii}) < \infty$, while a state which is recurrent but not positive recurrent is called **null recurrent**.

Let T_{ij} be the time it takes to get to state j if we are now in i. Note that this is measured from the time that we enter state i to the time we enter state j.

One can show that in the discrete time case, a state i is recurrent if and only if

$$\sum_{n=0}^{\infty} P(T_{ii} = n) = \infty \tag{11.71}$$

This can be easily seen in the "only if" case: Let A_n denote the indicator random variable for the event $T_{ii} = n$ (Section 3.6). Then $P(T_{ii} = n) = EA_n$, so the left-hand side of (11.71) is the expected value of the total number of visits to state i. If state i is recurrent, then we will visit i infinitely often, and thus that sum should be equal to infinity.

Consider an **irreducible** Markov chain, meaning one which has the property that one can get from any state to any other state (though not necessarily in one step). One can show that in an irreducible chain, if one state is recurrent then they all are. The same statement holds if "recurrent" is replaced by "positive recurrent."

Again, this should make intuitive sense to you for the recurrent case: We make infinitely many visits to state i, and each time we have a nonzero probability of going to state j from there. Thus we should make infinitely many visits to j as well.

11.5.2 Example: Random Walks

Consider the famous **random walk** on the full set of integers: At each time step, one goes left one integer or right one integer (e.g. to +3 or +5 from +4), with probability 1/2 each. In other words, we flip a coin and go left for heads, right for tails.

If we start at 0, then we return to 0 when we have accumulated an equal number of heads and tails. So for even-numbered n, i.e. n = 2m, we have

$$P(T_{ii} = n) = P(\text{m heads and m tails}) = \binom{2m}{m} \frac{1}{2^{2m}} \tag{11.72}$$

One can use Stirling's approximation,

$$m! \approx \sqrt{2\pi} e^{-m} m^{m+1/2} \tag{11.73}$$

to show that the series (11.71) diverges in this case. So, this chain (meaning all states in the chain) is recurrent. However, it turns out not to be not positive recurrent, as we'll see below.

The same is true for the corresponding random walk on the two-dimensional integer lattice (moving up, down, left or right with probability 1/4 each). However, in the three-dimensional case, the chain is not even null recurrent; it is transient.

11.5.3 Finding Hitting and Recurrence Times

For a positive recurrent state i in a discrete-time Markov chain,

$$\pi_i = \frac{1}{E(T_{ii})} \tag{11.74}$$

The approach to deriving this is similar to that of Section 11.1.5.1. Define alternating On and Off subcycles, where On means we are at state i and Off means we are elsewhere. An On subcycle has duration 1, and an Off subcycle has duration $T_{ii} - 1$. Define a full cycle to consist of an On subcycle followed by an Off subcycle.

Then intuitively the proportion of time we are in state i is

$$\pi_i = \frac{E(\text{On})}{E(\text{On}) + E(\text{Off})} = \frac{1}{ET_{ii}} \tag{11.75}$$

The equation is similar for the continuous-time case. Here $E(\text{On}) = 1/\lambda_i$. The Off subcycle has mean duration $ET_{ii} - 1/\lambda_i$. Note again that T_{ii} is measured from the time we enter state i once until the time we enter it again. We then have

$$\pi_i = \frac{1/\lambda_i}{E(T_{ii})} \tag{11.76}$$

Thus positive recurrence means that $\pi_i > 0$. For a null recurrent chain, the limits in Equation (11.3) are 0, which means that there may be rather little one can say of interest regarding the long-run behavior of the chain.

We are often interested in finding quantities of the form $E(T_{ij})$. We can do so by setting up systems of equations similar to the balance equations used for finding stationary distributions.

First consider the discrete case. Conditioning on the first step we take after being at state i, and using the

Law of Total Expectation, we have

$$E(T_{ij}) = \sum_{k \neq j} p_{ik}[1 + E(T_{kj})] + p_{ij} \cdot 1 = 1 + \sum_{k \neq j} p_{ik} E(T_{kj}) \tag{11.77}$$

By varying i and j in (11.77), we get a system of linear equations which we can solve to find the ET_{ij}. Note that (11.74) gives us equations we can use here too.

The continuous version uses the same reasoning:

$$E(T_{ij}) = \sum_{k \neq j} p_{ik} \left[\frac{1}{\lambda_i} + E(T_{kj}) \right] + p_{ij} \cdot \frac{1}{\lambda_i} = \frac{1}{\lambda_i} + \sum_{k \neq j} p_{ik} E(T_{kj}) \tag{11.78}$$

One can use a similar analysis to determine the probability of ever reaching a state, in chains in which this probability is not 1. (Some chains have have transient or even **absorbing** states, i.e. states u such that $p_{uv} = 0$ whenever $v \neq u$.)

For fixed j define

$$\alpha_{ij} = P(T_{ij} < \infty) \tag{11.79}$$

Then denoting by S the state we next visit after i, we have

$$\begin{aligned}
\alpha_{ij} &= P(T_{ij} < \infty) && (11.80)\\
&= \sum_k P(S = k \text{ and } T_{ij} < \infty) && (11.81)\\
&= \sum_{k \neq j} P(S = k \text{ and } T_{kj} < \infty) + P(S = j) && (11.82)\\
&= \sum_{k \neq j} P(S = k)\ P(T_{kj} < \infty | S = k) + P(S = j) && (11.83)\\
&= \sum_{k \neq j} p_{ik}\ \alpha_{kj} + p_{ij} && (11.84)\\
& && (11.85)
\end{aligned}$$

So, again we have a system of linear equations that we can solve for the α_{ij}.

11.5.4 Example: Finite Random Walk

Let's go back to the example in Section 11.1.1.

Suppose we start our random walk at 2. How long will it take to reach state 4? Set $b_i = E(T_{i4}|\text{start at i})$. From (11.77) we could set up equations like

$$b_2 = \frac{1}{3}(1 + b_1) + \frac{1}{3}(1 + b_2) + \frac{1}{3}(1 + b_3) \tag{11.86}$$

Now change the model a little, and make states 1 and 6 absorbing. Suppose we start at position 3. What is the probability that we eventually are absorbed at 6 rather than 1? We could set up equations like (11.80) to find this.

11.5.5 Example: Tree-Searching

Consider the following Markov chain with infinite state space {0,1,2,3,...}.[8] The transition matrix is defined by $p_{i,i+1} = q_i$ and $p_{i0} = 1 - q_i$. This kind of model has many different applications, including in computer science tree-searching algorithms. (The state represents the level in the tree where the search is currently, and a return to 0 represents a backtrack. More general backtracking can be modeled similarly.)

The question at hand is, What conditions on the q_i will give us a positive recurrent chain?

Assuming $0 < q_i < 1$ for all i, the chain is clearly irreducible. Thus, to check for recurrence, we need check only one state, say state 0.

For state 0 (and thus the entire chain) to be recurrent, we need to show that $P(T_{00} < \infty) = 1$. But

$$P(T_{00} > n) = \Pi_{i=0}^{n-1} q_i \tag{11.87}$$

Therefore, the chain is recurrent if and only if

$$\lim_{n\to\infty} \Pi_{i=0}^{n-1} q_i = 0 \tag{11.88}$$

For positive recurrence, we need $E(T_{00}) < \infty$. Now, one can show that for any nonnegative integer-valued

[8] Adapted from *Performance Modelling of Communication Networks and Computer Architectures*, by P. Harrison and N. Patel, pub. by Addison-Wesley, 1993.

random variable Y

$$E(Y) = \sum_{n=0}^{\infty} P(Y > n) \tag{11.89}$$

Thus for positive recurrence, our condition on the q_i is

$$\sum_{n=0}^{\infty} \Pi_{i=0}^{n-1} q_i < \infty \tag{11.90}$$

Exercises

1. Consider a "wraparound" variant of the random walk in Section 11.1.1. We still have a reflecting barrier at 1, but at 5, we go back to 4, stay at 5 or "wrap around" to 1, each with probability 1/3. Find the new set of stationary probabilities.

2. Consider the Markov model of the shared-memory multiprocessor system in our PLN. In each part below, your answer will be a function of $q_1, ..., q_m$.

(a) For the case m = 3, find $p_{(2,0,1),(1,1,1)}$.

(b) For the case m = 6, give a compact expression for $p_{(1,1,1,1,1,1),(i,j,k,l,m,n)}$.
Hint: We have an instance of a famous parametric distribution family here.

3. This problem involves the analysis of call centers. This is a subject of much interest in the business world, with there being commercial simulators sold to analyze various scenarios. Here are our assumptions:

- Calls come in according to a Poisson process with intensity parameter λ.
- Call duration is exponentially distributed with parameter η.
- There are always at least b operators in service, and at most b+r.
- Operators work from home, and can be brought into or out of service instantly when needed. They are paid only for the time in service.
- If a call comes in when the current number of operators is larger than b but smaller than b+r, another operator is brought into service to process the call.
- If a call comes in when the current number of operators is b+r, the call is rejected.

- When an operator completes processing a call, and the current number of operators (including this one) is greater than b, then that operator is taken out of service.

Note that this is a birth/death process, with the state being the number of calls currently in the system.

(a) Find approximate closed-form expressions for the π_i for large b+r, in terms of b, r, λ and η. (You should not have any summation symbols.)

(b) Find the proportion of rejected calls, in terms of π_i and b, r, λ and η.

(c) An operator is paid while in service, even if he/she is idle, in which case the wages are "wasted." Express the proportion of wasted time in terms of the π_i and b, r, λ and η.

(d) Suppose b = r = 2, and $\lambda = \eta = 1.0$. When a call completes while we are in state b+1, an operator is sent away. Find the mean time until we make our next summons to the reserve pool.

4. The **bin-packing problem** arises in many computer science applications. Items of various sizes must be placed into fixed-sized bins. The goal is to find a packing arrangement that minimizes unused space. Toward that end, work the following problem.

We are working in one dimension, and have a continuing stream of items arriving, of lengths $L_1, L_2, L_3, ...$ We place the items in the bins in the order of arrival, i.e. without optimizing. We continue to place items in a bin until we encounter an item that will not fit in the remaining space, in which case we go to the next bin.

Suppose the bins are of length 5, and an item has length 1, 2, 3 or 4, with probability 0.25 each. Find the long-run proportion of wasted space.

Hint: Set up a discrete-time Markov chain, with "time" being the number of items packed so far, and the state being the amount of occupied space in the current bin. Define T_n to be 1 or 0, according to whether the n^{th} item causes us to begin packing a new bin, so that the number of bins used by "time" n is $T_1 + ... + T_n$.

5. Suppose we keep rolling a die. Find the mean number of rolls needed to get three consecutive 4s.

Hint: Use the material in Section 11.5.

6. A system consists of two machines, with exponentially distributed lifetimes having mean 25.0. There is a single repairperson, but he is not usually on site. When a breakdown occurs, he is summoned (unless he is already on his way or on site), and it takes him a random amount of time to reach the site, exponentially distributed with mean 2.0. Repair time is exponentially distributed with mean 8.0. If after completing a repair the repairperson finds that the other machine needs fixing, he will repair it; otherwise he will leave. Repair is performed on a First Come, First Served schedule. Find the following:

(a) The long-run proportion of the time that the repairperson is on site.

(b) The rate per unit time of calls to the repairperson.

(c) The mean time to repair, i.e. the mean time between a breakdown of a machine and completion of repair of that machine.

(d) The probability that, when two machines are up and one of them goes down, the second machine fails before the repairperson arrives.

7. Consider again the random walk in Section 11.1.1. Find

$$\lim_{n\to\infty} \rho(X_n, X_{n+1}) \tag{11.91}$$

Hint: Apply the Law of Total Expectation to $E(X_n X_{n+1})$.

8. Consider a random variable X that has a continuous density. That implies that $G(u) = P(X > u)$ has a derivative. Differentiate (11.49) with respect to r, then set r = 0, resulting in a differential equation for G. Solve that equation to show that the only continuous densities that produce the memoryless property are those in the exponential family.

9. Suppose we model a certain database as follows. New items arrive according to a Poisson process with intensity parameter α. Each item stays in the database for an exponentially distributed amount of time with parameter σ, independently of the other items. Our state at time t is the number of items in the database at that time. Find closed-form expressions for the stationary distribution π and the long-run average size of the database.

10. Consider our machine repair example in Section 11.4.4, with the following change: The repairperson is offsite, and will not be summoned unless both machines are down. Once the repairperson arrives, she will not leave until both machines are up. So for example, if she arrives and repairs machine B, then while repairing A finds that B has gone down again, she will start work on B immediately after finishing with A. Travel time to the site from the maintenance office is 0. Repair is performed on a First Come, First Served schedule. The time a machine is in working order has an exponential distribution with rate ω, and repair is exponentially distributed with rate ρ. Find the following in terms of ω and ρ:

(a) The long-run proportion of the time that the repairperson is on site.

(b) The rate per unit time of calls to the repairperson.

(c) The mean time to repair, i.e. the mean time between a breakdown of a machine and completion of repair of that machine. (Hint: The best approach is to look at rates. First, find the number of breakdowns per unit time. Then, ask how many of these occur during a time when both machines are up, etc. In each case, what is the mean time to repair for the machine that breaks?)

11. There is a town with two social groups, with the following dynamics:

- Everyone is in exactly one group at a time.
- People arrive from outside town, with exponentially distributed interarrival times at rate α, and join one of the groups with probability 0.5 each.
- Each person will occasionally switch groups, with one possible "group" being to leave town entirely (never to return). A person's time before switching groups is exponentially distributed with rate σ. The switch will either be to the other group or to the outside world, with probabilities q and 1-q, respectively.

Let the state of the system be (i,j), where i and j are the number of current members in groups 1 and 2, respectively. Answer in terms of α, λ, τ and π:

(a) Give the balance equation for the state (8,8).

(b) Fill in the blank: The president of Group 1 tells reporter, "We've found over the years that _____% of entries into our group come as transfers from the other group."

Chapter 12

Introduction to Queuing Models

Seems like we spend large parts of our lives standing in line (or as they say in New York, standing "on" line). This can be analyzed probabilistically, a subject we will be introduced in this chapter.

12.1 Introduction

Like other areas of applied stochastic processes, queuing theory has a vast literature, covering a huge number of variations on different types of queues. Our tutorial here can only just scratch the surface to this field.

Here is a rough overview of a few of the large categories of queuing theory:

- Single-server queues.
- Networks of queues, including **open** networks (in which jobs arrive from outside the network, visit some of the servers in the network, then leave) and **closed** networks (in which jobs continually circulate within the network, never leaving).
- Non-First Come, First Served (FCFS) service orderings. For example, there are Last Come, First Served (i.e. stacks) and Processor Sharing (which models CPU timesharing).

In this brief introduction, we will not discuss non-FCFS queues, and will only scratch the surface on the other tops.

12.2 M/M/1

The first M here stands for "Markov" or "memoryless," alluding to the fact that arrivals to the queue are Markovian, i.e. interarrivals are i.i.d. exponentially distributed. The second M means that the service times are also i.i.d. exponential. Denote the reciprocal-mean interarrival and service times by λ and μ.

The 1 in M/M/1 refers to the fact that there is a single server. We will assume FCFS job scheduling here, but close inspection of the derivation will show that it applies to some other kinds of scheduling too.

This system is a continuous-time Markov chain, with the state X_t at time t being the number of jobs in the system (not just in the queue but also including the one currently being served, if any).

12.2.1 Steady-State Probabilities

Intuitively the steady-state probabilities π_i will exist only if $\lambda < \mu$. Otherwise jobs would come in faster than they could be served, and the queue would become infinite. So, we assume that $u < 1$, where $u = \frac{\lambda}{\mu}$.

Clearly this is a birth-and-death chain. For state k, the birth rate $\rho_{k,k+1}$ is λ and the death rate $\rho_{k,k-1}$ is μ, k = 0,1,2,... (except that the death rate at state 0 is 0). Using the formula derived for birth/death chains, we have that

$$\pi_i = u^i \pi_0, \ i \geq 0 \tag{12.1}$$

and

$$\pi_0 = \frac{1}{\sum_{j=0}^{\infty} u^j} = 1 - u \tag{12.2}$$

In other words,

$$\pi_i = u^i (1 - u), \ i \geq 0 \tag{12.3}$$

Note by the way that since $\pi_0 = 1 - u$, then u is the *utilization* of the server, i.e. the proportion of the time the server is busy. In fact, this can be seen intuitively: Think of a very long period of time of length t. During this time approximately λt jobs having arrived, keeping the server busy for approximately $\lambda t \cdot \frac{1}{\mu}$ time. Thus the fraction of time during which the server is busy is approximantely

$$\frac{\lambda t \cdot \frac{1}{\mu}}{t} = \frac{\lambda}{\mu} \tag{12.4}$$

12.2.2 Mean Queue Length

Another way to look at Equation (12.3) is as follows. Let the random variable N have the long-run distribution of X_t, so that

$$P(N = i) = u^i(1 - u),\ i \geq 0 \tag{12.5}$$

Then this says that N+1 has a geometric distribution, with "success" probability 1-u. (N itself is not quite geometrically distributed, since N's values begin at 0 while a geometric distribution begins at 1.)

Thus the long-run average value E(N) for X_t will be the mean of that geometric distribution, minus 1, i.e.

$$EN = \frac{1}{1-u} - 1 = \frac{u}{1-u} \tag{12.6}$$

The long-run mean queue length E(Q) will be this value minus the mean number of jobs being served. The latter quantity is $1 - \pi_0 = u$, so

$$EQ = \frac{u^2}{1-u} \tag{12.7}$$

12.2.3 Distribution of Residence Time/Little's Rule

Let R denote the **residence time** of a job, i.e. the time elapsed from the job's arrival to its exit from the system. Little's Rule says that

$$EN = \lambda ER \tag{12.8}$$

This property holds for a variety of queuing systems, including this one. It can be proved formally, but here is the intuition:

Think of a particular job (in the literature of queuing theory, it is called a "tagged job") at the time it has just exited the system. If this is an "average" job, then it has been in the system for ER amount of time, during which an average of λER new jobs have arrived behind it. These jobs now comprise the total number of jobs in the system, which in the average case is EN.

Applying Little's Rule here, we know EN from Equation (12.6), so we can solve for ER:

$$ER = \frac{1}{\lambda}\frac{u}{1-u} = \frac{1/\mu}{1-u} \tag{12.9}$$

With a little more work, we can find the actual distribution of R, not just its mean. This will enable us to obtain quantities such as Var(R) and P(R > z). Here is our approach:

When a job arrives, say there are N jobs ahead of it, including one in service. Then this job's value of R can be expressed as

$$R = S_{self} + S_{1,resid} + S_2 + ... + S_N \tag{12.10}$$

where S_{self} is the service time for this job, $S_{1,resid}$ is the remaining time for the job now being served (i.e. the residual life), and for $i > 1$ the random variable S_i is the service time for the i^{th} waiting job.

Then the Laplace transform of R, evaluated at say w, is

$$\begin{aligned} E(e^{-wR}) &= E[e^{-w(S_{self}+S_{1,resid}+S_2+...+S_N)}] && (12.11)\\ &= E\left(E[e^{-w(S_{self}+S_{1,resid}+S_2+...+S_N)}|N]\right) && (12.12)\\ &= E[\{E(e^{-wS})\}^{N+1}] && (12.13)\\ &= E[g(w)^{N+1}] && (12.14) \end{aligned}$$

where

$$g(w) = E(e^{-wS}) \tag{12.15}$$

is the Laplace transform of the service variable, i.e. of an exponential distribution with parameter equal to the service rate μ. Here we have made use of these facts:

- The Laplace transform of a sum of independent random variables is the product of their individual Laplace transforms.
- Due to the memoryless property, $S_{1,resid}$ has the same distribution as do the other S_i.
- The distribution of the service times S_i and queue length N observed by our tagged job is the same as the distributions of those quantities at all times, not just at arrival times of tagged jobs. This property can be proven for this kind of queue and many others, and is called PASTA—Poisson Arrivals See Time Averages.

  (Note that the PASTA property is not obvious. On the contrary, given our experience with the Bus Paradox and length-biased sampling in Section 6.3, we should be wary of such things. But the PASTA property does hold and can be proven.)

But that last term in (12.14), $E[g(w)^{N+1}]$, is the generating function of N+1, evaluated at g(w). And we know from Section 12.2.2 that N+1 has a geometric distribution. The generating function for a nonnegative-integer valued random variable K with success probability p is

$$g_K(s) = E(s^K) = \sum_{i=1}^{\infty} s^i (1-p)^{i-1} p = \frac{ps}{1-s(1-p)} \tag{12.16}$$

In (12.14), we have p = 1 -u and s = g(w). So,

$$E(v^{N+1}) = \frac{g(w)(1-u)}{1-u[g(w)]} \tag{12.17}$$

Finally, by definition of Laplace transform,

$$g(w) = E(e^{-wS}) = \int_0^{\infty} e^{-wt} \mu e^{-\mu t} dt = \frac{\mu}{w+\mu} \tag{12.18}$$

So, from (12.11), (12.17) and (12.18), the Laplace transform of R is

$$\frac{\mu(1-u)}{w+\mu(1-u)} \tag{12.19}$$

In principle, Laplace transforms can be inverted, and we could use numerical methods to retrieve the distribution of R from (12.19). But hey, look at that! Equation (12.19) has the same form as (12.18). In other words, we have discovered that R has an exponential distribution too, only with parameter $\mu(1-u)$ instead of μ.

This is quite remarkable. The fact that the service and interarrival times are exponential doesn't mean that everything else will be exponential too, so it is surprising that R does turn out to have an exponential distribution.

It is even more surprising in that R is a sum of independent exponential random variables, as we saw in (12.10), and we know that such sums have Erland distributions. The resolution of this seeming paradox is that the number of terms N in (12.10) is itself random. Conditional on N, R has an Erlang distribution, but unconditionally R has an exponential distribution.

12.3 Multi-Server Models

Here we have c servers, with a common queue. There are many variations.

12.3.1 M/M/c

Here the servers are homogeneous. When a job gets to the head of the queue, it is served by the first available server.

The state is again the number of jobs in the system, including any jobs at the servers. Again it is a birth/death chain, with $u_{i,i+1} = \lambda$ and

$$u_{i,i-1} = \begin{cases} i\mu, & \text{if 0 < i < c} \\ c\mu, & \text{if i} \geq \text{c} \end{cases} \tag{12.20}$$

The solution turns out to be

$$\pi_k = \begin{cases} \pi_0 (\frac{\lambda}{\mu})^k \frac{1}{k!}, & \text{k} < c \\ \pi_0 (\frac{\lambda}{\mu})^k \frac{1}{c! c^{k-c}}, & \text{k} \geq c \end{cases} \tag{12.21}$$

where

$$\pi_0 = \left[\sum_{k=0}^{c-1} \frac{(cu)^k}{k!} + \frac{(cu)^c}{c!} \frac{1}{1-u} \right]^{-1} \tag{12.22}$$

and

$$u = \frac{\lambda}{c\mu} \tag{12.23}$$

Note that the latter quantity is still the utilization per server, using an argument similar to that which led to (12.4).

Recalling that the Taylor series for e^z is $\sum_{k=0}^{\infty} z^k/k!$ we see that

$$\pi_0 \approx e^{-cu} \tag{12.24}$$

12.3.2 M/M/2 with Heterogeneous Servers

Here the servers have different rates. We'll treat the case in which c = 2. Assume $\mu_1 < \mu_2$. When a job reaches the head of the queue, it chooses machine 2 if that machine is idle, and otherwise waits for the first available machine. Once it starts on a machine, it cannot be switched to the other.

Denote the state by (i,j,k), where

- i is the number of jobs at server 1
- j is the number of jobs at server 2
- k is the number of jobs in the queue

The key is to notice that states 111, 112, 113, ... act like the M/M/k queue. This will reduce finding the solution of the balance equations to solving a finite system of linear equations.

For $k \geq 1$ we have

$$(\lambda + \mu_1 + \mu_2)\pi_{11k} = (\mu_1 + \mu_2)\pi_{11,k+1} + \lambda\pi_{11,k-1} \tag{12.25}$$

Collecting terms as in the derivation of the stationary distribution for birth/death processes, (12.25) becomes

$$\lambda(\pi_{11k} - \pi_{11,k-1}) = (\mu_1 + \mu_2)(\pi_{11,k+1} - \pi_{11k}),\ k = 1, 2, ... \tag{12.26}$$

Then we have

$$(\mu_1 + \mu_2)\pi_{11,k+1} - \lambda\pi_{11k} = (\mu_1 + \mu_2)\pi_{11k} - \lambda\pi_{11,k-1} \tag{12.27}$$

So, we now have all the π_{11i}, i = 2,3,... in terms of π_{111} and π_{110}, thus reducing our task to solving a finite set of linear equations, as promised. Here are the rest of the equations:

$$\lambda\pi_{000} = \mu_2\pi_{010} + \mu_1\pi_{100} \tag{12.28}$$

$$(\lambda + \mu_2)\pi_{010} = \lambda\pi_{000} + \mu_1\pi_{110} \tag{12.29}$$

$$(\lambda + \mu_1)\pi_{100} = \mu_2\pi_{110} \tag{12.30}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{110} = \lambda\pi_{010} + \lambda\pi_{100} + (\mu_1 + \mu_2)\pi_{111} \tag{12.31}$$

From (3.46), we have

$$(\mu_1 + \mu_2)\pi_{111} - \lambda\pi_{110} = (\mu_1 + \mu_2)\pi_{110} - \lambda(\pi_{010} + \pi_{100}) \tag{12.32}$$

Look at that last term, $\lambda(\pi_{010} + \pi_{100})$. By adding (12.29) and (12.30), we have that

$$\lambda(\pi_{010} + \pi_{100}) = \lambda\pi_{000} + \mu_1\pi_{110} + \mu_2\pi_{110} - \mu_1\pi_{100} - \mu_2\pi_{010} \tag{12.33}$$

Substituting (12.28) changes (12.33) to

$$\lambda(\pi_{010} + \pi_{100}) = \mu_1\pi_{110} + \mu_2\pi_{110} \tag{12.34}$$

So...(12.32) becomes

$$(\mu_1 + \mu_2)\pi_{111} - \lambda\pi_{110} = 0 \tag{12.35}$$

By induction in (12.27), we have

$$(\mu_1 + \mu_2)\pi_{11,k+1} - \lambda\pi_{11k} = 0,\ k = 1, 2, ... \tag{12.36}$$

and

$$\pi_{11i} = \delta^i\pi_{110},\ i = 0, 1, 2, ... \tag{12.37}$$

where

$$\delta = \frac{\lambda}{\mu_1 + \mu_2} \tag{12.38}$$

$$\begin{aligned} 1 &= \sum_{i,j,k} \pi_{ijk} && (12.39)\\ &= \pi_{000} + \pi_{010} + \pi_{100} + \sum_{i=0}^{\infty} \pi_{11i} && (12.40)\\ &= \pi_{000} + \pi_{010} + \pi_{100} + \pi_{110} \sum_{i=0}^{\infty} \delta^i && (12.41)\\ &= \pi_{000} + \pi_{010} + \pi_{100} + \pi_{110} \cdot \frac{1}{1-\delta} && (12.42) \end{aligned}$$

Finding close-form expressions for the π_i is then straightforward.

12.4 Loss Models

One of the earliest queuing models was M/M/c/c: Markovian interarrival and service times, c servers and a buffer space of c jobs. Any job which arrives when c jobs are already in the system is lost. This was used by telephone companies to find the proportion of lost calls for a bank of c trunk lines.

12.4.1 Cell Communications Model

Let's consider a more modern example of this sort, involving cellular phone systems. (This is an extension of the example treated in K.S. Trivedi, *Probability and Statistics, with Reliability and Computer Science Applications* (second edition), Wiley, 2002, Sec. 8.2.3.2, which is in turn is based on two papers in the *IEEE Transactions on Vehicular Technology*.)

We consider one particular cell in the system. Mobile phone users drift in and out of the cell as they move around the city. A call can either be a **new call**, i.e. a call which someone has just dialed, or a **handoff call**, i.e. a call which had already been in progress in a neighboring cell but now has moved to this cell.

Each call in a cell needs a **channel**.[1] There are n channels available in the cell. We wish to give handoff calls priority over new calls.[2] This is accomplished as follows.

The system always reserves g channels for handoff calls. When a request for a new call (i.e. a non-handoff call) arrives, the system looks at X_t, the current number of calls in the cell. If that number is less than n-g, so that there are more than g idle channels available, the new call is accepted; otherwise it is rejected.

We assume that new calls originate from within the cells according to a Poisson process with rate λ_1, while handoff calls drift in from neighboring cells at rate λ_2. Meanwhile, call durations are exponential with rate μ_1, while the time that a call remains within the cell is exponential with rate μ_2.

12.4.1.1 Stationary Distribution

We again have a birth/death process, though a bit more complicated than our earlier ones. Let $\lambda = \lambda_1 + \lambda_2$ and $\mu = \mu_1 + \mu_2$. Then here is a sample balance equation, focused on transitions into (left-hand side in the equation) and out of (right-hand side) state 1:

$$\pi_0 \lambda + \pi_2 2\mu = \pi_1(\lambda + \mu) \tag{12.43}$$

Here's why: How can we enter state 1? Well, we could do so from state 0, where there are no calls; this

[1] This could be a certain frequency or a certain time slot position.

[2] We would rather give the caller of a new call a polite rejection message, e.g. "No lines available at this time, than suddenly terminate an existing conversation.

occurs if we get a new call (rate λ_1) or a handoff call (rate λ_2. In state 2, we enter state 1 if one of the two calls ends (rate μ_1) or one of the two calls leaves the cell (rate μ_2). The same kind of reasoning shows that we leave state 1 at rate $\lambda + \mu$.

As another example, here is the equation for state n-g:

$$\pi_{n-g}[\lambda_2 + (n-g)\mu] = \pi_{n-g+1} \cdot (n-g+1)\mu + \pi_{n-g-1}\lambda \tag{12.44}$$

Note the term λ_2 in (12.44), rather than λ as in (12.43).

Using our birth/death formula for the π_i, we find that

$$\pi_k = \begin{cases} \pi_0 \frac{A^k}{k!}, & \text{k} \leq \text{n-g} \\ \pi_0 \frac{A^{n-g}}{k!} A_1^{k-(n-g)}, & \text{k} \geq \text{n-g} \end{cases} \tag{12.45}$$

where $A = \lambda/\mu$, $A_1 = \lambda_2/\mu$ and

$$\pi_0 = \left[\sum_{k=0}^{n-g-1} \frac{A^k}{k!} + \sum_{k=n-g}^{n} \frac{A^{n-g}}{k!} A_1^{k-(n-g)} \right]^{-1} \tag{12.46}$$

12.4.1.2 Going Beyond Finding the π

One can calculate a number of interesting quantities from the π_i:

- The probability of a handoff call being rejected is π_n.
- The probability of a new call being dropped is

$$\sum_{k=n-g}^{n} \pi_k \tag{12.47}$$

- Since the per-channel utilization in state i is i/n, the overall long-run per-channel utilization is

$$\sum_{i=0}^{n} \pi_i \frac{i}{n} \tag{12.48}$$

- The long-run proportion of accepted calls which are handoff calls is the rate at which handoff calls are accepted, divided by the rate at which calls are accepted:

$$\frac{\lambda_2 \sum_{i=0}^{n-1} \pi_i}{\lambda_1 \sum_{i=0}^{n-g-1} \pi_i + \lambda_2 \sum_{i=0}^{n-1} \pi_i} \tag{12.49}$$

12.5 Nonexponential Service Times

The Markov property is of course crucial to the analyses we made above. Thus dropping the exponential assumption presents a major analytical challenge.

One queuing model which has been found tractable is M/G/1: Exponential interarrival times, general service times, one server. In fact, the mean queue length and related quantities can be obtained fairly easily, as follows.

Consider the residence time R for a tagged job. R is the time that our tagged job must first wait for completion of service of all jobs, if any, which are ahead of it—queued or now in service—plus the tagged job's own service time. Let $T_1, T_2, ...$ be i.i.d. with the distribution of a generic service time random variable S. T_1 represents the service time of the tagged job itself. $T_2, ..., T_N$ represent the service times of the queued jobs, if any.

Let N be the number of jobs in the system, either being served or queued; B be either 1 or 0, depending on whether the system is busy (i.e. N > 0) or not; and $S_{1,resid}$ be the remaining service time of the job currently being served, if any. Finally, we define, as before, $u = \frac{\lambda}{1/ES}$, the utilization. Note that that implies the EB = u.

Then the distribution of R is that of

$$BS_{1,resid} + (T_1 + ... + T_N) + (1 - B)T_1 \tag{12.50}$$

Note that if N = 0, then $T_1 + ... + T_N$ is considered to be 0, i.e. not present in (12.50).

Then

$$\begin{aligned} E(R) &= uE(S_{1,resid}) + E(T_1 + ... + T_N) + (1 - u)ET_1 && (12.51)\\ &= uE(S_{1,resid}) + E(N)E(S) + (1 - u)ES && (12.52)\\ &= uE(S_{1,resid}) + \lambda E(R)E(S) + (1 - u)ES && (12.53) \end{aligned}$$

The last equality is due to Little's Rule. Note also that we have made use of the PASTA property here, so that the distribution of N is the same at arrival times as general times.

Then

$$E(R) = \frac{uE(S_{1,resid})}{1-u} + ES \tag{12.54}$$

Note that the two terms here represent the mean residence time as the mean queuing time plus the mean service time.

So we must find $E(S_{1,resid})$. This is just the mean of the remaining-life distribution which we saw in Section 6.4 of our unit on renewal theory. Then

$$\begin{aligned} E(S_{1,resid}) &= \int_0^\infty t\frac{1-F_S(t)}{ES}\,dt & (12.55)\\ &= \frac{1}{ES}\int_0^\infty t\int_t^\infty f_S(u)\,du\,dt & (12.56)\\ &= \frac{1}{ES}\int_0^\infty f_S(u)\int_0^u t\,dt\,du & (12.57)\\ &= \frac{1}{2ES}E(S^2) & (12.58) \end{aligned}$$

So,

$$E(R) = \frac{uE(S^2)}{2ES(1-u)} + ES \tag{12.59}$$

What is remarkable about this famous formula is that E(R) depends not only on the mean service time but also on the variance. This result, which is not so intuitively obvious at first glance, shows the power of modeling. We might observe the dependency of E(R) on the variance of service time empirically if we do simulation, but here is a compact formula that shows it for us.

12.6 Reversed Markov Chains

We can get insight into some kinds of queuing systems by making use of the concepts of **reversed** Markov chains, which involve "playing the Markov chain backward," just as we could play a movie backward.

Consider a continuous-time, irreducible, positive recurrent Markov chain X(t).[3] For any fixed time τ (typ-

[3]Recall that a Markov chain is irreducible if it is possible to get from each state to each other state in a finite number of steps, and that the term *positive recurrent* means that the chain has a long-run state distribution π. Also, concerning our assumption here of continuous time, we should note that there are discrete-time analogs of the various points we'll make below.

ically thought of as large), define the **reversed** version of X(t) as Y(t) = X(τ-t), for $0 \leq t \leq \tau$. We will discuss a number of properties of reversed chains. These properties will enable what mathematicians call "soft analysis" of some Markov chains, especially those related to queues. This term refers to short, simple, elegant proofs or derivations.

12.6.1 Markov Property

The first property to note is that Y(t) is a Markov chain! Here is our first chance for soft analysis.

The "hard analysis" approach would be to start with the definition, which in continuous time would be that

$$P\left(Y(t) = k | Y(u), u \leq s\right) = P\left(Y(t) = k | Y(s)\right) \tag{12.60}$$

for all $0 < s < t$ and all k, using the fact that X(t) has the same property. That would involve making substitutions in Equation (12.60) like Y(t) = X(τ-t), etc.

But it is much easier to simply observe that the Markov property holds if and only if, conditional on the present, the past and the future are independent. Since that property holds for X(t), it also holds for Y(t) (with the roles of the "past" and the "future" interchanged).

12.6.2 Long-Run State Proportions

Clearly, if the long-run proportion of the time X(t) = k is π_i, the same long-run proportion will hold for Y(t). This of course only makes sense if you think of larger and larger τ.

12.6.3 Form of the Transition Rates of the Reversed Chain

Let $\tilde{\rho}_{ij}$ denote the number of transitions from state i to state j per unit time in the reversed chain. That number must be equal to the number of transitions from j to i in the original chain. Therefore,

$$\pi_i \tilde{\rho}_{ij} = \pi_j \rho_{ji} \tag{12.61}$$

This gives us a formula for the $\tilde{\rho}_{ij}$:

$$\tilde{\rho}_{ij} = \frac{\pi_j}{\pi_i} \rho_{ji} \tag{12.62}$$

12.6.4 Reversible Markov Chains

In some cases, the reversed chain has the same probabilistic structure as the original one! Note carefully what that would mean. In the continuous-time case, it would mean that $\tilde{\rho}_{ij} = \rho_{ij}$ for all i and j, where the $\tilde{\rho}_{ij}$ are the transition rates of Y(t).[4] If this is the case, we say that X(t) is **reversible**.

That is a very strong property. An example of a chain which is not reversible is the tree-search model in Section 11.5.5.[5] There the state space consists of all the nonnegative integers, and transitions were possible from states n to n+1 and from n to 0. Clearly this chain is not reversible, since we can go from n to 0 in one step but not vice versa.

12.6.4.1 Conditions for Checking Reversibility

Equation (12.61) shows that the original chain X(t) is reversible if and only if

$$\pi_i \rho_{ij} = \pi_j \rho_{ji} \tag{12.63}$$

for all i and j. These equations are called the **detailed balance equations**, as opposed to the general **balance equations**,

$$\sum_{j \neq i} \pi_j \rho_{ji} = \pi_i \lambda_i \tag{12.64}$$

which are used to find the π values. Recall that (12.64) arises from equating the flow into state i with the flow out of it. By contrast, Equation (12.63) equates the flow into i from a particular state j to the flow from i to j. Again, that is a much stronger condition, so we can see that most chains are not reversible. However, a number of important ones are reversible, as we'll see.

For example, consider birth/death chains. Here, the only cases in which ρ_{rs} is nonzero are those in which $|i - j| = 1$. Now, Equation (11.64) in our derivation of π for birth/death chains is exactly (12.63)! So we see that birth/death chains are reversible.

More generally, equations (12.63) may not be so easy to check, since for complex chains we may not be able to find closed-form expressions for the π values. Thus it is desirable to have another test available for reversibility. One such test is **Kolmogorov's Criterion**:

[4] Note that for a continuous-time Markov chain, the transition rates do indeed uniquely determined the probabilistic structure of the chain, not just the long-run state proportions. The short-run behavior of the chain is also determined by the transition rates, and at least in theory can be calculated by solving differential equations whose coefficients make use of those rates.

[5] That is a discrete-time example, but the principle here is the same.

The chain is reversible if and only if for any **loop** of states, the product of the transition rates is the same in both the forward and backward directions.

For example, consider the loop $i \to j \to k \to i$. Then we would check whether $\rho_{ij}\rho_{jk}\rho_{ki} = \rho_{ik}\rho_{kj}\rho_{ji}$.

Technically, we do have to check *all* loops. However, in many cases it should be clear that just a few loops are representative, as the other loops have the same structure.

Again consider birth/death chains. Kolmogorov's Criterion trivially shows that they are reversible, since any loop involves a path which is the same path when traversed in reverse.

12.6.4.2 Making New Reversible Chains from Old Ones

Since reversible chains are so useful (when we are lucky enough to have them), a very useful trick is to be able to form new reversible chains from old ones. The following two properties are very handy in that regard:

(a) Suppose U(t) and V(t) are reversible Markov chains, and define W(t) to be the tuple [U(t),V(t)]. Then W(t) is reversible.

(b) Suppose X(t) is a reversible Markov chain, and A is an irreducible subset of the state space of the chain, with long-run state distribution π. Define a chain W(t) with transition rates $\rho\prime_{ij}$ for $i \in A$, where $\rho\prime_{ij} = \rho_{ij}$ if $j \in A$ and $\rho\prime_{ij} = 0$ otherwise. Then W(t) is reversible, with long-run state distribution given by

$$\pi\prime_i = \frac{\pi_i}{\sum_{j \in A} \pi_j} \tag{12.65}$$

12.6.4.3 Example: Distribution of Residual Life

In Section 6.4.3, we used Markov chain methods to derive the age distribution at a fixed observation point in a renewal process. From remarks made there, we know that residual life has the same distribution. This could be proved similarly, at some effort, but it comes almost immediately from reversibility considerations. After all, the residual life in the reversed process is the age in the original process.

12.6.4.4 Example: Queues with a Common Waiting Area

Consider two M/M/1 queues, with chains G(t) and H(t), with independent arrival streams but having a common waiting area, with jobs arriving to a full waiting area simply being lost.[6]

First consider the case of an infinite waiting area. Let u_1 and u_2 be the utilizations of the two queues, as in (12.3). G(t) and H(t), being birth/death processes, are reversible. Then by property (a) above, the chain [G(t),H(t)] is also reversible. Long-run proportion of the time that there are m jobs in the first queue and n jobs in the second is

$$\pi_{mn} = (1-u_1)^m u_1 (1-u_2)^n u_2 \tag{12.66}$$

for m,n = 0,1,2,3,...

Now consider what would happen if these two queues were to have a common, finite waiting area. Denote the amount of space in the waiting area by w. The new process is the restriction of the original process to a subset of states A as in (b) above. (The set A will be precisely defined below.) It is easily verified from the Kolmogorov Criterion that the new process is also reversible.

Recall that the state m in the original queue U(t) is the number of jobs, including the one in service if any. That means the number of jobs waiting is $(m-1)^+$, where $x^+ = \max(x,0)$. That means that for our new system, with the common waiting area, we should take our subset A to be

$$\{(m,n) : m,n \geq 0, (m-1)^+ + (n-1)^+ \leq w\} \tag{12.67}$$

So, by property (b) above, we know that the long-run state distribution for the queue with the finite common waiting area is

$$\pi_{mn} = \frac{1}{a}(1-u_1)^m u_1 (1-u_2)^n u_2 \tag{12.68}$$

where

$$a = \sum_{(i,j) \in A} (1-u_1)^i u_1 (1-u_2)^j u_2 \tag{12.69}$$

In this example, reversibility was quite useful. It would have been essentially impossible to derive (12.68) algebraically. And even if intuition had suggested that solution as a guess, it would have been quite messy to verify the guess.

[6] Adapted from Ronald Wolff, *Stochastic Modeling and the Theory of Queues* Prentice Hall, 1989.

12.6.4.5 Closed-Form Expression for π for Any Reversible Markov Chain

(Adapted from Ronald Nelson, *Probability, Stochastic Processes and Queuing Theory*, Springer-Verlag, 1995.)

Recall that most Markov chains, especially those with infinite state spaces, do not have closed-form expressions for the steady-state probabilities. But we can always get such expressions for reversible chains, as follows.

Choose a fixed state s, and find paths from s to all other states. Denote the path to i by

$$s = j_{i1} \to j_{i2} \to ... \to j_{im_i} = i \tag{12.70}$$

Define

$$\psi_i = \begin{cases} 1, & \text{i = s} \\ \Pi_{k=1}^{m_i} r_{ik}, & \text{i} \neq \text{s} \end{cases} \tag{12.71}$$

where

$$r_{ik} = \frac{\rho(j_{ik}, j_{i,k+1})}{\rho(j_{i,k+1}, j_{i,k})} \tag{12.72}$$

Then the steady-state probabilities are

$$\pi_i = \frac{\psi_i}{\sum_k \psi_k} \tag{12.73}$$

You may notice that this looks similar to the derivation for birth/death processes, which as has been pointed out, are reversible.

12.7 Networks of Queues

12.7.1 Tandem Queues

Let's first consider an M/M/1 queue. As mentioned earlier, this is a birth/death process, thus reversible. This has an interesting and very useful application, as follows.

Think of the times at which jobs *depart* this system, i.e. the times at which jobs finish service. In the reversed process, these times are *arrivals*. Due to the reversibility, that means that the distribution of departure times is the same as that of arrival times. In other words:

- Departures from this system behave as a Poisson process with rate λ.

Also, let the initial state X(0) be distributed according to the steady-state probabilities π.[7] Due to the PASTA property of Poisson arrivals, the distribution of the system state at arrival times is the same as the distribution of the system state at nonrandom times t. Then by reversibility, we have that:

- The state distribution at departure times is the same as at nonrandom times.

And finally, noting as in Section 12.6.1 that, given X(t), the states $\{X(s), s \leq t\}$ of the queue before time t are statistically independent of the arrival process after time t, reversibility gives us that:

- Given t, the departure process before time t is statistically independent of the states $\{X(s), s \geq t\}$ of the queue after time t.

Let's apply that to **tandem** queues, which are queues acting in series. Suppose we have two such queues, with the first, $X_1(t)$ feeding its output to the second one, $X_2(t)$, as input. Suppose the input into $X_1(t)$ is a Poisson process with rate λ, and service times at both queues are exponentially distributed, with rates μ_1 and μ_2.

$X_1(t)$ is an M/M/1 queue, so its steady-state probabilities for $X_1(t)$ are given by Equation (12.3), with $u = \lambda/\mu_1$.

By the first bulleted item above, we know that the input into $X_2(t)$ is also Poisson. Therefore, $X_2(t)$ also is an M/M/1 queue, with steady-state probabilities as in Equation (12.3), with $u = \lambda/\mu_2$.

Now, what about the joint distribution of $[X_1(t), X_2(t)]$? The third bulleted item above says that the input to $X_2(t)$ up to time t is independent of $\{X_1(s), s \geq t\}$. So, using the fact that we are assuming that $X_1(0)$ has the steady-state distribution, we have that

$$P[X_1(t) = i, X_2(t) = j] = (1 - u_1)u_1^i P[X_2(t) = j] \tag{12.74}$$

Now letting $t \to \infty$, we get that the long-run probability of the vector $[X_1(t), X_2(t)]$ being equal to (i,j) is

$$(1 - u_1)u_1^j(1 - u_2)u_2^j \tag{12.75}$$

[7] Recall Section 11.1.2.4.

In other words, the steady-state distribution for the vector has the two components of the vector being independent.

Equation (12.75) is called a **product form solution** to the balance equations for steady-state probabilities.

By the way, the vector $[X_1(t), X_2(t)]$ is *not* reversible.

12.7.2 Jackson Networks

The tandem queues discussed in the last section comprise a special case of what are known as **Jackson networks**. Once again, there exists an enormous literature of Jackson and other kinds of queuing networks. The material can become very complicated (even the notation is very complex), and we will only present an introduction here. Our presentation is adapted from I. Mitrani, *Modelling of Computer and Communcation Systems*, Cambridge University Press, 1987.

Our network consists of N nodes, and jobs move from node to node. There is a queue at each node, and service time at node i is exponentially distributed with mean $1/\mu_i$.

12.7.2.1 Open Networks

Each job originally arrives externally to the network, with the arrival rate at node i being γ_i. After moving among various nodes, the job will eventually leave the network. Specifically, after a job completes service at node i, it moves to node j with probability q_{ij}, where

$$\sum_j q_{ij} < 1 \tag{12.76}$$

reflecting the fact that the job will leave the network altogether with probability $1 - \sum_j q_{ij}$.[8] It is assumed that the movement from node to node is memoryless.

As an example, you may wish to think of movement of packets among routers in a computer network, with the packets being jobs and the routers being nodes.

Let λ_i denote the total traffic rate into node i. By the usual equating of flow in and flow out, we have

$$\lambda_i = \gamma_i + \sum_{j=1}^{N} \lambda_j q_{ji} \tag{12.77}$$

[8] By the way, q_{ii} can be nonzero, allowing for feedback loops at nodes.

Note the in Equations (12.77), the knowns are γ_i and the q_{ji}. We can solve this system of linear equations for the unknowns, λ_i.

The utilization at node i is then $u_i = \lambda_i/\mu_i$, as before. Jackson's Theorem then says that in the long run, node i acts as an M/M/1 queue with that utilization, and that the nodes are independent in the long run:[9]

$$\lim_{t\to\infty} P[X_1(t) = i_1, ..., X_N(t) = i_N] = \Pi_{i=1}^{N}(1 - u_i)u^i \tag{12.78}$$

So, again we have a product form solution.

Let L_i denote the average number of jobs at node i. From Equation (12.6), we have $L_i = u_i/(1 - u_i)$. Thus the mean number of jobs in the system is

$$L = \sum_{i=1}^{N} \frac{u_i}{1 - u_i} \tag{12.79}$$

From this we can get the mean time that jobs stay in the network, W: From Little's Rule, $L = \gamma W$, so

$$W = \frac{1}{\gamma} \sum_{i=1}^{N} \frac{u_i}{1 - u_i} \tag{12.80}$$

where $\gamma = \gamma_1 + ... + \gamma_N$ is the total external arrival rate.

Jackson networks are not generally reversible. The reversed versions of Jackson networks are worth studying for other reasons, but we cannot pursue them here.

12.7.3 Closed Networks

In a closed Jackson network, we have for all i, $\gamma_i = 0$ and

$$\sum_j q_{ij} = 1 \tag{12.81}$$

In other words, jobs never enter or leave the network. There have been many models like this in the computer performance modeling literature. For instance, a model might consist of some nodes representing CPUs, some representing disk drives, and some representing users at terminals.

[9] We do not present the proof here, but it really is just a matter of showing that the distribution here satisfies the balance equations.

It turns out that we again get a product form solution.[10] The notation is more involved, so we will not present it here.

Exercises

1. Investigate the robustness of the M/M/1 queue model with respect to the assumption of exponential service times, as follows. Suppose the service time is actually uniformly distributed on (0,c), so that the mean service time would be c/2. Assume that arrivals do follow the exponential model, with mean interarrival time 1.0. Find the mean residence time, using (12.9), and compare it to the true value obtained from (12.59). Do this for various values of c, and graph the two curves using R.

2. Many mathematical analyses of queuing systems use **finite source** models. There are always a fixed number j of jobs in the system. A job queues up for the server, gets served in time S, then waits a random time W before queuing up for the server again.

A typical example would be a file server with j clients. The time W would be the time a client does work before it needs to access the file server again.

(a) Use Little's Rule, on two or more appropriately chosen boxes, to derive the following relation:

$$ER = \frac{jES}{U} - EW \tag{12.82}$$

where R is residence time (time spent in the queue plus service time) in one cycle for a job and U is the utilization fraction of the server.

(b) Set up a continuous time Markov chain, assuming exponential distributions for S and W, with state being the number of jobs currently at the server. Derive closed-form expressions for the π_i.

3. Consider the following variant of an M/M/1 queue: Each customer has a certain amount of patience, varying from one customer to another, exponentially distributed with rate η. When a customer's patience wears out while the customer is in the queue, he/she leaves (but not if his/her job is now in service). Arrival and service rates are λ and ν, respectively.

(a) Express the π_i in terms of λ, ν and η.

(b) Express the proportion of lost jobs as a function of the π_i, λ, ν and η.

4. A shop has two machines, with service time in machine i being exponentially distributed with rate μ_i, i = 1,2. Here $\mu_1 > \mu_2$. When a job reaches the head of the queue, it chooses machine 1 if that machine is

[10] This is confusing, since the different nodes are now not independent, due to the fact that the number of jobs in the overall system is constant.

idle, and otherwise waits for the first available machine. If when a job finishes on machine 1 there is a job in progress at machine 2, the latter job will be transferred to machine 1, getting priority over any queued jobs. Arrivals follow the usual Poisson process, parameter λ.

(a) Find the mean residence time.

(b) Find the proportion of jobs that are originally assigned to machine 2.

Appendix A

Review of Matrix Algebra

This book assumes the reader has had a course in linear algebra (or has self-studied it, always the better approach). This appendix is intended as a review of basic matrix algebra, or a quick treatment for those lacking this background.

A.1 Terminology and Notation

A **matrix** is a rectangular array of numbers. A **vector** is a matrix with only one row (a **row vector** or only one column (a **column vector**).

The expression, "the (i,j) element of a matrix," will mean its element in row i, column j.

Please note the following conventions:

- Capital letters, e.g. A and X, will be used to denote matrices and vectors.
- Lower-case letters with subscripts, e.g. $a_{2,15}$ and x_8, will be used to denote their elements.
- Capital letters with subscripts, e.g. A_{13}, will be used to denote submatrices and subvectors.

If A is a **square** matrix, i.e. one with equal numbers n of rows and columns, then its **diagonal** elements are a_{ii}, i = 1,...,n.

The **norm** (or **length**) of an n-element vector **X** is

$$\| X \| = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{A.1}$$

A.1.1 Matrix Addition and Multiplication

- For two matrices have the same numbers of rows and same numbers of columns, addition is defined elementwise, e.g.

$$\begin{pmatrix} 1 & 5 \\ 0 & 3 \\ 4 & 8 \end{pmatrix} + \begin{pmatrix} 6 & 2 \\ 0 & 1 \\ 4 & 0 \end{pmatrix} = \begin{pmatrix} 7 & 7 \\ 0 & 4 \\ 8 & 8 \end{pmatrix} \tag{A.2}$$

- Multiplication of a matrix by a **scalar**, i.e. a number, is also defined elementwise, e.g.

$$0.4 \begin{pmatrix} 7 & 7 \\ 0 & 4 \\ 8 & 8 \end{pmatrix} = \begin{pmatrix} 2.8 & 2.8 \\ 0 & 1.6 \\ 3.2 & 3.2 \end{pmatrix} \tag{A.3}$$

- The **inner product** or **dot product** of equal-length vectors X and Y is defined to be

$$\sum_{k=1}^{n} x_k y_k \tag{A.4}$$

- The product of matrices A and B is defined if the number of rows of B equals the number of columns of A (A and B are said to be **conformable**). In that case, the (i,j) element of the product C is defined to be

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \tag{A.5}$$

  For instance,

$$\begin{pmatrix} 7 & 6 \\ 0 & 4 \\ 8 & 8 \end{pmatrix} \begin{pmatrix} 1 & 6 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 19 & 66 \\ 8 & 16 \\ 24 & 80 \end{pmatrix} \tag{A.6}$$

  It is helpful to visualize c_{ij} as the inner product of row i of A and column j of B, e.g. as shown in bold face here:

$$\begin{pmatrix} \mathbf{7} & \mathbf{6} \\ 0 & 4 \\ 8 & 8 \end{pmatrix} \begin{pmatrix} \mathbf{1} & 6 \\ \mathbf{2} & 4 \end{pmatrix} = \begin{pmatrix} \mathbf{7} & 70 \\ 8 & 16 \\ 8 & 80 \end{pmatrix} \tag{A.7}$$

- Matrix multiplicatin is associative and distributive, but in general not commutative:

$$A(BC) = (AB)C \tag{A.8}$$

$$A(B + C) = AB + AC \tag{A.9}$$

$$AB \neq BA \tag{A.10}$$

A.2 Matrix Transpose

- The transpose of a matrix A, denoted A' or A^T, is obtained by exchanging the rows and columns of A, e.g.

$$\begin{pmatrix} 7 & 70 \\ 8 & 16 \\ 8 & 80 \end{pmatrix}' = \begin{pmatrix} 7 & 8 & 8 \\ 70 & 16 & 80 \end{pmatrix} \tag{A.11}$$

- If A + B is defined, then

$$(A + B)' = A' + B' \tag{A.12}$$

- If A and B are conformable, then

$$(AB)' = B'A' \tag{A.13}$$

A.3 Linear Independence

Equal-length vectors X_1,...,X_k are said to be **linearly independent** if it is impossible for

$$a_1X_1 + ... + a_kX_k = 0 \tag{A.14}$$

unless all the a_i are 0.

A.4 Determinants

Let A be an nxn matrix. The definition of the determinant of A, det(A), involves an abstract formula featuring permutations. It will be omitted here, in favor of the following computational method.

Let $A_{-(i,j)}$ denote the submatrix of A obtained by deleting its ith row and jth column. Then the determinant can be computed recursively across the kth row of A as

$$det(A) = \sum_{m=1}^{n} (-1)^{k+m} det(A_{-(k,m)}) \tag{A.15}$$

where

$$det \begin{pmatrix} s & t \\ u & v \end{pmatrix} = sv - tu \tag{A.16}$$

A.5 Matrix Inverse

- The **identity** matrix I of size n has 1s in all of its diagonal elements but 0s in all off-diagonal elements. It has the property that AI = A and IA = A whenever those products are defined.
- The A is a square matrix and AB = I, then B is said to be the **inverse** of A, denoted A^{-1}. Then BA = I will hold as well.
- A^{-1} exists if and only if its rows (or columns) are linearly independent.
- A^{-1} exists if and only if $det(A) \neq 0$.
- If A and B are square, conformable and invertible, then AB is also invertible, and

$$(AB)^{-1} = B^{-1}A^{-1} \tag{A.17}$$

A.6 Eigenvalues and Eigenvectors

Let A be a square matrix.[1]

[1]For nonsquare matrices, the discussion here would generalize to the topic of **singular value decomposition**.

- A scalar λ and a nonzero vector X that satisfy

$$AX = \lambda X \tag{A.18}$$

  are called an **eigenvalue** and **eigenvector** of A, respectively.

- A matrix U is said to be **orthogonal** if its rows have norm 1 and are orthogonal to each other, i.e. their inner product is 0. U thus has the property that $UU' = I$ i.e. $U^{-1} = U$.

- If A is symmetric and real, then it is **diagonalizable**, i.e there exists an orthogonal matrix U such that

$$U'AU = D \tag{A.19}$$

  for a diagonal matrix D. The elements of D are the eigenvalues of A, and the columns of U are the eigenvectors of A.

Appendix B

R Quick Start

Here we present a five-minute introduction to the R data/statistical programming language. Further learning resources are available at `http://heather.cs.ucdavis.edu/~/matloff/r.html`.

R syntax is similar to that of C. It is object-oriented (in the sense of encapsulation, polymorphism and everything being an object) and is a functional language (i.e. almost no side effects, every action is a function call, etc.).

B.1 Correspondences

aspect	C	R
assignment	=	<− (or =)
array terminology	array	vector (1-D), matrix (2-D), array (2-D+)
subscripts	start at 0	start at 1
array notation	m[2][3]	m[12,7]
storage	2-D arrays in row-major order	matrices in column-major order
mixed container	struct, members acessed by .	list, members acessed by $ or [[ ]]

B.2 Starting R

To invoke R, just type "R" into a terminal window. On a Windows machine, you probably have an R icon to click.

If you prefer to run from an IDE, the easiest one for a quick install is probably RStudio, `www.rstudio.org`.

R is (normally) interactive, with > as the prompt.

B.3 First Sample Programming Session

Below is a commented R session, to introduce the concepts. I had a text editor open in another window, constantly changing my code, then loading it via R's **source()** command.[1] The original contents of the file **odd.R** were:

```
1  oddcount <- function(x)  {
2     k <- 0  # assign 0 to k
3     for (n in x)  {
4        if (n %% 2 == 1) k <- k+1  # %% is the modulo operator
5     }
6     return(k)
7  }
```

By the way, we could have written that last statement as simply

```
1     k
```

because the last computed value of an R function is returned automatically.

The R session is shown below. You may wish to type it yourself as you go along, trying little experiments of your own along the way.[2]

```
1  > source("odd.R")  # load code from the given file
2  > ls()  # what objects do we have?
3  [1] "oddcount"
4  > # what kind of object is oddcount (well, we already know)?
5  > class(oddcount)
6  [1] "function"
7  > # while in interactive mode, can print any object by typing its name;
8  > # otherwise use print(), e.g. print(x+y)
9  > oddcount
10 function(x)  {
11    k <- 0  # assign 0 to k
12    for (n in x)  {
13       if (n %% 2 == 1) k <- k+1  # %% is the modulo operator
```

[1] I personally am not a big fan of using IDEs for my programming activities. If you use one, it probably has a button to click as an alternative to using **source()**.

[2] The source code for this file is at `http://heather.cs.ucdavis.edu/~matloff/MiscPLN/R5MinIntro.tex`.

```
14      }
15      return(k)
16  }
17  > # test it
18  > y <- c(5,12,13,8,88)  # c() is the concatenate function
19  > y
20  [1]  5 12 13  8 88
21  > oddcount(y)  # should report 2 odd numbers
22  [1] 2
23  > # change code to vectorize the count operation
24  > source("odd.R")
25  > oddcount
26  function(x)  {
27      x1 <- (x %% 2) == 1  # x1 now a vector of TRUEs and FALSEs
28      x2 <- x[x1]  # x2 now has the elements of x that were TRUE in x1
29      return(length(x2))
30  }
31  > # try subset of y, elements 2 through 3
32  > oddcount(y[2:3])
33  [1] 1
34  > # try subset of y, elements 2, 4 and 5
35  > oddcount(y[c(2,4,5)])
36  [1] 0
37  > # compactify the code
38  > source("odd.R")
39  > oddcount
40  function(x)  {
41      length(x[x %% 2 == 1])  # last value computed auto returned
42  }
43  > oddcount(y)
44  [1] 2
45  > # now have ftn return odd count AND the odd numbers themselves
46  > source("odd.R")
47  > oddcount
48  function(x)  {
49      x1 <- x[x %% 2 == 1]
50      return(list(odds=x1, numodds=length(x1)))
51  }
52  > # R's list type can contain any type; components delineated by $
53  > oddcount(y)
```

```
$odds
[1]  5 13

$numodds
[1] 2

> ocy <- oddcount(y)
> ocy
$odds
[1]  5 13

$numodds
[1] 2

> ocy$odds
[1]  5 13
> ocy[[1]]
[1]  5 13
> ocy[[2]]
[1] 2
```

Note that the R function **function()** produces functions! Thus assignment is used. For example, here is what **odd.R** looked like at the end of the above session:

```
oddcount <- function(x)  {
    x1 <- x[x %% 2 == 1]
    return(list(odds=x1, numodds=length(x1)))
}
```

We created some code, and then used function to create a function object, which we assigned to **oddcount**.

Note that we eventually **vectorized** our function **oddcount()**. This means taking advantage of the vector-based, functional language nature of R, exploiting R's built-in functions instead of loops. This changes the venue from interpreted R to C level, with a potentially large increase in speed. For example:

```
> x <- runif(1000000)  # 1000000 random numbers from the interval (0,1)
> system.time(sum(x))
   user  system elapsed
  0.008   0.000   0.006
> system.time({s <- 0; for (i in 1:1000000) s <- s + x[i]})
   user  system elapsed
  2.776   0.004   2.859
```

B.4 Second Sample Programming Session

A matrix is a special case of a vector, with added class attributes, the numbers of rows and columns.

```
> # "rowbind() function combines rows of matrices; there's a cbind() too
> m1 <- rbind(1:2,c(5,8))
> m1
     [,1] [,2]
[1,]    1    2
[2,]    5    8
> rbind(m1,c(6,-1))
     [,1] [,2]
[1,]    1    2
[2,]    5    8
[3,]    6   -1
> m2 <- matrix(1:6,nrow=2)
> m2
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> ncol(m2)
[1] 3
> nrow(m2)
[1] 2
> m2[2,3]
[1] 6
# get submatrix of m2, cols 2 and 3, any row
> m3 <- m2[,2:3]
> m3
     [,1] [,2]
[1,]    3    5
[2,]    4    6
> m1 * m3  # elementwise multiplication
     [,1] [,2]
[1,]    3   10
[2,]   20   48
> 2.5 * m3  # scalar multiplication (but see below)
     [,1] [,2]
[1,]  7.5 12.5
[2,] 10.0 15.0
> m1 %*% m3  # linear algebra matrix multiplication
```

```
38        [,1] [,2]
39 [1,]    11   17
40 [2,]    47   73
41 > # matrices are special cases of vectors, so can treat them as vectors
42 > sum(m1)
43 [1] 16
44 > ifelse(m2 %%3 == 1,0,m2) # (see below)
45        [,1] [,2] [,3]
46 [1,]    0    3    5
47 [2,]    2    0    6
```

The "scalar multiplication" above is not quite what you may think, even though the result may be. Here's why:

In R, scalars don't really exist; they are just one-element vectors. However, R usually uses **recycling**, i.e. replication, to make vector sizes match. In the example above in which we evaluated the express `2.5 * m3`, the number 2.5 was recycled to the matrix

$$\begin{pmatrix} 2.5 & 2.5 \\ 2.5 & 2.5 \end{pmatrix} \tag{B.1}$$

in order to conform with **m3** for (elementwise) multiplication.

The **ifelse()** function's call has the form

```
ifelse(boolean vectorexpression1, vectorexpression2, vectorexpression3)
```

All three vector expressions must be the same length, though R will lengthen some via recycling. The action will be to return a vector of the same length (and if matrices are involved, then the result also has the same shape). Each element of the result will be set to its corresponding element in **vectorexpression2** or **vectorexpression3**, depending on whether the corresponding element in **vectorexpression1** is TRUE or FALSE.

In our example above,

```
> ifelse(m2 %%3 == 1,0,m2) # (see below)
```

the expression `m2 %%3 == 1` evaluated to the boolean matrix

$$\begin{pmatrix} T & F & F \\ F & T & F \end{pmatrix} \tag{B.2}$$

(TRUE and FALSE may be abbreviated to T and F.)

The 0 was recycled to the matrix

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{B.3}$$

while **vectorexpression3**, **m2**, evaluated to itself.

B.5 Online Help

R's **help()** function, which can be invoked also with a question mark, gives short descriptions of the R functions. For example, typing

```
> ?rep
```

will give you a description of R's **rep()** function.

An especially nice feature of R is its **example()** function, which gives nice examples of whatever function you wish to query. For instance, typing

```
> example(wireframe())
```

will show examples—R code and resulting pictures—of **wireframe()**, one of R's 3-dimensional graphics functions.