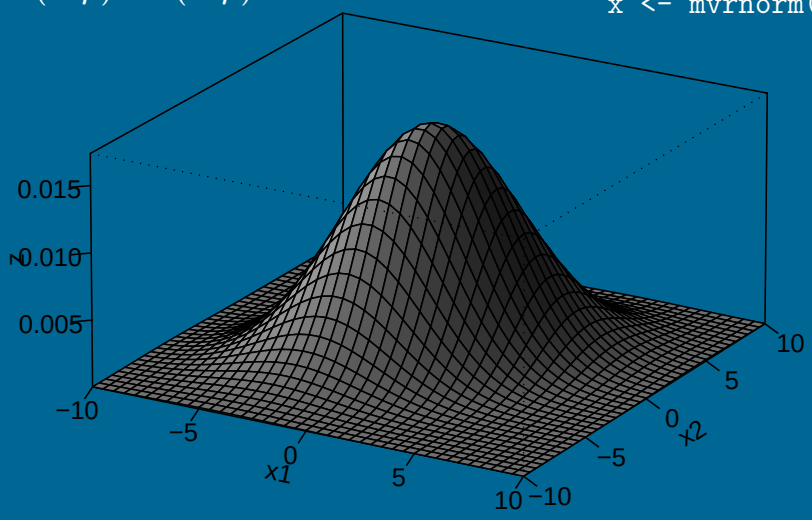
From Algorithms to Z-Scores:

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)
```



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, statistical computing, and regression methodology.

Prof. Matloff is a former appointed member of IFIP Working Group 11.3, an international committee concerned with database software security, established under 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 2011. 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")
```

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 10.)

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

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

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

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

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

Table 2.1: Sample Space for the Dice Example

Thus

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

2.2 The Crucial Notion of a Repeatable Experiment

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

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

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

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

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

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

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

2.3 Our Definitions

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

- We assume an "experiment" which is (at least in concept) repeatable. The experiment of rolling two dice is repeatable, and even the ALOHA experiment is so. (We simply watch the network for a long time, collecting data on pairs of consecutive epochs in which there are two active stations at the beginning.) On the other hand, the econometricians, in forecasting

2009, cannot "repeat" 2008. Yet all of the econometricians' tools assume that events in 2008 were affected by various sorts of randomness, and we think of repeating the experiment in a conceptual sense.

- We imagine performing the experiment a large number of times, recording the result of each repetition on a separate line in a notebook.
- We say A is an **event** for this experiment if it is a possible boolean (i.e. yes-or-no) outcome of the experiment. In the above example, here are some events:
  * X+Y = 6
  * X = 1
  * Y = 3
  * X-Y = 4
- A **random variable** is a numerical outcome of the experiment, such as X and Y here, as well as X+Y, 2XY and even sin(XY).
- For any event of interest A, imagine a column on A in the notebook. The k^{th} line in the notebook, k = 1,2,3,..., will say Yes or No, depending on whether A occurred or not during the k^{th} repetition of the experiment. For instance, we have such a column in our table above, for the event {A = blue+yellow = 6}.
- For any event of interest A, we define P(A) to be the long-run fraction of lines with Yes entries.
- For any events A, B, imagine a new column in our notebook, labeled "A and B." In each line, this column will say Yes if and only if there are Yes entries for both A and B. P(A and B) is then the long-run fraction of lines with Yes entries in the new column labeled "A and B."[1]
- 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.

[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.

Think of probabilities in this "notebook" context:

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

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

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

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

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

Here are some more very important definitions and properties:

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

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

  Again, this terminology *disjoint* stems from the set-theoretic sample space approach, where it means that A ∩ B = ϕ. That mathematical terminology works fine for our dice example,

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

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*

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

- 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).

Note too that (2.5) implies

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

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.

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

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

$$P(X_1 = 2) = p^2 + (1-p)^2 = 0.52 \tag{2.7}$$

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

$$
\begin{aligned}
P(X_1 = 2) &= P(\underbrace{C_1 \text{ and } C_2}\text{ or } \underbrace{\text{not } C_1 \text{ and not } C_2}) && (2.8)\\
&= P(C_1 \text{ and } C_2) + P(\text{ not } C_1 \text{ and not } C_2) \text{ (from (2.2))} && (2.9)\\
&= P(C_1)P(C_2) + P(\text{ not } C_1)P(\text{ not } C_2) \text{ (from (2.4))} && (2.10)\\
&= p^2 + (1-p)^2 && (2.11)
\end{aligned}
$$

(The underbraces in (2.8) 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:

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

(2.9): 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.

(2.10): 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.

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:

$$
\begin{aligned}
P(X_2 = 2) &= P(X_1 = 0 \text{ and } X_2 = 2 \text{ or } X_1 = 1 \text{ and } X_2 = 2 \text{ or } X_1 = 2 \text{ and } X_2 = 2)\\
&= P(X_1 = 0 \text{ and } X_2 = 2) && (2.12)\\
&+ P(X_1 = 1 \text{ and } X_2 = 2)\\
&+ P(X_1 = 2 \text{ and } X_2 = 2)
\end{aligned}
$$

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

$$
P(X_1 = 1 \text{ and } X_2 = 2) = P(X_1 = 1)P(X_2 = 2|X_1 = 1) \tag{2.13}
$$

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

$$P(X_1 = 1) = 2p(1-p) = 0.48 \tag{2.14}$$

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

$$P(X_2 = 2|X_1 = 1) = q[p^2 + (1-p)^2] = 0.41 \tag{2.15}$$

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 (2.6), we know that

$$P(X_1 = 1|X_2 = 2) = \frac{P(X_1 = 1 \textit{ and } X_2 = 2)}{P(X_2 = 2)} \tag{2.16}$$

We computed both numerator and denominator here before, in Equations (2.13) and (2.12), 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 (2.16), 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 (2.16) that

$$P(X_1 = 1|X_2 = 2) = 0.43 \tag{2.17}$$

yet from (2.14) we have

$$P(X_1 = 1) = 0.48. \tag{2.18}$$

2.6 Bayes' Rule

(This section should not be confused with Section 12.9. The latter is highly controversial, while the material in this section is not controversial at all.)

Following (2.16) above, we noted that the ingredients had already been computed, in (2.13) and (2.12). If we go back to the derivations in those two equations and substitute in (2.16), we have

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

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

$$= \frac{P(X_1 = 1)P(X_2 = 2|X_1 = 1)}{P(X_1 = 1)P(X_2 = 2|X_1 = 1) + P(X_1 = 2)P(X_2 = 2|X_1 = 2)} \tag{2.21}$$

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

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)} \tag{2.22}$$

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.

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

- 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.

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

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

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

2.8 Solution Strategies

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

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

$$P(X_1 = 2) = P(\underbrace{C_1 \text{ and } C_2}\ \text{ or }\ \underbrace{\text{not } C_1 \text{ and not } C_2}) \tag{2.23}$$

  above.
- Ask the famous question, "How can it happen?" Break big events down into small events; in the above case the event $X_1 = 2$ can happen if C_1 and C_2 or not C_1 and not C_2.
- 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 Equations (2.8)ff. After you gain more experience, you can start skipping steps, but not in the initial learning period.
- Solving probability problems—and even more so, building useful probability models—is like computer programming: It's a creative process.

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

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

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

Moreover, if you need P(B|A), there is no magic rule on how to find it. On the one hand, you might calculate it from (2.6), as we did in (2.16), but on the other hand you may be able to reason out the value of P(B|A), as we did following (2.14). Just try some cases until you find one that works, in the sense that you can evaluate both factors. It's the same as trying various programming ideas until you find one that works.

2.9 Example: Divisibility of Random Integers

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

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

$$
\begin{aligned}
P(N=2) &= P(\text{fail in step 5 and succeed in step 4}) \quad (\text{"How can it happen?"}) && (2.24)\\
&= P(\text{fail in step 5}) \quad P(\text{succeed in step 4} \mid \text{fail in step 5}) \quad ((2.5)) && (2.25)\\
&= [1-q(5)]q(4) && (2.26)\\
&= \frac{4}{5} \cdot \frac{1}{4} && (2.27)\\
&= \frac{1}{5} && (2.28)
\end{aligned}
$$

But there's more.

First, note that q(i) is either equal or approximately equal to 1/i. Then following the derivation in (2.24), you'll find that

$$P(N=j) \approx \frac{1}{5} \tag{2.29}$$

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.

2.10 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 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

$$
\begin{aligned}
P(\text{no complete circuit}) &= P(R+B \leq 7) && (2.30)\\
&= P(R \leq 6, R \neq 3, B = 0 \text{ or } R = 3, B \leq 4) && (2.31)\\
&= P(R \leq 6, R \neq 3, B = 0) + P(R = 3, B \leq 4) && (2.32)\\
&= P(R \leq 6, R \neq 3) + P(R = 3, B \leq 4) && (2.33)\\
&= P(R \leq 6, R \neq 3) + P(R = 3)\ P(B \leq 4) && (2.34)\\
&= \frac{5}{6} + \frac{1}{6} \cdot \frac{4}{6} && (2.35)\\
&= \frac{17}{18} && (2.36)
\end{aligned}
$$

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

$$
\begin{aligned}
P(B > 0 | R + B = 4) &= \frac{P(R + B = 4, B > 0)}{P(R + B = 4)} && (2.37)\\
&= \frac{P(R + B = 4, B > 0)}{P(R + B = 4, B > 0) + P(R + B = 4, B = 0)} && (2.38)\\
&= \frac{P(R = 3, B = 1)}{P(R = 3, B = 1) + P(R = 4)} && (2.39)\\
&= \frac{\frac{1}{6} \cdot \frac{1}{6}}{\frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6}} && (2.40)\\
&= \frac{1}{7} && (2.41)
\end{aligned}
$$

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

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

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

2.11 Example: Bus Ridership

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

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

$$\begin{aligned} P(L_2 = 0) &= P(B_1 = 0 \text{ and } L_2 = 0 \text{ or } B_1 = 1 \text{ and } L_2 = 0 \text{ or } B_1 = 2 \text{ and } L_2 = 0) && (2.42) \\ &= \sum_{i=0}^{2} P(B_1 = i \text{ and } L_2 = 0) && (2.43) \\ &= \sum_{i=0}^{2} P(B_1 = i) P(L_2 = 0 | B_1 = i) && (2.44) \\ &= 0.5^2 + (0.4)(0.2)(0.5) + (0.1)(0.2^2)(0.5) && (2.45) \\ &= 0.292 && (2.46) \end{aligned}$$

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 Appendix B.

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 Example: Rolling Dice

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

```
1  # roll d dice; find P(total = k)
2
3  # 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.

2.12.2 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.

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

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.

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]

[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.

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.

Even better:

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

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

The various improvements shown above compactify the code, and in many cases, make it much faster.[7] Note, though, that this comes at the expense of using more memory.

2.12.3 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.

```
1  # finds P(X1 = 2), P(X2 = 2) and P(X2 = 2|X1 = 1) in ALOHA example
2  sim <- function(p,q,nreps) {
3     countx2eq2 <- 0
4     countx1eq1 <- 0
5     countx1eq2 <- 0
6     countx2eq2givx1eq1 <- 0
7     # simulate nreps repetitions of the experiment
8     for (i in 1:nreps) {
9        numsend <- 0  # no messages sent so far
10       # simulate A and B's decision on whether to send in epoch 1
11       for (i in 1:2)
12          if (runif(1) < p) numsend <- numsend + 1
```

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

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

Note that each of the **nreps** iterations of the main **for** loop is analogous to one line in our hypothetical notebook. So, 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 (2.22) 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.

2.12.4 Example: Bus Ridership, cont'd.

Consider the example in Section 2.11. Let's find the probability that after visiting the tenth stop, the bus is empty. This is too complicated to solve analytically, but can easily be similated:

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

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

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

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

2.12.5 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 12.

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:

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

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

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

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

2.13.2 "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** 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

[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.

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

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:

Multinomial Coefficients: Suppose we have c objects in r categories, with c_i objects in category

[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.

i, i = 1,...,r. Then the number of ways to arrange them is

$$\frac{c!}{c_1!...c_r!}, \quad c_1 + ... + c_r = c \tag{2.52}$$

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 2.12.3.

(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 2.10, 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 2.11 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.

23. Consider the ALOHA simulation in Section 22.

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

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

Chapter 3

Discrete Random Variables

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

3.1 Random Variables

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

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

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

3.2 Discrete Random Variables

In our dice example, the random variable X could take on six values in the set {1,2,3,4,5,6}. 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 X and Y are said to be **independent** if for any sets I and J, the events {X is in I} and {Y is in J} are independent, i.e. P(X is in I and Y is in J) = P(X is in I) P(Y is in J).

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

3.4 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.

[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.

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

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 3.13.2 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

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

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,

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

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 8.3.1.

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 9.7 for details.)

As with expected values, the properties of variance discussed above, and also in Section 7.1.1 below, are key to the entire remainder of this book. You should notice immediately when you are in a setting in which they are applicable. For instance, if you see the variance of the sum of two random variables, you should instinctively think of (3.64) right away.

3.5.2 Central Importance of the Concept of Variance

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

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

3.5.3 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.

3.5.3.1 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.

3.5.3.2 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 i^{th} 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 8, 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 V 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,

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

By the way, (3.64) is actually the Pythagorean Theorem in a certain esoteric, infinite-dimesional vector space (related to a similar remark made earlier). This is pursued in Section 9.7 for the mathematically inclined.

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:

```
1  # finds E(X1), E(X2), Var(X2), Cov(X1,X2)
2  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 2.10. Below is simulation code to find the probability in (2.37):

```
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
```

```
9        } else bonus <- FALSE
10       if (position == 4) {
11          count4 <- count4 + 1
12          if (bonus) countbousngiven4 <- countbousngiven4 + 1
13       }
14    }
15    return(countbousngiven4/count4)
16 }
```

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.)

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

3.12.1 Example: Toss Coin Until First Head

In (3.68),

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

3.12.2 Example: Sum of Two Dice

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

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

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}.

3.12.3 Example: Watts-Strogatz Random Graph Model

Random graph models are used to analyze many types of link systems, such as power grids, social networks and even movie stars. The following is a variation on a famous model of that type, due to Duncan Watts and Steven Strogatz.

We have a graph of n nodes (e.g. each node is a person).[3] Think of them as being linked in a circle,

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

so we already have n links. One can thus reach any node in the graph from any other, by following the links of the circle. (We'll assume all links are bidirectional.)

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

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

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

This problem is similar in spirit to (though admittedly more difficult to think about than) kings-and-hearts example of Section 2.13.1. Other than the two neighboring links in the original circle, there aren-2 possible shortcut links to attach to our given node. We're interested in the probability that r-2 of them are chosen, and that k-(r-2) are chosen from the other possible links. Thus our probability is:

$$p_M(r) = \frac{\binom{n-2}{r-2}\binom{n^2/2-3n/2-(n-2)}{k-(r-2)}}{\binom{n^2/2-3n/2}{k}} = \frac{\binom{n-2}{r-2}\binom{n^2/2-5n/2+2}{k-(r-2)}}{\binom{n^2/2-3n/2}{k}} \tag{3.72}$$

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. In fact, we just had an example above, in Section 3.12.3. Since we get a different function p_M for each different value of k, that was a parametric family of pmfs, indexed by k.

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

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

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

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

$$p_N(k) = \left(1 - \frac{1}{6}\right)^{k-1} \frac{1}{6}, k = 1, 2, ... \tag{3.74}$$

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 events. We call the occurrence of the event **success** and the nonoccurrence **failure** (just convenient terms, not value judgments). The associated indicator random variable are denoted B_i, i = 1,2,3,... So B_i is 1 for success on the i^{th} trial, 0 for failure, with success probability p. For instance, p is 1/2 in the coin case, and 1/6 in the die example.

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

$$p_W(k) = (1 - p)^{k-1} p, k = 1, 2, ... \tag{3.75}$$

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.[4]

It should make good intuitive sense to you that

$$E(W) = \frac{1}{p} \tag{3.76}$$

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

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 $r \leq s$,

$$\sum_{i=r}^{s} t^i = t^r \frac{1 - t^{s-r+1}}{1 - t} \tag{3.77}$$

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 $|t| < 1$,

$$\sum_{i=0}^{\infty} t^i = \frac{1}{1 - t} \tag{3.78}$$

To prove this, just take r = 0 and let $s \to \infty$ in (3.77).

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

$$\sum_{i=1}^{\infty} i t^{i-1} = \frac{1}{(1 - t)^2} \tag{3.79}$$

This is derived by applying $\frac{d}{dt}$ to (3.78).[5]

Deriving (3.76) is then easy, using (3.79):

$$\begin{aligned} EW &= \sum_{i=1}^{\infty} i(1-p)^{i-1} p && (3.80) \\ &= p \sum_{i=1}^{\infty} i(1-p)^{i-1} && (3.81) \\ &= p \cdot \frac{1}{[1-(1-p)]^2} && (3.82) \\ &= \frac{1}{p} && (3.83) \end{aligned}$$

[5]To be more carefully, we should differentiate (3.77) and take limits.

Using similar computations, one can show that

$$Var(W) = \frac{1-p}{p^2} \tag{3.84}$$

We can also find a closed-form expression for the quantities $P(W \leq m)$, m = 1,2,... (This has a formal name, as will be seen later in Section 4.2.) For any positive integer m we have

$$\begin{aligned} F_W(m) &= P(W \leq m) && (3.85)\\ &= 1 - P(W > m) && (3.86)\\ &= 1 - P(\text{the first m trials are all failures}) && (3.87)\\ &= 1 - (1-p)^m && (3.88) \end{aligned}$$

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

3.13.1.1 R Functions

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

For example, if you run

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

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

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

- **dgeom(i,p)**, to find $P(X = i)$
- **pgeom(i,p)**, to find $P(X \leq i)$

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

Important note: Some books define geometric distributions slightly differently, as the number of failures before the first success, rather than the number of trials to the first success. Thus for example in calling **dgeom()**, subtract 1 from the value used in our definition.

3.13.1.2 Example: a Parking Space Problem

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

To solve this problem, you might at first think that D follows a geometric distribution, but 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

$$D = \begin{cases} 11 - N, & N \leq 10 \\ N - 11, & N > 10 \end{cases} \tag{3.89}$$

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

$$ED = \sum_{i=1}^{10} (11 - i) 0.85^{i-1} 0.15 + \sum_{i=11}^{\infty} (i - 11) 0.85^{i-1} 0.15 \tag{3.90}$$

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

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

```
1 parksim <- function(nreps) {
2    # do the experiment nreps times, recording the values of N
3    nvals <- rgeom(nreps,0.2) + 1
4    # now find the values of D
5    dvals <- ifelse(nvals <= 10,11-nvals,nvals-11)
6    # return ED
7    return(mean(dvals))
8 }
```

Note the vectorized addition and recycling (Section 2.12.2 in the line

```
nvals <- rgeom(nreps,0.2) + 1
```

The call to **ifelse()** is another instance of R's vectorization, a vectorized if-then-else. The first argument evaluates to a vector of TRUE and FALSE values. For each TRUE, the corresponding element of **dvals** will be set to the corresponding element of the vector **11-nvals** (again involving vectorized addition and recycling), and for each false, the element of **dvals** will be set to the element of **nvals-11**.

3.13.2 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).[6]

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

$$P(X=2) = \binom{5}{2} 0.5^2(1-0.5)^3 = \binom{5}{2}/32 = 5/16 \tag{3.91}$$

For general n and p,

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.92}$$

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:

$$X = \sum_{i=1}^{n} B_i \tag{3.93}$$

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

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

$$EB_i = p \tag{3.94}$$

and

$$Var(B_i) = p(1-p) \tag{3.95}$$

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:

$$EX = E(B_1 + ..., +B_n) = EB_1 + ... + EB_n = np \tag{3.96}$$

and from (3.64),

$$Var(X) = Var(B_1 + ..., +B_n) = Var(B_1) + ... + Var(B_n) = np(1-p) \tag{3.97}$$

Again, (3.96) should make good intuitive sense to you.

3.13.2.1 R Functions

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

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

3.13.2.2 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

$$\begin{aligned} p_X(m) &= P(X = m) && (3.98)\\ &= \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} P(X = m \text{ and } Y = i) && (3.99)\\ &= \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} P(X = m|Y = i)\ P(Y = i) && (3.100)\\ &= \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} \binom{i}{m} 0.5^i \binom{k}{i} 0.5^k && (3.101)\\ &= 0.5^k \sum_{i=\text{ceil}(m/2)}^{\min(m,k)} \frac{k!}{m!(i-m)!(k-i)!} 0.5^i && (3.102) \end{aligned}$$

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

3.13.2.3 Example: Analysis of Social Networks

Let's continue our earlier discussion from Section 3.12.3.

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.

Recall the notion of degree distribution from Section 3.12.3. Clearly the degree distribution here 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

$$k\binom{n-k}{2} + \binom{k}{2} \tag{3.103}$$

trials and success probability p.

3.13.3 The Poisson Family of Distributions

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

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, ... \tag{3.104}$$

It turns out that

$$EX = \lambda \tag{3.105}$$

$$Var(X) = \lambda \tag{3.106}$$

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:

$$e^t = \sum_{i=0}^{\infty} \frac{t^i}{i!} \tag{3.107}$$

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 4.4.4.5.

3.13.3.1 R Functions

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

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

3.13.4 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:

$$\{N = 5\} = \{\text{2 heads in the first 4 tosses and head on the } 5^{th} \text{ toss}\} \tag{3.108}$$

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

$$P(\text{2 heads in the first 4 tosses}) = \binom{4}{2}\left(\frac{1}{2}\right)^4 \tag{3.109}$$

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

$$P(N = 5) = \binom{4}{2}\left(\frac{1}{2}\right)^5 = \frac{3}{16} \tag{3.110}$$

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

$$P(N = k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r, k = r, r+1, ... \tag{3.111}$$

We can write

$$N = G_1 + ... + G_r \tag{3.112}$$

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

$$E(N) = r \cdot \frac{1}{p} \tag{3.113}$$

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

$$Var(N) = r \cdot \frac{1-p}{p^2} \tag{3.114}$$

3.13.5 The Power Law Family of Distributions

Here

$$p_X(k) = ck^{-\gamma},\ k = 1, 2, 3, ... \tag{3.115}$$

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:

$$1.0 = \sum_{k=1}^{\infty} ck^{-\gamma} \approx c \int_1^{\infty} k^{-\gamma}\, dk = c/(\gamma - 1) \tag{3.116}$$

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

$$p_X(k) = ck^{-\gamma}q^k \tag{3.117}$$

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

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

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

write it mathematically:

$$P(Y_6 > X_6) = \sum_{i=1}^{4} \sum_{j=0}^{i-1} P(Y_6 = i \text{ and } X_6 = j) \tag{3.118}$$

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,

$$P(Y_6 = i \text{ and } X_6 = j) = P(Y_6 = i) \cdot P(X_6 = j) \tag{3.119}$$

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

$$P(Y_6 = i) = \binom{4}{i} 0.5^i (1 - 0.5)^{4-i} \tag{3.120}$$

and

$$P(X_6 = j) = \binom{6}{j} 0.5^j (1 - 0.5)^{6-j} \tag{3.121}$$

We would then substitute (3.120) and (3.121) in (3.118). 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

$$p = \binom{4}{2} 0.5^4 = \frac{3}{8} \tag{3.122}$$

Thus using the fact that N has a geometric distribution,

$$P(N = 5) = (1 - p)^4 p = 0.057 \tag{3.123}$$

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.[7] For instance, the probability that there is a successful transmission is equal to the probability that exactly one of the m nodes attempts to send,

$$\binom{m}{1} p(1 - p)^{m-1} = mp(1 - p)^{m-1} \tag{3.124}$$

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 (3.124).

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.

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

3.15 A Preview of Markov Chains

Here we introduce Markov chains, a topic covered in much more detail in Chapter 16. 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
>
> $$P(X_6 = j | X_1 = i_1, X_2 = i_2, X_3 = i_3, X_4 = i_4, X_5 = i) \tag{3.125}$$
>
> does not depend on $i_m, m = 1, ..., 4$. Thus we can write (3.125) 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:

$$P = \begin{pmatrix} (1-q)^2 + 2q(1-q)p & 2q(1-q)(1-p) + 2q^2p(1-p) & q^2[p^2 + (1-p)^2] \\ (1-q)p & 2qp(1-p) + (1-q)(1-p) & q[p^2 + (1-p)^2] \\ 0 & 2p(1-p) & p^2 + (1-p)^2 \end{pmatrix} \tag{3.126}$$

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 16.1.2, in typical applications

$$\pi_i = \lim_{t \to \infty} \frac{N_{it}}{t} \tag{3.127}$$

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

$$\lim_{t \to \infty} P(X_t = i) = \pi_i \tag{3.128}$$

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

In Chapter 16 it is shown that the π_i are easy to find (in the case of finite state spaces, the subject of this section here), by solving the matrix equation

$$(I - P')\pi = 0 \tag{3.129}$$

subject to the constraint

$$\sum_i \pi_i = 1 \tag{3.130}$$

Here I is the identity matrix, and ' denotes matrix transpose. R code to do all this (after some algebraic manipulations), **findpi1()**, is provided in Section 16.1.2.2, reproduced here for convenience:

```
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 }
```

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

$$0 \cdot 0.47 + 1 \cdot 0.43 + 2 \cdot 0.10 = 0.63 \tag{3.131}$$

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

$$\pi = (0.5714286, 0.2857143, 0.1428571) \tag{3.132}$$

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

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

$$p_{00} = 0.5 \tag{3.133}$$

$$p_{01} = 0.4 \tag{3.134}$$

$$p_{20} = (0.2)^2(0.5) = 0.02 \tag{3.135}$$

$$p_{20,20} = (0.8)^2 0 + 20 \cdot (0.2)(0.8)^{19}(0.4) + 190 \cdot (0.2)^2 (0.8^{18}(0.1) \tag{3.136}$$

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

$$\sum_{i=0}^{20} \pi_i i \tag{3.137}$$

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

$$1 \cdot [\pi_{19}(0.1) + \pi_{20}(0.4)] + 2 \cdot [\pi_{20}(0.1)] \tag{3.138}$$

3.15.4 An Inventory Model

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

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

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

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

Let's write a function **inventory(p,q,r)** that returns the π vector for this Markov chain. It will call **findpi1()**, similarly to the two code snippets on page **??**.

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

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

$$P(A_1 \text{ and } A_2) = P(A_1)P(A_2) \tag{3.139}$$

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

$$P(A_1) = P(C_1)P(A_1|C_1) + P(C_2)P(A_1|C_2) = (0.5)(0.9) + (0.5)(0.1) = 0.5 \tag{3.140}$$

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 (3.139) is equal to 0.25. What about the left hand side?

$$
\begin{aligned}
P(A_1 \text{ and } A_2) &= P(A_1 \text{ and } A_2 \text{ and } C_1) + P(A_1 \text{ and } A_2 \text{ and } C_2) && (3.141)\\
&= P(A_1 \text{ and } A_2|C_1)P(C_1) + P(A_1 \text{ and} A_2|C_2)P(C_2) && (3.142)\\
&= (0.9)^2(0.5) + (0.1)^2(0.5) && (3.143)\\
&= 0.41 && (3.144)
\end{aligned}
$$

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,

$$P(Y \geq d) \leq \frac{EY}{d} \tag{3.145}$$

To prove (3.145), let Z be the indicator random variable for the event $Y \geq d$ (Section 3.6).

Now note that

$$Y \geq dZ \tag{3.146}$$

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

So

$$EY \geq dEZ \tag{3.147}$$

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

(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 (3.147) is $dP(Y \geq d)$, so (3.145) follows.

Now to prove (3.41), define

$$Y = (X - \mu)^2 \tag{3.148}$$

and set $d = c^2\sigma^2$. Then (3.145) says

$$P[(X - \mu)^2 \geq c^2\sigma^2] \leq \frac{E[(X - \mu)^2]}{c^2\sigma^2} \tag{3.149}$$

Since

$$(X - \mu)^2 \geq c^2\sigma^2 \text{ if and only if } |X - \mu| \geq c\sigma \tag{3.150}$$

the left-hand side of (3.149) is the same as the left-hand side of (3.41). The numerator of the right-hand size of (3.149) 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*

$$P(\cup_i A_i) = \sum_i P(A_i) \tag{3.151}$$

A **random variable** *is any function $X : S \to \mathcal{R}$.*[8]

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*

$$\lim_{n \to \infty} \frac{U_1 + ... + U_n}{n} = E(U) \text{ with probability 1} \tag{3.152}$$

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.

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

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

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

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

4. Derive (3.84).

5. Finish the computation in (3.90).

6. Derive the facts that for a Poisson-distributed random variable X with parameter λ, $EX = Var(X) = \lambda$. Use the hints in Section 3.13.3.

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

$$EX = \sum_{i=1}^{\infty} P(X \geq i) \tag{3.153}$$

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

$$\lim_{n \to \infty} P(X = n/2) = 0 \tag{3.154}$$

Use Stirling's approximation,

$$k! \approx \sqrt{2\pi k} \left(\frac{k}{e} \right)^k \tag{3.155}$$

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 30.

(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 3.13.1.2. 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 3.13.3, 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.

29. Supppose we have a nonnegative random variable X, and define a new random variable Y, which is equal to X if X > 8 and equal to 0 otherwise. Assume X takes on only a finite number of values (just a mathematical nicety, not really an issue). Which one of the following is true:

(i) $EY \leq EX$.

(ii) $EY \geq EX$.

(iii) Either of EY and EX could be larger than the other, depending on the situation.

(iv) EY is undefined.

30. Say we roll two dice, a blue one and a yellow one. Let B and Y denote the number of dots we get, respectively. Now let G denote the indicator random variable for the event S = 2. Find E(G).

31. Suppose I_1, I_2 and I_3 are independent indicator random variables, with $P(I_j = 1) = p_j$, j = 1,2,3. Find the following in terms of the p_j, writing your derivation with reasons in the form of mailing tube numbers.

32. Consider the ALOHA example, Section 3.14.3 . Write a call to the built-in R function **dbinom()** to evaluate (3.124) for general m and p.

33. Consider the bus ridership example, Section 2.11. Suppose upon arrival to a certain stop, there are 2 passengers. Let A denote the number of them who choose to alight at that stop.

(a) State the parametric family that the distribution of A belongs to.

(b) Find $p_A(1)$ and $F_A(1)$, writing each answer in decimal expression form e.g. $12^8 \cdot 0.32 + 0.3333$.

34. Suppose you have a large disk farm, so heavily used that the lifetimes L are measured in months. They come from two different factories, in proportions q and 1-q. The disks from factory i have geometrically distributed lifetime with parameter p_i, i = 1,2. Find Var(L) in terms of q and the p_i.

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 (4.2) 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 for t = 0.23,

$$F_D(0.23) = P(D \leq 0.23) = P(0 \leq D \leq 0.23) = 0.23 \tag{4.5}$$

Also,

$$F_D(-10.23) = P(D \leq -10.23) = 0 \tag{4.6}$$

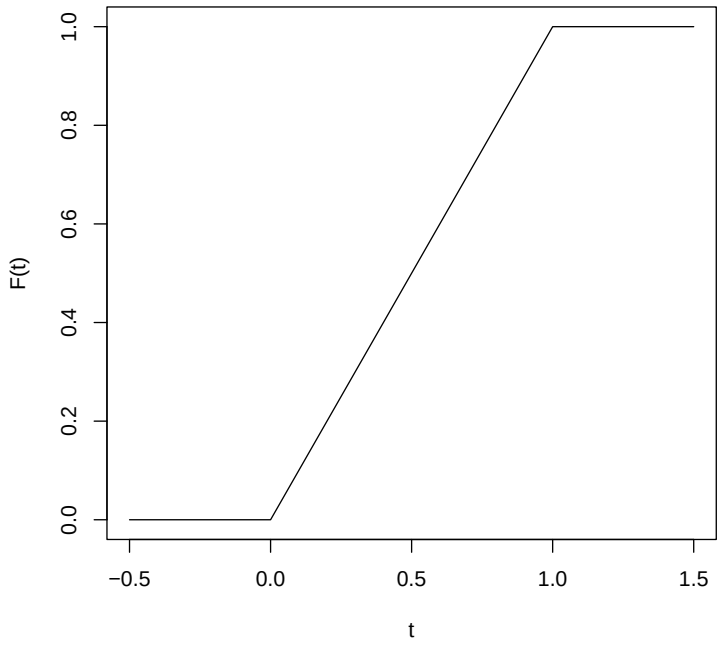
and

$$F_D(10.23) = P(D \leq 10.23) = 1 \tag{4.7}$$

In general for our dart,

$$F_D(t) = \begin{cases} 0, & \text{if } t \leq 0 \\ t, & \text{if } 0 < t < 1 \\ 1, & \text{if } t \geq 1 \end{cases} \tag{4.8}$$

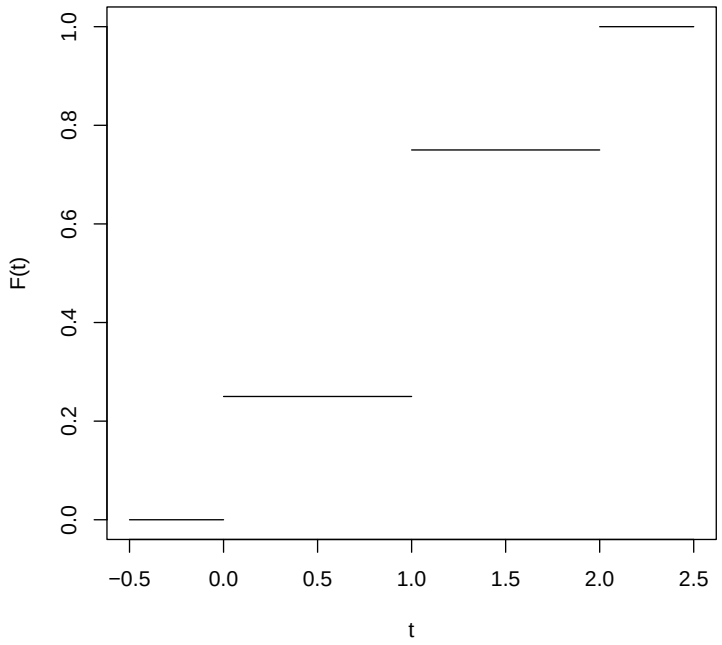
Here is the graph of F_D:



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

$$F_Z(t) = \begin{cases} 0, & \text{if } t < 0 \\ 0.25, & \text{if } 0 \leq t < 1 \\ 0.75, & \text{if } 1 \leq t < 2 \\ 1, & \text{if } t \geq 2 \end{cases} \tag{4.9}$$

For instance, $F_Z(1.2) = P(Z \leq 1.2) = P(Z = 0 \text{ or } Z = 1) = 0.25 + 0.50 = 0.75$. (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 (4.8) is continuous while F_Z in (4.9) has jumps. For this reason, we call random variables like D—ones which have 0 probability for individual points—**continuous random variables**.

Students sometimes ask, "What is t?" The answer is that it's simply the argument of a mathematical function, just like the role of t in, say, $g(t) = sin(\pi t), -\infty < t < \infty$. $F_Z()$ is a function, just like this g(t) or the numerous functions that you worked with in calculus. Each input yields an ouput; the input 1.2 yields the output 0.75 in the case of $F_Z()$ while the input 1 yields the output 0 in the case of g(t).

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*

$$f_W(t) = \frac{d}{dt} F_W(t), -\infty < t < \infty \tag{4.10}$$

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:

$$\begin{aligned}
2(0.1) f_X(1.3) &\approx \int_{1.2}^{1.4} f_X(t)\, dt \quad \text{(rect. approx. to slice of area)} && (4.11)\\
&= F_X(1.4) - F_X(1.2) \quad (f_X = F_X') && (4.12)\\
&= P(1.2 < X \leq 1.4) \quad \text{(def. of } F_X\text{)} && (4.13)\\
&= P(1.2 < X < 1.4) \quad \text{(prob. of single pt. is 0)} && (4.14)
\end{aligned}$$

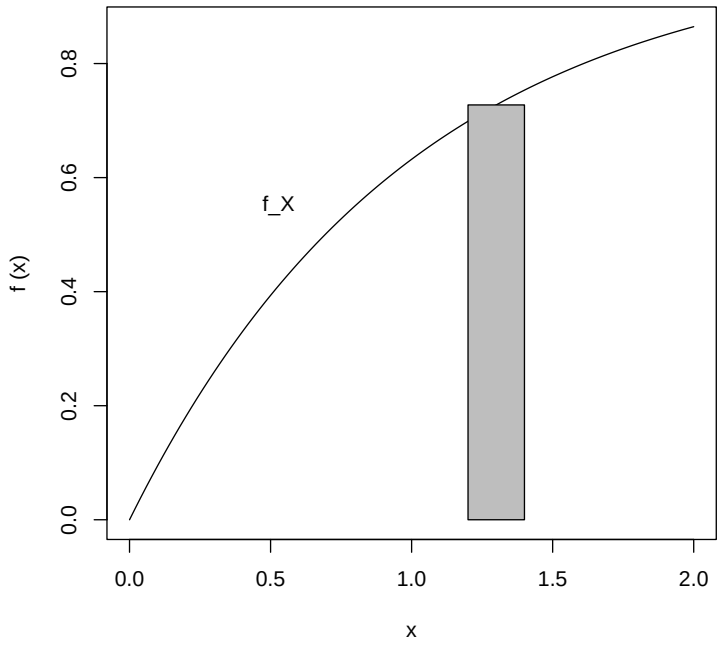


Figure 4.1: Approximation of Probability by a Rectangle

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

$$2cf_X(t) \approx P(t - c < X < t + c) \tag{4.15}$$

Thus we have:

Interpretation of Density Functions

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

$$\frac{P(r - c < X < r + c)}{P(s - c < X < s + c)} \approx \frac{f_X(r)}{f_X(s)} \tag{4.16}$$

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.)

4.3.2 Properties of Densities

Equation (4.10) implies

Property A:

$$P(a < W \leq b) = F_W(b) - F_W(a) = \int_a^b f_W(t)\ dt \tag{4.17}$$

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

Property B:

$$P(a < W \leq b) = P(a \leq W \leq b) = P(a \leq W < b) = P(a < W < b) = \int_a^b f_W(t)\ dt \tag{4.18}$$

This in turn implies:

Property C:

$$\int_{-\infty}^{\infty} f_W(t)\ dt = 1 \tag{4.19}$$

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

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

$$E(W) = \sum_c c p_W(c) \tag{4.20}$$

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:

$$E(W) = \int_t t f_W(t)\ dtE \tag{4.21}$$

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

$$E(W) = \int_{-\infty}^{\infty} f_W(t)\ dtE \tag{4.22}$$

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

And of course,

$$E(W^2) = \int_t t^2 f_W(t)\ dt \tag{4.23}$$

and in general, similarly to (3.24):

Property E:

$$E[g(W)] = \int_t g(t) f_W(t)\ dt \tag{4.24}$$

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), (4.4.2.1) 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:

$$EX = \int_1^4 t \cdot 2t/15 \; dt = 2.8 \tag{4.25}$$

$$P(X > 2.5) = \int_{2.5}^4 2t/15 \; dt = 0.65 \tag{4.26}$$

$$F_X(s) = \int_1^s 2t/15 \; dt = \frac{s^2 - 1}{15} \quad \text{for s in (1,4) (cdf is 0 for t} < 1\text{, and 1 for t} > 4) \tag{4.27}$$

$$\begin{aligned} Var(X) &= E(X^2) - (EX)^2 \quad \text{(from (3.29)} && (4.28)\\ &= \int_1^4 t^2 2t/15 \; dt - 2.8^2 \quad \text{(from (4.25))} && (4.29)\\ &= 5.7 && (4.30) \end{aligned}$$

$$\begin{aligned} P(\text{tenths digit of X is even}) &= \sum_{i=0}^{28} P\left[1 + i/10 < X < 1 + (i+1)/10\right] && (4.31)\\ &= \sum_{i=0}^{28} \int_{1+i/10}^{1+(i+1)/10} 2t/15 \; dt && (4.32)\\ &= \text{... (integration left to the reader)} && (4.33) \end{aligned}$$

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

Proportion of bulbs with lifetime less than the mean lifetime:

$$P(L < 2.8) = \int_1^{2.8} 2t/15 \; dt = (2.8^2 - 1)/15 \tag{4.34}$$

Mean of 1/L:

$$E(1/L) = \int_1^4 \frac{1}{t} \cdot 2t/15 \; dt = \frac{2}{5} \tag{4.35}$$

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

Use (3.111) with k = 2 and p = 0.65.

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

$$f_D(t) = \frac{1}{r-q} \tag{4.36}$$

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).

4.4.1.2 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

4.4.1.3 Example: Modeling of Disk Performance

Uniform distributions are often used to model computer disk requests. Recall that a disk consists of a large number of concentric rings, called **tracks**. When a program issues a request to read or

write a file, the **read/write head** must be positioned above the track of the first part of the file. This move, which is called a **seek**, can be a significant factor in disk performance in large systems, e.g. a database for a bank.

If the number of tracks is large, the position of the read/write head, which I'll denote 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**.

4.4.1.4 Example: Modeling of Denial-of-Service Attack

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

4.4.2 The Normal (Gaussian) Family of Continuous Distributions

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

4.4.2.1 Density and Properties

Density and Parameters:

The density for a normal distribution is

$$f_W(t) = \frac{1}{\sqrt{2\pi}\sigma}\ e^{-0.5\left(\frac{t-\mu}{\sigma}\right)^2}, -\infty < t < \infty \tag{4.37}$$

[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.

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 (4.4.2.1) *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 (4.37), now with the new mean and variance.

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

$$\begin{aligned} F_Y(t) &= P(Y \leq t) \text{ (definition of } F_Y) && (4.38)\\ &= P(cX + d \leq t) \text{ (definition of Y)} && (4.39)\\ &= P\left(X \leq \frac{t-d}{c}\right) \text{ (algebra)} && (4.40)\\ &= F_X\left(\frac{t-d}{c}\right) \text{ (definition of } F_X) && (4.41) \end{aligned}$$

Therefore

$$\begin{aligned} f_Y(t) &= \frac{d}{dt}F_Y(t) \text{ (definition of } f_Y) && (4.42)\\ &= \frac{d}{dt}F_X\left(\frac{t-d}{c}\right) \text{ (from (4.41))} && (4.43)\\ &= f_X\left(\frac{t-d}{c}\right) \cdot \frac{d}{dt}\frac{t-d}{c} \text{ (definition of } f_X \text{ and the Chain Rule)} && (4.44)\\ &= \frac{1}{c} \cdot \frac{1}{\sqrt{2\pi}\sigma}\, e^{-0.5\left(\frac{\frac{t-d}{c}-\mu}{\sigma}\right)^2} \text{ (from (4.37)} && (4.45)\\ &= \frac{1}{\sqrt{2\pi}(c\sigma)}\, e^{-0.5\left(\frac{t-(c\mu+d)}{c\sigma}\right)^2} \text{ (algebra)} && (4.46) \end{aligned}$$

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

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

Closure Under Independent Summation

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

This is a pretty remarkable phenomenon, not true for most other parametric families. If for instance X and Y each with, say, a U(0,1) distribution, then the density of S turns out to be triangle-shaped, NOT another uniform distribution. (This can be derived using the methods of Section 8.3.2.)

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

Evaluating the Normal cdf

The function in (4.37) 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

$$\frac{X - \mu}{\sigma} \tag{4.47}$$

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.

4.4.2.2 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

$$P(X \geq 535) = P\left(Z \geq \frac{535 - 500}{15}\right) = 1 - \Phi(35/15) = 0.01 \tag{4.48}$$

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.

4.4.2.3 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

$$
\begin{aligned}
P(N \geq 30 | N \geq 25) &= \frac{P(N \geq 30 \text{ and } N \geq 25)}{P(N \geq 25)} \quad ((2.5)) && (4.49)\\
&= \frac{P(N \geq 30)}{P(N \geq 25)} && (4.50)\\
&= \frac{1 - \Phi\left[(30 - 28.8)/3.1\right]}{1 - \Phi\left[(25 - 28.8)/3.1\right]} && (4.51)\\
&= 0.39 && (4.52)
\end{aligned}
$$

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 4.4.2.7.

4.4.2.4 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.

4.4.2.5 Example: Cumulative Roundoff Error

Suppose that computer roundoff error in computing the square roots of numbers in a certain range is distributed uniformly on (-0.5,0.5), and that we will be computing the sum of n such square roots. Suppose we compute a sum of 50 square roots. Let's find the approximate probability that the

[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 4.4.2.9.

sum is more than 2.0 higher than it should be. (Assume that the error in the summing operation is negligible compared to that of the square root operation.)

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

$$T = U_1 + ... + U_{50} \tag{4.53}$$

By the Central Limit Theorem, T has an approximately normal distribution, with mean 50 EU and variance 50 Var(U), where U is a random variable having the distribution of the U_i. From Section 4.4.1.1, we know that

$$EU = (-0.5 + 0.5)/2 = 0, \quad Var(U) = \frac{1}{12}[0.5 - (-0.5)]^2 = \frac{1}{12} \tag{4.54}$$

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

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

4.4.2.6 Example: Bug Counts

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

Here X_i is the number of bugs in the ith section of code, and T is the total number of bugs. 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
```

4.4.2.7 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):

$$T = T_1 + ... + T_n \tag{4.55}$$

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

$$P(X > 12) = P(Z > \frac{12 - 10}{\sqrt{5}}) \approx 1 - \Phi(0.894) = 0.186 \tag{4.56}$$

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 (4.56), we might write

$$P(X > 12) = P(X \geq 13) = P(Z > \frac{13 - 10}{\sqrt{5}}) \approx 1 - \Phi(1.342) = 0.090 \tag{4.57}$$

That value of 0.090 is considerably smaller than the 0.186 we got from (4.56). We could "split the difference" this way:

$$P(X > 12) = P(X \geq 12.5) = P(Z > \frac{12.5 - 10}{\sqrt{5}}) \approx 1 - \Phi(1.118) = 0.132 \tag{4.58}$$

(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.

4.4.2.8 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.

4.4.2.9 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*

$$\lim_{n \to \infty} P(L_n \leq t) = P(M \leq t), \text{ for all } t \tag{4.59}$$

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*

$$Z = \frac{X_1 + ... + X_n - nm}{v\sqrt{n}} \tag{4.60}$$

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

4.4.2.10 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.

4.4.3 The Chi-Square Family of Distributions

4.4.3.1 Density and Properties

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

$$Y = Z_1^2 + ... + Z_k^2 \tag{4.61}$$

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 4.4.5 below, with r = k/2 and $\lambda = 0.5$.

The R functions **dchisq()**, **pchisq()**, **qchisq()** and **rchisq()** give us the density, cdf, quantile function and random number generator for the chi-squared family. The second argument in each case is the number of degrees of freedom. The first argument is the argument to the corresponding math function in all cases but **rchisq()**, in which it is the number of random variates to be generated.

For instance, to get the value of $f_X(5.2)$ for a chi-squared random variable having 3 degrees of freedom, we make the following call:

```
> dchisq(5.2,3)
[1] 0.06756878
```

4.4.3.2 Example: Error in Pin Placement

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. Let's find $P(W > 0.6)$.

Since a distance is the square root of a sum of squares, this sounds like the chi-squared distribution might be relevant. So, let's first convert the problem to one involving squared distance:

$$P(W > 0.6) = P(W^2 > 0.36) \tag{4.62}$$

But $W^2 = X^2 + Y^2$, so

$$P(W > 0.6) = P(X^2 + Y^2 > 0.36) \tag{4.63}$$

This is not quite chi-squared, as that distribution involves the sum of squares of independent N(0,1) random variables. But due to the normal family's closure under affine transformations (page 95), we know that X/0.2 and Y/0.2 do have N(0,1) distributions. So write

$$P(W > 0.7) = P[(X/0.2)^2 + (Y/0.2)^2 > 0.36/0.2^2] \tag{4.64}$$

Now evaluate the right-hand side:

```
> 1 - pchisq(0.36/0.04,2)
[1] 0.01110900
```

4.4.3.3 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]

4.4.4 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

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

confused, though, with the term *exponential family* that arises in mathematical statistics, which includes exponential distributions but is much broader.

4.4.4.1 Density and Properties

The densities in this family have the form

$$f_W(t) = \lambda e^{-\lambda t}, 0 < t < \infty \tag{4.65}$$

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

4.4.4.2 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

4.4.4.3 Example: Refunds on Failed Components

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

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

$$P(L < 500) = \int_0^{500} 0.0001 e^{-0.0001t}\, dt = 0.05 \tag{4.66}$$

Then plug p = 0.05 into (3.83) and (3.84).

4.4.4.4 Example: Overtime Parking Fees

A certain public parking garage charges parking fees of $1.50 for the first hour, and $1 per hour after that. Suppose parking times T are exponentially distributed with mean 1.5 hours. Let W denote the total fee paid. Let's find E(W) and Var(W).

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

$$W = \begin{cases} 1.5T, & \text{if } T \leq 1 \\ 1.5 + 1 \cdot (T-1) = T + 0.5, & \text{if } T > 1 \end{cases} \tag{4.67}$$

That's good, because we know how to find the expected value of a function of a continuous random variable, from (4.24). Defining g() as in (4.67) above, we have

$$EW = \int_0^\infty g(t) \ \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt = \int_0^1 1.5t \ \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt + \int_1^\infty (t+0.5) \ \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt \tag{4.68}$$

The integration is left to the reader.

Now, what about Var(W)? As is often the case, it's easier to use (3.29), so we need to find $E(W^2)$. The above integration becomes

$$E(W^2) = \int_0^\infty g^2(t) \ \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt = \int_0^1 1.5^2 t \ \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt + \int_1^\infty (t+0.5)^2 \ \frac{1}{1.5} e^{-\frac{1}{1.5}t} dt \tag{4.69}$$

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

4.4.4.5 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

$$T_r = X_1 + ... + X_r \tag{4.70}$$

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*

$$T_r = X_1 + ... + X_r \tag{4.71}$$

and

$$N(t) = \max\{k : T_k \leq t\} \tag{4.72}$$

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

$$\begin{aligned} P(X_1 > t) &= P[N(t) = 0] \quad \text{(see above equiv.)} && (4.73)\\ &= \frac{(\lambda t)^0}{0!} \cdot e^{-\lambda t} \quad \text{((3.104)} && (4.74)\\ &= e^{-\lambda t} && (4.75) \end{aligned}$$

Then

$$f_{X_1}(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = \lambda e^{-\lambda t} \tag{4.76}$$

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,

$$P[N(u) = k] = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{4.77}$$

The proof for the case k = 0 just reverses (4.73) 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 (4.78) from u to infinity. One needs to perform k-1 integrations by parts, and eventually one arrives at (4.77), 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 (4.65).

4.4.4.6 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 5. Due to this property, exponential distributions play a central role in Markov chains (Chapter 16).

4.4.5 The Gamma Family of Distributions

4.4.5.1 Density and Properties

Recall Equation (4.70), 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

$$f_{T_r}(t) = \frac{1}{(r-1)!}\lambda^r t^{r-1} e^{-\lambda t},\ t > 0 \tag{4.78}$$

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 (4.78).

then

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

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} \, dx \tag{4.79}$$

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

$$f_W(t) = \frac{1}{\Gamma(r)} \lambda^r t^{r-1} e^{-\lambda t}, \; t > 0 \tag{4.80}$$

(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 (4.70) 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 $\lambda = 0.5$.

4.4.5.2 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

$$P(Y > 552) = \int_{552}^{\infty} \frac{1}{4!} 0.01^5 t^4 e^{-0.01t} \, dt \tag{4.81}$$

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
```

4.4.5.3 Importance in Modeling

As seen in (4.70), 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.

4.4.6 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.

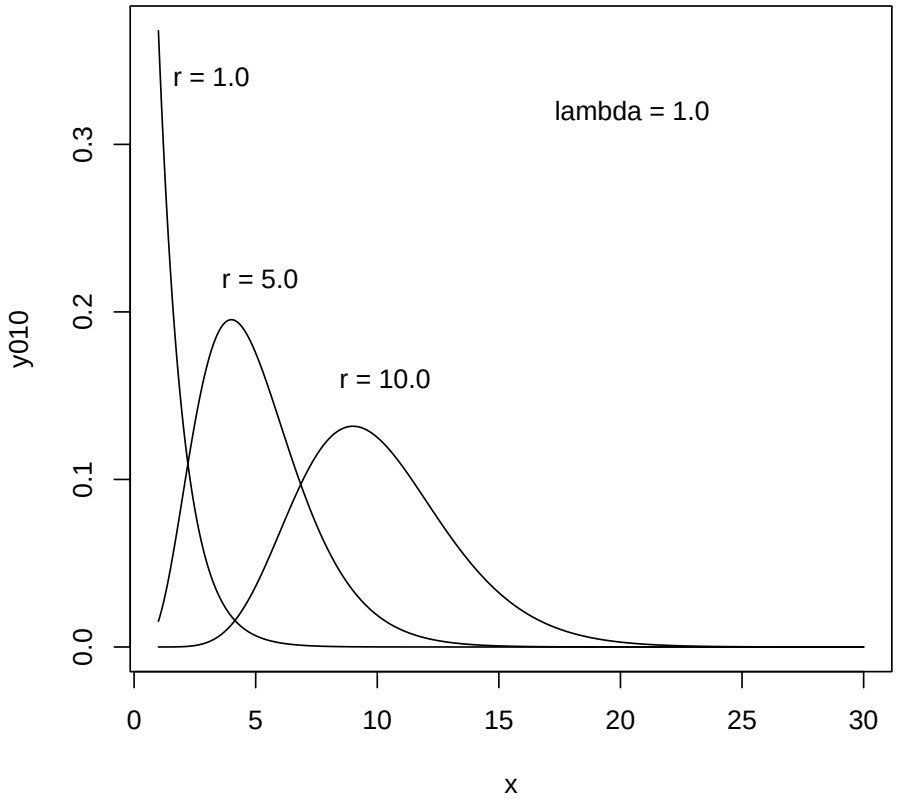
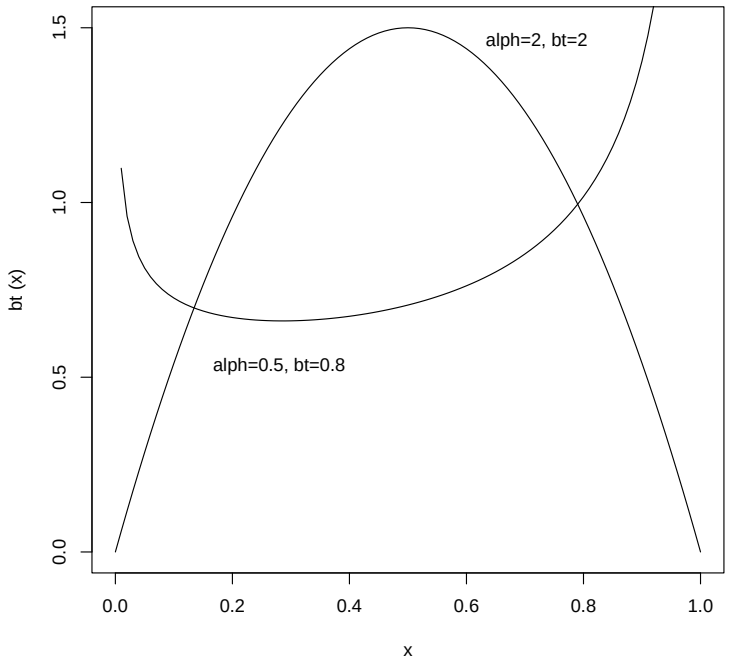


Figure 4.2: Various Gamma Densities

The densities of the family have the following form:

$$\frac{\Gamma(\alpha+\lambda)}{\Gamma(\alpha)\Gamma(\lambda)}(1-t)^{\alpha-1}t^{\beta-1} \tag{4.82}$$

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



The mean and variance are

$$\frac{\alpha}{\alpha+\beta} \tag{4.83}$$

and

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4.84}$$

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 13.

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

$$\begin{aligned} P[G(U) \leq t] & \\ &= P[U \leq G^{-1}(t)] \\ &= P[U \leq H(t)] \\ &= H(t) \end{aligned} \tag{4.85}$$

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

$$f_X(t) = \frac{1}{\pi c} \frac{1}{1 + (\frac{t-b}{c})^2}, \ \infty < t < \infty \tag{4.86}$$

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.

10. 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.

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

12. 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 $0 < t < 1$.

13. 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)

14. 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, $E[(Z - EZ)^3]$.

15. 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?

16. 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).

17. 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

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.

5.1 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**.

5.1.1 Derivation and Intuition

What the term *memoryless* means for a random variable W is that for all positive t and u

$$P(W > t + u | W > t) = P(W > u) \tag{5.1}$$

Any exponentially distributed random variable has this property. Let's derive this:

$$
\begin{aligned}
P(W > t+u | W > t) &= \frac{P(W > t+u \text{ and } W > t)}{P(W > t)} && (5.2)\\
&= \frac{P(W > t+u)}{P(W > t)} && (5.3)\\
&= \frac{\int_{t+u}^{\infty} \lambda e^{-\lambda s}\, ds}{\int_{t}^{\infty} \lambda e^{-\lambda s}\, ds} && (5.4)\\
&= e^{-\lambda u} && (5.5)\\
&= P(W > u) && (5.6)
\end{aligned}
$$

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 (5.1) 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 (5.1), we would have

$$R(t+u)]/R(t) = R(u) \tag{5.7}$$

or

$$R(t+u) = R(t)R(u) \tag{5.8}$$

Differentiating both sides with respect to t, we'd have

$$R\prime(t+u) = R\prime(t)R(u) \tag{5.9}$$

Setting t to 0, this would say

$$R\prime(u) = R\prime(0)R(u) \tag{5.10}$$

This is a well-known differential equation, whose solution is

$$R(u) = e^{-cu} \tag{5.11}$$

which is exactly 1 minus the cdf for an exponentially distributed random variable.

5.1.2 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 4.4.4.5 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 16.

5.1.3 Example: Light Bulbs

Suppose the lifetimes in years of light bulbs have the density 2t/15 on (1,4), 0 elsewhere. Say I've been using bulb A for 2.5 years now in a certain lamp, and am continuing to use it. But at this

time I put a new bulb, B, in a second lamp. I am curious as to which bulb is more likely to burn out within the next 1.2 years. Let's find the two probabilities.

For bulb A:

$$P(L > 3.7|L > 2.5) = \frac{P(L > 3.7)}{P(L > 2.5)} = 0.24 \tag{5.12}$$

For bulb B:

$$P(X > 1.2) = \int_{1.2}^{4} 2t/15 \; dt = 0.97 \tag{5.13}$$

5.2 Hazard Functions

5.2.1 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

$$P(L = 81|L > 80) = \frac{P(L = 81 \text{ and } L > 80)}{P(L > 80)} = \frac{P(L = 81)}{P(L > 80)} = \frac{p_L(81)}{1 - F_L(80)} \tag{5.14}$$

In general, for discrete L, we define its **hazard function** as

$$h_L(i) = \frac{p_L(i)}{1 - F_L(i-1)} \tag{5.15}$$

By analogy, for continuous L we define

$$h_L(t) = \frac{f_L(t)}{1 - F_L(t)} \tag{5.16}$$

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 (5.16) 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.

Some parametric families of distributions have strictly increasing failure rates (IFR). Some of strictly decreasing failure rates (DFR). 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 (5.16) is the derivative of $-ln[1-F_L(t)]$. Therefore

$$\int_0^t h_L(s)\ ds = -\ln[1 - F_L(t)] \tag{5.17}$$

so that

$$1 - F_L(t) = e^{-\int_0^t h_L(s)\ ds} \tag{5.18}$$

and thus[1]

$$f_L(t) = h_L(t)e^{-\int_0^t h_L(s)\ ds} \tag{5.19}$$

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.

[1]Recall that the derivative of the integral of a function is the original function!

5.2.2 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.

5.3 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 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.*

5.3.1 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.

5.3.2 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.

$$p_X(2) = p_X(6) = 0.1 \tag{5.20}$$

Intuitively, one would then reason that

$$p_Y(6) = 3p_Y(2) \tag{5.21}$$

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

$$p_Y(j) \propto jp_X(j) \tag{5.22}$$

where $\propto$ is the "is proportional to" symbol. Thus

$$p_Y(j) = cjp_X(j) \tag{5.23}$$

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

$$1 = \sum_j p_Y(j) = \sum_j cjp_X(j) \tag{5.24}$$

That last term is c E(X)! So, c = 1/EX, and

$$p_Y(j) = \frac{1}{EX} \cdot jp_X(j) \tag{5.25}$$

The continuous analog of (5.25) is

$$f_Y(t) = \frac{1}{EX} \cdot t f_X(t) \tag{5.26}$$

So, for our bus example, in which $f_X(t) = 0.1e^{-0.1t},\ t > 0$ and EX = 10,

$$f_Y(t) = 0.01te^{-0.1t} \tag{5.27}$$

You may recognize this as an Erlang density with r = 2 and $\lambda = 0.1$. That distribution does indeed have mean 20.

5.4 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**.

5.4.1 Renewal Theory

Recall the light bulb example of Section 4.4.4.5. Every time a light bulb burns out, we immediately replace it with a new one. The time of the r^{th} replacement is denoted by T_r, and satisfies the relation

$$N(t) = \max\{k : T_k \leq t\} \tag{5.28}$$

where N(t) is the number of replacements that have occurred by time t and X_i is the lifetime of the i^{th} 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

$$N(t) \geq k \text{ if and only if } T_k \leq t \tag{5.29}$$

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 (5.29) 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.

5.4.2 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,

$$f_{D|V}(s,t) = \frac{1}{t}, \quad 0 < s < t \tag{5.30}$$

Thus (9.2) yields

$$f_{D,V}(s,t) = \frac{1}{t} \cdot f_V(t), \quad 0 < s < t \tag{5.31}$$

Then (8.17) shows

$$f_D(s) = \int_s^\infty \frac{1}{t} \cdot f_V(t)\ dt \tag{5.32}$$

$$= \int_s^\infty \frac{1}{EX} \cdot f_X(t)\ dt \tag{5.33}$$

$$= \frac{1 - F_X(s)}{EX} \tag{5.34}$$

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 (5.32) is really the limiting density of $f_{D(t)}$ as $t \to \infty$.

5.4.3 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, 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

$$\lim_{t \to \infty} f_{A(t)}(w) = \frac{1 - F_L(w)}{E(L)} \tag{5.35}$$

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:

[2] If you are wondering about that first equality in (5.32), 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 9.1.3.

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,

$$p_{i,0} = P(L = i + 1 | L > i) = \frac{p_L(i+1)}{1 - F_L(i)} \tag{5.36}$$

$$p_{i,i+1} = \frac{1 - F_L(i+1)}{1 - F_L(i)} \tag{5.37}$$

Define

$$q_i = \frac{1 - F_L(i+1)}{1 - F_L(i)} \tag{5.38}$$

and write

$$\pi_{i+1} = \pi_i q_i \tag{5.39}$$

Applying (5.39) recursively, we have

$$\pi_{i+1} = \pi_0 q_i q_{i-1}) \cdots q_0 \tag{5.40}$$

But the right-hand side of (5.40) telescopes down to

$$\pi_{i+1} = \pi_0 [1 - F_L(i+1)] \tag{5.41}$$

Then

$$1 = \sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} [1 - F_L(i)] = \pi_0 E(L) \tag{5.42}$$

Thus

$$\pi_i = \frac{1 - F_L(i+1)}{EL} \tag{5.43}$$

in analogy to (5.35).

5.4.4 Mean of the Residual and Age Distributions

Taking the expected value of (5.32) or (5.35), we get a double integral. Reversing the order of integration, we find that the mean residual life or age is given by

$$\frac{E(L^2)}{2EL} \tag{5.44}$$

5.4.5 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.

5.4.6 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 (5.35), we have

$$f_A(w) = \frac{1 - \frac{w}{3}}{1.5} = \frac{2}{3} - \frac{2}{9}w \tag{5.45}$$

Then

$$P(A > 1) = \int_1^3 \left(\frac{2}{3} - \frac{2}{9}w\right)\ dw = \frac{4}{9} \tag{5.46}$$

5.4.7 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

$$s(\tau) = \lim_{t \to \infty} E[S(t, \tau)] \tag{5.47}$$

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 (5.43).

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. $A_i(t) < \tau$.

[3] Note, though, tht all random variables here are discrete, not continuous.

Thus, using (5.43) and letting 1_i be 1 or 0 according to whether or not $A_i(t) < \tau$, we have that

$$\begin{aligned} s(\tau) &= \lim_{t \to \infty} E(\sum_{i=1}^{n} 1_i) \\ &= \lim_{t \to \infty} \sum_{i=1}^{n} P(A_i(t) < \tau) \\ &= \sum_{i=1}^{n} \sum_{j=0}^{\tau-1} \frac{1 - F_i(j)}{E(L_i)} \end{aligned} \tag{5.48}$$

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 5.3. 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 5.1, 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 (5.1).

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 30. Find the hazard function $h_V(t)$, where V is the number of dots obtained on one roll (1, 2 or 3).

6. 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 $t > 0.5$.

Chapter 6

Stop and Review

There's quite a lot of material in the preceding chapters, but it's crucial that you have a good command of it before proceeding, as the coming chapters will continue to build on it.

With that aim, here are the highlights of what we've covered so far, with links to the places at which they were covered:

- **expected value** (Section 3.4):

  Consider random variables X and Y (not assumed independent), and constants c_1 and c_2. We have:

  $$E(X+Y) = EX + EY \tag{6.1}$$

  $$E(c_1 X) = c_1 X \tag{6.2}$$

  $$E(c_1 X + c_2 Y) = c_1 EX + c_2 EY \tag{6.3}$$

  By induction,

  $$E(a_1 U_1 + ... + a_k U_k) = a_1 EX_1 + ... + a_k EX_k \tag{6.4}$$

  for random variables U_i and constants a_i.

- **variance** (Section 3.5):

Consider random variables X and Y (now assumed independent), and constants c_1 and c_2. We have:

$$Var(X+Y) = Var(X) + Var(Y) \tag{6.5}$$

$$Var(c_1 X) = c_1^2 Var(X) \tag{6.6}$$

By induction,

$$Var(a_1 U_1 + ... + a_k U_k) = a_1^2 Var(U_1) + ... + a_k^2 Var(U_k) \tag{6.7}$$

for independent random variables U_i and constants a_i.

- **indicator random variables** (Section 3.6):

  Equal 1 or 0, depending on whether a specified event A occurs.

  If T is an indicator random variable for the event A, then

$$ET = P(A), \quad Var(T) = P(A)[1 - P(A)] \tag{6.8}$$

- **distributions**:

  - **cdfs**(Section 4.2):
    For any random variable X,

$$F_X(t) = P(X \leq t), \qquad -\infty < t < \infty \tag{6.9}$$

  - **pmfs** (Section 3.12):
    For a discrete random variable X,

$$p_X(k) = P(X = k) \tag{6.10}$$

  - **density functions** (Section 3.12):
    For a continuous random variable X,

$$f_X(t) = \frac{d}{dt} F_X(t), \qquad -\infty < t < \infty \tag{6.11}$$

    and

$$P(X \text{ in } A) = \int_A f_X(s)\, ds \tag{6.12}$$

- **famous parametric families of distributions**:

  Just like one can have a family of curves, say $sin(2\pi n\theta(t)$ (different curve for each n), certain families of distributions have been found useful. They're called *parametric families*, because they are indexed by one or more parameters, anlagously to n above.

  discrete:

  - **geometric** (Section 3.13.1)
    Number of i.i.d. trials until first success. For success probability p:

$$p_N(k) = (1-p)^k p \tag{6.13}$$

$$EN = 1/p, \quad Var(N) = \frac{1-p}{p^2} \tag{6.14}$$

  - **binomial** (Section 3.13.2):
    Number of successes in n i.i.d. trials, probability p of success per trial:

$$p_N(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6.15}$$

$$EN = np, \quad Var(N) = np(1-p) \tag{6.16}$$

  - **Poisson** (Section 3.13.3):
    Has often been found to be a good model for counts over time periods.
    One parameter, often called λ. Then

$$p_N(k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, ... \tag{6.17}$$

$$EN = Var(N) = \lambda \tag{6.18}$$

  - **negative binomial** (Section 3.13.4):
    Number of i.i.d. trials until r^{th} success. For success probability p:

$$p_N(k) = \binom{k-1}{r-1} (1-p)^{k-r} p^r, k = r, r+1, ... \tag{6.19}$$

$$E(N) = r \cdot \frac{1}{p}, \quad Var(N) = r \cdot \frac{1-p}{p^2} \tag{6.20}$$

  continuous:

– **uniform** (Section 4.4.1.1):
All points "equally likely." If the interval is (q,r),

$$f_X(t) = \frac{1}{r-q}, \quad q < t < r \tag{6.21}$$

$$EX == \frac{q+r}{2}, \quad Var(D) = \frac{1}{12}(r-q)^2 \tag{6.22}$$

– **normal (Gaussian)** (Section 4.4.2):
"Bell-shaped curves." Useful due to Central Limit Theorem (Section 4.4.2.4. (Thus good approximation to binomial distribution.)
Closed under affine transformations (Section 4.4.2.1)!
Parameterized by mean and variance, μ and σ^2:

$$f_X(t) = \frac{1}{\sqrt{2\pi}\sigma}\ e^{-0.5\left(\frac{t-\mu}{\sigma}\right)^2}, -\infty < t < \infty \tag{6.23}$$

exponential (Section 4.4.4):
Memoryless! One parameter, usually called λ. Connectedt to Poisson family.

$$f_X(t) == \lambda e^{-\lambda t}, 0 < t < \infty \tag{6.24}$$

$$EX = 1/\lambda, \quad Var(X) = 1/\lambda^2 \tag{6.25}$$

- **gamma** (Section 4.4.5):
Special case, Erlang family, arises as the distribution of the sum of i.i.d. exponential random variables.

$$f_X(t) = \frac{1}{\Gamma(r)}\lambda^r t^{r-1} e^{-\lambda t}, \ t > 0 \tag{6.26}$$

Chapter 7

Covariance and Random Vectors

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, **which forms the very core of statistics**.

7.1 Measuring Co-variation of Random Variables

7.1.1 Covariance

Definition 16 *The* **covariance** *between random variables X and Y is defined as*

$$Cov(X,Y) = E[(X - EX)(Y - EY)] \tag{7.1}$$

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 (7.1) 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.

Linearity in both arguments:

$$Cov(aX + bY, cU + dV) = acCov(X, U) + adCov(X, V) + bcCov(Y, U) + bdCov(Y, V) \quad (7.2)$$

for any constants a, b, c and d.

Insensitivity to additive constants:

$$Cov(X, Y + q) = Cov(X, Y) \quad (7.3)$$

for any constant q and so on.

Covariance of a random variable with itself:

$$Cov(X, X) = Var(X) \quad (7.4)$$

for any X with finite variance.

Shortcut calculation of covariance:

$$Cov(X, Y) = E(XY) - EX \cdot EY \quad (7.5)$$

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 (7.5).

$$\begin{aligned} Cov(X, Y) &= E[(X - EX)(Y - EY)] \quad \text{(definition)} && (7.6)\\ &= E\left[XY - EX \cdot Y - EY \cdot X + EX \cdot EY\right] \quad \text{(algebra)} && (7.7)\\ &= E(XY) + E[-EX \cdot Y] + E[-EY \cdot X] + E[EX \cdot EY] \quad \text{(E[U+V]=EU+EV)} && (7.8)\\ &= E(XY) - EX \cdot EY \quad \text{(E[cU] = cEU, Ec = c)} && (7.9) \end{aligned}$$

Variance of sums:

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) \quad (7.10)$$

This comes from (7.5), the relation $Var(X) = E(X^2) - EX^2$ and the corresponding one for Y. Just substitute and do the algebra.

By induction, (7.10) generalizes for more than two variables:

$$Var(W_1 + ... + W_r) = \sum_{i=1}^{r} Var(W_i) + 2 \sum_{1 \leq j < i \leq r} Cov(W_i, W_j) \tag{7.11}$$

7.1.2 Example: Variance of Sum of Nonindependent Variables

Consider random variables X_1 and X_2, for which $Var(X_i) = 1.0$ for i = 1,2, and $Cov(X_1, X_2) = 0.5$. Let's find $Var(X_1 + X_2)$.

This is quite straightforward, from (7.10):

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1, X_2) = 3 \tag{7.12}$$

7.1.3 Example: the Committee Example Again

Let's find Var(M) in the committee example of Section 3.7. In (3.52), we wrote M as a sum of indicator random variables:

$$M = G_1 + G_2 + G_3 + G_4 \tag{7.13}$$

and found that

$$P(G_i = 1) = \frac{2}{3} \tag{7.14}$$

for all i.

You should review why this value is the same for all i, as this reasoning will be used again below. Also review Section 3.6.

Applying (7.11) to (7.13), we have

$$Var(M) = 4Var(G_1) + 12Cov(G_1.G_2) \tag{7.15}$$

Finding that first term is easy, from (3.44):

$$Var(G_1) = \frac{2}{3} \cdot \left(1 - \frac{2}{3}\right) = \frac{2}{9} \tag{7.16}$$

Now, what about $Cov(G_1.G_2)$? Equation (7.5) will be handy here:

$$Cov(G_1.G_2) = E(G_1G_2) - E(G_1)E(G_2) \tag{7.17}$$

That first term in (7.17) is

$$\begin{aligned} E(G_1G_2) &= P(G_1 = 1 \text{ and } G_2 = 1) && (7.18)\\ &= P(\text{choose a man on both the first and second pick}) && (7.19)\\ &= \frac{6}{9} \cdot \frac{5}{8} && (7.20)\\ &= \frac{5}{12} && (7.21) \end{aligned}$$

That second term in (7.17) is, again from Section 3.6,

$$\left(\frac{2}{3}\right)^2 = \frac{4}{9} \tag{7.22}$$

All that's left is to put this together in (7.15), left to the reader.

7.1.4 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 (7.2) 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

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \tag{7.23}$$

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

7.1.5 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  nturns <- 10
8  xyvals <- matrix(nrow=nreps,ncol=2)
9  for (rep in 1:nreps) {
10    x <- 0
11    y <- 0
12    for (turn in 1:nturns) {
13       # x's turn
14       x <- x + taketurn(x,y)
15       # y's turn
16       y <- y + taketurn(y,x)
17    }
18    xyvals[rep,] <- c(x,y)
19 }
20 print(cor(xyvals[,1],xyvals[,2]))
```

The output is 0.65. 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 (7.5). The reader should make sure he/she understands how this would be done.

7.2 Sets of Independent Random Variables

Recall from Section 3.3:

Definition 17 *Random variables X and Y are said to be* **independent** *if for any sets I and J, the events* $\{X \text{ is in } I\}$ *and* $\{Y \text{ is in } J\}$ *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.

7.2.1 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.

7.2.1.1 Expected Values Factor

If X and Y are independent, then

$$E(XY) = E(X)E(Y) \tag{7.24}$$

7.2.1.2 Covariance Is 0

If X and Y are independent, we have

$$Cov(X, Y) = 0 \tag{7.25}$$

and thus

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

This follows from (7.24) and (7.5).

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 (7.5).

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 8.4.2.1), then zero covariance does imply independence.

7.2.1.3 Variances Add

If X and Y are independent, then we have

$$Var(X+Y) = Var(X) + Var(Y). \tag{7.26}$$

This follows from (7.10) and (7.24).

7.2.2 Examples Involving Sets of Independent Random Variables

7.2.2.1 Example: Dice

In Section 7.1.1, 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

$$\begin{aligned} Cov(X,S) &= Cov(X, X+Y) \text{ (def. of S)} && (7.27)\\ &= Cov(X,X) + Cov(X,Y) \text{ (from (7.2))} && (7.28)\\ &= Var(X) + 0 \text{ (from (7.4), (7.25))} && (7.29) \end{aligned}$$

Also, from (7.26),

$$Var(S) = Var(X+Y) = Var(X) + Var(Y) \tag{7.30}$$

But Var(Y) = Var(X). So the correlation between X and S is

$$\rho(X,S) = \frac{Var(X)}{\sqrt{Var(X)}\sqrt{2Var(X)}} = 0.707 \tag{7.31}$$

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.

7.2.2.2 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 $Var(X_1 X_2)$.

$$\begin{aligned} Var(X_1X_2) &= E(X_1^2X_2^2) - [E(X_1X_2)]^2 \quad (3.29) && (7.32)\\ &= E(X_1^2) \cdot E(X_2^2) - \mu_1^2\mu_2^2 \quad (7.24) && (7.33)\\ &= (\sigma_1^2 + \mu_1^2)(\sigma_2^2 + \mu_2^2) - \mu_1^2\mu_2^2 && (7.34)\\ &= \sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 && (7.35) \end{aligned}$$

7.2.2.3 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 (7.24), we have

$$EZ = E\left(\frac{X}{Y}\right) = \frac{1}{p} \cdot E\left[\frac{1}{Y}\right] \tag{7.36}$$

so we need to find E(1/Y):

$$E(\frac{1}{Y}) = \sum_{i=1}^{\infty} \frac{1}{i}(1-p)^{i-1}p \tag{7.37}$$

Unfortunately, no further simplification seems possible.

Now let's find $F_Z(m)$ for a positive integer m.

$$
\begin{aligned}
F_Z(m) &= P\left(\frac{X}{Y} \leq m\right) && (7.38)\\
&= P(X \leq mY) && (7.39)\\
&= \sum_{i=1}^{\infty} P(Y=i)\ P(X \leq mY | Y=i) && (7.40)\\
&= \sum_{i=1}^{\infty} (1-p)^{i-1} p\ P(X \leq mi) && (7.41)\\
&= \sum_{i=1}^{\infty} (1-p)^{i-1} p\ [1-(1-p)^{mi}] && (7.42)
\end{aligned}
$$

this last step coming from (3.88).

We can actually reduce (7.42) to closed form, by writing

$$(1-p)^{i-1}(1-p)^{mi} = (1-p)^{mi+i-1} = \frac{1}{1-p}\left[(1-p)^{m+1}\right]^i \tag{7.43}$$

and then using (3.78). Details are left to the reader.

7.3 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

$$a_{i1}x_1 + ... + a_{in}x_n = b_i, \quad i = 1, ..., n, \tag{7.44}$$

where the x_i are the unknowns to be solved for.

This system can be represented compactly as

$$AX = B, \tag{7.45}$$

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.

7.3.1 Properties of Mean Vectors

Definition 18 *The expected value of W is defined to be the vector*

$$EW = (EW_1, ..., EW_k)' \tag{7.46}$$

The linearity of the components implies that of the vectors:

For any scalar constants c and d, and any random vectors V and W, we have

$$E(cV + dW) = cEV + dEW \tag{7.47}$$

where the multiplication and equality is now in the vector sense.

Also, multiplication by a constant matrix factors:

If A is a nonrandom matrix having k columns, then

$$E(AW) = AEW \tag{7.48}$$

7.3.2 Covariance Matrices

Definition 19 *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:

Say c is a constant scalar. Then cW is a k-component random vector like W, and

$$Cov(cW) = c^2 Cov(W) \tag{7.49}$$

If A is an r x k but nonrandom matrix. Then AW is an r-component random vector, and

$$Cov(AW) = A\ Cov(W)\ A' \tag{7.50}$$

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

$$Cov(V + W) = Cov(V) + Cov(W) \tag{7.51}$$

Of course, this is also true for sums of any (nonrandom) number of independent random vectors.

In analogy with (3.29), for any random vector Q,

$$Cov(Q) = E(QQ') - EQ\ (EQ)' \tag{7.52}$$

7.3.3 Example: Easy Sum Again

Let's redo the example in Section 7.1.2 again, this time using matrix methods.

First note that

$$X_1 + X_2 = (1,1) \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \tag{7.53}$$

so take A = (1,1). Then from (7.50),

$$Var(X_1 + X_2) = (1,1) \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3 \tag{7.54}$$

Of course using the matrix formulation didn't save us much time here, but for complex problems it's invaluable.

7.3.4 Example: (X,S) Dice Example Again

Recall Sec. 7.2.2.1. We rolled two dice, getting X and Y dots, and set S to X+Y. We then found $\rho(X,S)$. Let's find $\rho(X,S)$ using matrix methods.

The key is finding a proper choice for A in (7.50). A little thought shows that

$$\begin{pmatrix} X \\ S \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \tag{7.55}$$

Thus the covariance matrix of (X,S)' is

$$\begin{aligned} Cov[(X,S)'] &= \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} Var(X) & 0 \\ 0 & Var(Y) \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} && (7.56) \\ &= \begin{pmatrix} Var(X) & 0 \\ Var(X) & Var(Y) \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} && (7.57) \\ &= \begin{pmatrix} Var(X) & Var(X) \\ Var(X) & Var(X)+Var(Y) \end{pmatrix} && (7.58) \end{aligned}$$

since X and Y are independent. We would then proceed as before.

This matches what we found earlier, as it should, but shows how matrix methods can be used. This example was fairly simple, so those methods did not produce a large amount of streamlining, but in other examples later in the book, the matrix approach will be key.

7.3.5 Example: Dice Game

This example will necessitate a sneak preview of material in Chapter 8, but it will be worthwhile to present this example now, , in order to show why covariance matrices and their properties are so important.

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)
- $P(X > Y > Z)$

As will be shown in Section 8.4.2, the triple (X,Y,Z) has an approximate multivariate normal distribution. The latter is a generalization of the normal distribution, again covered in that section, but all we need to know here is that:

(a) If a random vector W has a multivariate normal distribution, and A is a constant matrix, then the new random vector AW is also multivariate normally distributed.

(b) R provides functions that compute probabilities involving this family of distributions.

Just as the univariate normal family is parameterized by the mean and variance, the multivariate normal family has as its parameters the mean *vector* and the covariance *matrix*.

We'll of course need to know the mean vector and covariance matrix of the random vector $(X, Y, Z)'$. Once again, this will be shown later (using (8.89) and (8.102)), but for now take them on faith:

$$E[(X, Y, Z)] = (50/6, 50/3, 50/2) \tag{7.59}$$

and

$$Cov[(X, Y, Z)] = 50 \begin{pmatrix} 5/36 & -1/18 & -1/12 \\ -1/18 & 2/9 & -1/6 \\ -1/12 & -1/6 & 1/4 \end{pmatrix} \tag{7.60}$$

We use the R function **pmvnorm()**, which computes probabilities of "rectangular" regions for multivariate normally distributed random vectors W.[1] The arguments we'll use for this function here are:

- **mean**: the mean vector
- **sigma**: the covariance matrix
- **lower, upper**: bounds for a multidimensional "rectangular" region of interest

[1] You must first load the **mvtnorm** library to use this function.

Since a multivariate normal distribution is characterized by its mean vector and covariance matrix, the first two arguments above shouldn't suprise you. But what about the other two?

The function finds the probability of our random vector falling into a multidimensional rectangular region that we specify, through the arguments are **lower** and **upper**. Note that these will typically be specified via R's **c()** function, but default values are recycled versions of **-Inf** and **Inf**, built-in R constants for $-\infty$ and ∞.

An important special case is that in which we specify **upper** but allow **lower** to be the default values, yielding:

$$P(W_1 \leq c_1, ..., W_r \leq c_r) \tag{7.61}$$

just what we need to find $P(X \leq 12 \text{ and } Y \leq 16)$.

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. (Recall the "correction for continuity," Section 4.4.2.7.) 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

$$P(X \leq 12 \text{ and } Y \leq 16) \approx 0.43 \tag{7.62}$$

Now, let's find the probability that our total winnings, T, is over \$90. We know that T = 5X + 2Y, and property (a) above applies. We simply choose the matrix to be

$$A = (5, 2, 0) \tag{7.63}$$

since

$$(5, 2, 0) \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = 5X + 2Y \tag{7.64}$$

Then property (a) tells us that 5X + 2Y also has an approximate multivariate normal random vector, which of course is univariate normal here. In other words, T has an approximate normal distribution, great since we know how to find probabilities involving that distribution!

We thus need the mean and variance of T. The mean is easy:

$$ET = E(5X + 2Y) = 5EX + 2EY = 250/6 + 100/3 = 75 \tag{7.65}$$

For the variance, use (7.50). (Since T is a 1-element vector, its covariance matrix reduces to simply Var(T).)

$$Var(T) = ACov \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} A' = (5,2)50 \begin{pmatrix} 5/36 & -1/18 & -1/12 \\ -1/18 & 2/9 & -1/6 \\ -1/12 & -1/6 & 1/4 \end{pmatrix} \begin{pmatrix} 5 \\ 2 \end{pmatrix} = 162.5 \tag{7.66}$$

So, proceeding as in Chapter 4, we have

$$P(T > 90) = 1 - \Phi\left(\frac{90 - 75}{162.5^{0.5}}\right) = 0.12 \tag{7.67}$$

Now to find $P(X > Y > Z)$, we need to work with $(U, V)' = (X - Y, Y - Z)$, so set

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \tag{7.68}$$

and then proceed as before to find $P(U > 0, V > 0)$. Now we take **lower** to be (0,0), and **upper** to be the default values, ∞ in **pmvnorm()**.

Exercises

1. Suppose the pair (X,Y)' has mean vector (0,2) and covariance matrix

$$\begin{pmatrix} 1 & 2 \\ 2 & 6 \end{pmatrix}$$

Find the covariance matrix of the pair U = (X+Y,X-2Y)'.

2. Show that

$$\rho(aX + b, cY + d) = \rho(X, Y) \tag{7.69}$$

for any constants a, b, c and d.

3. 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.

4. Using the properties of covariance in Section 7.1.1, show that for any random variables X and Y, Cov(X+Y,X-Y) = Var(X) - Var(Y).

5. 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

$$c = \frac{E(XY) - EX \cdot EY}{Var(X)} \tag{7.70}$$

$$d = \frac{E(X^2) \cdot EY - EX \cdot E(XY)}{Var(X)} \tag{7.71}$$

6. 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 $A(X, Y)$ 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.

7. 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.

8. Derive (7.3). Hint: A constant, q here, is a random variable, trivially, with 0 variance.

9. Consider a three-card hand drawn from a 52-card deck. Let X and Y denote the number of hearts and diamonds, respectively. Find $\rho(X, Y)$.

10. Consider the lightbulb example in Section 4.4.4.5. Use the "mailing tubes" on Var() and Cov() to find $\rho(X_1, T_2)$.

11. Find the following quantities for the dice example in Section 7.2.2.1:

(a) Cov(X,2S)

(b) Cov(X,S+Y)

(c) Cov(X+2Y,3X-Y)

(d) $p_{X,S}(3,8)$

12. 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 7.3, find the covariance matrix of $Y = (Y_1, Y_2, Y_3, Y_4)$.

Chapter 8

Multivariate PMFs and Densities

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

8.1 Multivariate Probability Mass Functions

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

$$p_{Y,B}(i,j) = P(Y = i \text{ and } B = j) = \frac{\binom{2}{i}\binom{3}{j}\binom{4}{4-i-j}}{\binom{9}{4}} \tag{8.1}$$

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

$$p_{U,V}(i,j) = P(U = i \text{ and } V = j) \tag{8.2}$$

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

$$p_{U,V,W}(i,j,k) = P(U = i \text{ and } V = j \text{ and } W = k) \tag{8.3}$$

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

$$P(X \in A) = \sum_{i \in A} p_X(i) \tag{8.4}$$

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

$$P[(U,V) \in A) = \sum_{(i,j) \in A} \sum p_{U,V}(i,j) \tag{8.5}$$

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 (8.5), we would set

$$A = \{(i,j) : i < j\} \tag{8.6}$$

and then

$$P(Y < B) = P[(Y, B) \in A) = \sum_{i=0}^{2} \sum_{j=i+1}^{3} p_{Y,B}(i, j) \tag{8.7}$$

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

$$E[g(X] = \sum_i g(i) p_X(i) \tag{8.8}$$

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

$$E[g(U, V)] = \sum_i \sum_j g(i, j) p_{U,V}(i, j) \tag{8.9}$$

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

$$E(YB) = \sum_{i=0}^{2} \sum_{j=0}^{3} ij \; p_{Y,B}(i, j) = 0.255 \tag{8.10}$$

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

$$p_U(i) = P(U = i) = \sum_j P(U = i, V = j) = \sum_j p_{U,V}(i, j) \tag{8.11}$$

For example, look at the table following (8.5). Evaluating (8.11) for i = 1, say, with U = Y and V = B, would give us 0.012 + 0.024 + 0.006 + 0.000 = 0.042. Then all that (8.11) tells us is the P(Y = 1) = 0.042, which is obvious from the table; (8.11) simply is an application of our old principle, "Break big events down into small events."

[1]Not so bizarre, we'll find in Section 7.1.1.

Needless to say, we can recover the marginal distribution of V similarly to (8.11):

$$p_V(j) = P(V = j) = \sum_i P(U = i, V = j) = \sum_i p_{U,V}(i, j) \tag{8.12}$$

8.2 Multivariate Densities

8.2.1 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

$$F_{X,Y}(u, v) = P(X \leq u \text{ and } Y \leq v) \tag{8.13}$$

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:

$$f_{X,Y}(u, v) = \frac{\partial^2}{\partial u \partial v} F_{X,Y}(u, v) \tag{8.14}$$

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.

8.2.2 Use of Multivariate Densities in Finding Probabilities and Expected Values

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

$$P[(X, Y) \epsilon A] = \iint_A f_{X,Y}(u, v)\ du\ dv \tag{8.15}$$

[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.

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.

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

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u,v) f_{X,Y}(u,v) \; du \; dv \tag{8.16}$$

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 (8.15). See (8.20) below for an example.

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

$$f_X(s) = \int_t f_{X,Y}(s,t) \; dt \tag{8.17}$$

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

8.2.3 Example: a Triangular Distribution

Suppose (X,Y) has the density

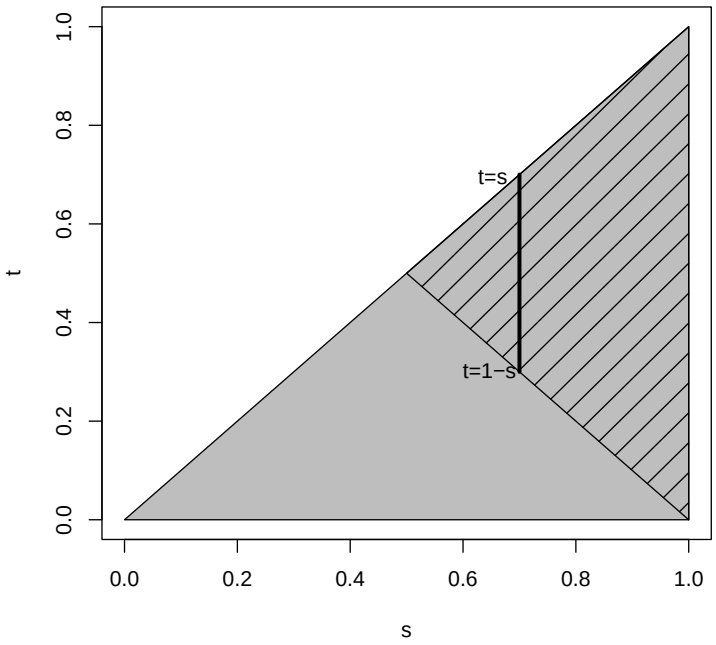
$$f_{X,Y}(s,t) = 8st, 0 < t < s < 1 \tag{8.18}$$

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 (8.15) 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 (8.7). 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 (8.15), 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.

$$P(X+Y>1) = \int_{0.5}^{1} \int_{1-s}^{s} 8st \; dt \; ds = \int_{0.5}^{1} 8s \cdot (s-0.5) \; ds = \frac{5}{6} \tag{8.19}$$

Following (8.16),

$$E[\sqrt{X+Y}] = \int_{0}^{1} \int_{0}^{s} \sqrt{s+t} \; 8st \; dt \; ds \tag{8.20}$$

Let's find the marginal density $f_Y(t)$. Just as we "summed out" in (8.11), in the continuous case

we must "integrate out" the s in (8.18):

$$f_Y(t) = \int_t^1 8st \; ds = 4t - 4t^3 \tag{8.21}$$

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

Let's find the correlation between X and Y for this density.

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^s st \cdot 8st \; dt \; ds && (8.22)\\ &= \int_0^1 8s^2 \cdot s^3/3 \; ds && (8.23)\\ &= \frac{4}{9} && (8.24) \end{aligned}$$

$$\begin{aligned} f_X(s) &= \int_0^s 8st \; dt && (8.25)\\ &= 4st^2\Big|_0^s && (8.26)\\ &= 4s^3 && (8.27) \end{aligned}$$

$$\begin{aligned} f_Y(t) &= \int_t^1 8st \; ds && (8.28)\\ &= 4t \cdot s^2\Big|_t^1 && (8.29)\\ &= 4t(1-t^2) && (8.30) \end{aligned}$$

$$EX = \int_0^1 s \cdot 4s^3 \; ds = \frac{4}{5} \tag{8.31}$$

$$E(X^2) = \int_0^1 s^2 \cdot 4s^3 \; ds = \frac{2}{3} \tag{8.32}$$

$$Var(X) = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = 0.027 \tag{8.33}$$

$$EY = \int_0^1 t \cdot (4t - 4t^3)\ ds = \frac{4}{3} - \frac{4}{5} = \frac{8}{15} \tag{8.34}$$

$$E(Y^2) = \int_0^1 t^2 \cdot (4t - 4t^3)\ dt = 1 - \frac{4}{6} = \frac{1}{3} \tag{8.35}$$

$$Var(Y) = \frac{1}{3} - \left(\frac{8}{15}\right)^2 = 0.049 \tag{8.36}$$

$$Cov(X,Y) = \frac{4}{9} - \frac{4}{5} \cdot \frac{8}{15} = 0.018 \tag{8.37}$$

$$\rho(X,Y) = \frac{0.018}{\sqrt{0.027 \cdot 0.049}} = 0.49 \tag{8.38}$$

8.2.4 Example: Train Rendezvouz

Train 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 trains. The bivariate density is

$$f_{X,Y}(s,t) = 2 - s - t,\ 0 < s,t < 1 \tag{8.39}$$

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. Let's find $P(W > 6)$.

First, convert this to a problem involving X and Y, since they are the random variables for which we have a density, and then use (8.15):

$$\begin{aligned} P(W > 0.1) &= P(Y + 0.1 < X) && (8.40) \\ &= \int_{0.1}^1 \int_0^{s-0.1} (2 - s - t)\ dtds && (8.41) \end{aligned}$$

8.3 More on Sets of Independent Random Variables

8.3.1 Probability Mass Functions and Densities Factor in the Independent Case

If X and Y are independent, then

$$p_{X,Y} = p_X p_Y \tag{8.42}$$

in the discrete case, and

$$f_{X,Y} = f_X f_Y \tag{8.43}$$

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:

$$\begin{aligned} p_{X,Y}(i,j) &= P(X = i \text{ and } Y = j) \text{ (definition)} && (8.44)\\ &= P(X = i)P(Y = j) \text{ (independence)} && (8.45)\\ &= p_X(i)p_Y(j) \text{ (definition))} && (8.46) \end{aligned}$$

Here is the proof for the continuous case;

$$\begin{aligned} f_{X,Y}(u,v) &= \frac{\partial^2}{\partial u \partial v} F_{X,Y}(u,v) && (8.47)\\ &= \frac{\partial^2}{\partial u \partial v} P(X \le u \text{ and } Y \le v) && (8.48)\\ &= \frac{\partial^2}{\partial u \partial v} \left[P(X \le u) \cdot P(Y \le v) \right] && (8.49)\\ &= \frac{\partial^2}{\partial u \partial v} F_X(u) \cdot F_Y(v) && (8.50)\\ &= f_X(u) f_Y(v) && (8.51) \end{aligned}$$

8.3.2 Convolution

Definition 21 *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*

[3]The reason for the asterisk, suggesting a product, will become clear in Section 9.4.3.

variable X+Y. In other words, convolution is a binary operation on the set of all densities.

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

$$f_Z(t) = \int_0^t g(s)h(t-s)\ ds \tag{8.52}$$

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;

$$\begin{aligned}
p_W(k) &= P(W=k) \text{ (by definition)} && (8.53)\\
&= P(U+V=k) \text{ (substitution)} && (8.54)\\
&= \sum_{i=0}^{k} P(U=i \text{ and } V=k-i) \text{ (“In what ways can it happen?”)} && (8.55)\\
&= \sum_{i=0}^{k} p_{U,V}(i,k-i) \text{ (by definition)} && (8.56)\\
&= \sum_{i=0}^{k} p_U(i)p_V(k-i) \text{ (from Section 8.3.1)} && (8.57)
\end{aligned}$$

Review the analogy between densities and pmfs in our unit on continuous random variables, Section 4.3.1, and then see how (8.52) is analogous to (8.53) through (8.57):

- k in (8.53) is analogous to t in (8.52)
- the limits 0 to k in (8.57) are analogous to the limits 0 to t in (8.52)
- the expression k-i in (8.57) is analogous to t-s in (8.52)
- and so on

8.3.3 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 (8.43), we know that their joint density is the product of their marginal densities,

$$f_{X,Y}(s,t) = 25e^{-5(s+t)}, s, t > 0 \tag{8.58}$$

Now

$$P(|X - Y| \leq 0.1) = 1 - P(|X - Y| > 0.1) = 1 - P(X > Y + 0.1) - P(Y > X + 0.1) \tag{8.59}$$

Look at that first probability. Applying (8.15) with $A = \{(s,t) : s > t + 0.1, 0 < s, t\}$, we have

$$P(X > Y + 0.1) = \int_0^\infty \int_{t+0.1}^\infty 25e^{-5(s+t)} \, ds \, dt = 0.303 \tag{8.60}$$

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.

8.3.4 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 (8.16) to be $|s - t|$.

$$\int_0^1 \int_0^1 |s - t| \cdot 1 \, ds \, dt = \frac{1}{3} \tag{8.61}$$

Let's find the density of the seek time $S = |X - Y|$:

$$\begin{aligned} F_S(v) &= P((X-Y| \leq v) && (8.62)\\ &= P(-v \leq X - Y \leq v) && (8.63)\\ &= 1 - P(X - Y < -v) - P(X - Y > v) && (8.64)\\ &= 1 - (1-v)^2 && (8.65) \end{aligned}$$

where for instance $P(X - Y > v)$ the integral of 1 on the triangle with vertices (v,0), (1,0) and (1,1-v), thus equal to the area of that triangle, $0.5(1-v)^2$.

Then

$$f_S(v) = \frac{d}{dt} F_S(v) = 2(1-v) \tag{8.66}$$

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 94, 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.

8.3.5 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

(8.3.2), we have

$$f_W(t) = \int_0^t f_X(s) f_Y(t-s)\ ds = \int_0^t 0.5 e^{-0.5s} e^{-(t-s)}\ ds = e^{-0.5t} - e^{-t},\ 0 < t < \infty \tag{8.67}$$

8.3.6 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 22 *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 4.4.4.5 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

$$\begin{aligned} F_Z(t) &= P(Z \leq t) \quad \text{(def. of cdf)} && (8.68)\\ &= 1 - P(Z > t) && (8.69)\\ &= 1 - P(W_1 > t \text{ and ... and } W_k > t) \quad (\min > t \text{ iff all } W_i > t) && (8.70)\\ &= 1 - \Pi_i \; P(W_i > t) \quad \text{(indep.)} && (8.71)\\ &= 1 - \Pi_i \; e^{-\lambda_i t} \quad \text{(expon. distr.)} && (8.72)\\ &= 1 - e^{-(\lambda_1 + ... + \lambda_n)t} && (8.73) \end{aligned}$$

Taking $\frac{d}{dt}$ of both sides shows (a).

For (b), suppose k = 2. we have that

$$\begin{aligned} P(Z = W_1) &= P(W_1 < W_2) && (8.74)\\ &= \int_0^\infty \int_t^\infty \lambda_1 e^{-\lambda_1 t} \lambda_2 e^{-\lambda_2 s} \; ds \; dt && (8.75)\\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2} && (8.76) \end{aligned}$$

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 16.4.

8.3.7 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 16.4, but all we need here is the result of Section 8.3.6.

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

$$\sum_{i=1}^{g-1} \frac{1}{i(g-i)} \tag{8.77}$$

Using a calculus approximation, we have

$$\int_1^{g-1} \frac{1}{x(g-x)}\ dx = \frac{1}{g} \int_1^{g-1} (\frac{1}{x} + \frac{1}{g-x})\ dx = \frac{2}{g} \ln(g-1) \tag{8.78}$$

The latter quantity goes to zero as $g \to \infty$. This confirms that the behavior seen by the student in simulations holds in general. In other words, (8.77) remains bounded as $g \to \infty$. This is a very interesting result, since it says that the mean time to alert is bounded no matter how big our group size is.

So, even though our model here was quite simple, probably overly so, it did explain why the student was seeing the surprising behavior in his simulations.

8.3.8 Example: Ethernet Again

In the Ethernet example in Section 8.3.3, we assumed that transmission time was a constant, 0.1. Now let's account for messages of varying sizes, by assuming that transmission time T for a message is random, exponentially distributed with mean 0.1. Let's find $P(X < Y$ and there is no collision).

That probability is equal to $P(X + T < Y)$. Well, this sounds like we're going to have to deal with triple integrals, but actually not. The derivation in Section 8.3.5 shows that the density of S = X+T is

$$f_S(t) = e^{-0.1t} - e^{-0.2t},\ 0 < t < \infty \tag{8.79}$$

Thus the joint density of S and Y is

$$f_{S,Y}(u,v) = (e^{-0.1u} - e^{-0.2u})0.2e^{-0.2v},\quad 0 < u, v, < \infty \tag{8.80}$$

We can then evaluate $P(S < Y)$ as a double integral, along the same lines as we did for instance in (8.19).

8.4 Parametric Families of Multivariate 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.

8.4.1 The Multinomial Family of Distributions

8.4.1.1 Probability Mass Function

This is a generalization of the binomial family.

Suppose one rolls 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 rolls occur in that order, i.e. the two 1s come first, then the 2, etc., then the probability is

$$\left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^1 \tag{8.81}$$

But there are many different orderings, in fact

$$\frac{8!}{2!1!0!1!3!1!} \tag{8.82}$$

of them, from Section 2.13.3, and thus

$$P(\text{two 1s, one 2, no 3s, one 4, three 5s, one 6}) = \frac{8!}{2!1!0!1!3!1!} \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^1 \tag{8.83}$$

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,

$$p_{X_1,...,X_r}(j_1, ..., j_r) = \frac{n!}{j_1!...j_r!} p_1^{j_1} ... p_r^{j_r} \tag{8.84}$$

Note that this family of distributions has r+1 parameters.

R has the function **dmultinom()** for the multinomial pmf. The call **dmultinom(x,n,prob**, **x** evaluates (8.84), where **x** is the vector $(j_1, ..., j_r)$ and **prob** is $(p_1, ..., p_r)$.

We can simulate multinomial random vectors in R using the **sample()** function:

```
# n is the number of trials, 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
```

```
7     # tabulate the counts (could be done more efficiently)
8     for (i in 1:n) {
9         j <- outcome[i]
10        counts[j] <- counts[j] + 1
11    }
12    return(counts)
13 }
```

8.4.1.2 Example: Component Lifetimes

Say the lifetimes of some electronic component, say a disk drive, are exponentially distributed with mean 4.5 years. If we have six of them, what is the probability that two fail before 1 year, two last between 1 and 2 years, and the remaining two last more than 2 years?

Let (X,Y,Z) be the number that last in the three time intervals. Then this vector has a multinomial distribution, with n = 6 trials, and

$$p_1 = \int_0^1 \frac{1}{4.5} e^{-t/4.5} \, dt = 0.20 \tag{8.85}$$

$$p_2 = \int_1^2 \frac{1}{4.5} e^{-t/4.5} \, dt = 0.16 \tag{8.86}$$

$$p_3 = \int_2^\infty \frac{1}{4.5} e^{-t/4.5} \, dt = 0.64 \tag{8.87}$$

We then use (8.84) to find the specified probability, which is:

$$\frac{6!}{2!2!2!} \, 0.20^2 0.16^2 0.64^2 \tag{8.88}$$

8.4.1.3 Mean Vectors and Covariance Matrices in the Multinomial Family

Consider a multinomially distributed random vector $X = (X_1, ..., X_r)'$, with n trials and category probabilities p_i. Let's find its mean vector and covariance matrix.

First, note that the marginal distributions of the X_i are binomial! So,

$$EX_i = np_i \text{ and } Var(X_i) = np_i(1 - p_i) \tag{8.89}$$

So we know EX now:

$$EX = \begin{pmatrix} np_1 \\ ... \\ np_r \end{pmatrix} \tag{8.90}$$

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? The derivation will follow in the footsteps of those of (3.97), but now in a vector context. Prepare to use your indicator random variable, random vector and covariance matrix skills! Also, this derivation will really build up your "probabilistic stamina level." So, it's good for you! But **now is the time to review (3.97), Section 3.6 and Section 7.3, before continuing.**

We'll continue the notation of the last section. In order to keep on eye on the concrete, we'll often illustrate the notation with the die example above; there we rolled a die 8 times, and defined 6 categories (one dot, two dots, etc.). We were interested in probabilities involving the number of trials that result in each of the 6 categories.

Define the random vector T_i to be the outcome of the ith trial. It is a vector of indicator random variables, one for each of the r categories. In the die example, for instance, consider the second roll, which is recorded in T_2. If that roll turns out to be, say, 5, then

$$T_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \tag{8.91}$$

Here is the key observation:

$$\begin{pmatrix} X_1 \\ ... \\ X_r \end{pmatrix} = \sum_{i=1}^{n} T_i \tag{8.92}$$

Keep in mind, (8.92) is a vector equation. In the die example, the first element of the left-hand side,X_1, is the number of times the 1-dot face turns up, and on the right-hand side, the first element of T_i is 1 or 0, according to whether the 1-dot face turns up on the ith roll. Make sure you believe this equation before continuing.

Since the trials are independent, (7.51) and (8.92) now tell us that

$$Cov[(X_1, ..., X_r)'| = \sum_{i=1}^{n} Cov(T_i) \tag{8.93}$$

But the trials are not only independent, but also identically distributed. (The die, for instance, has the same probabilities on each trial.) So the last equation becomes

$$Cov\left[\begin{pmatrix} X_1 \\ ... \\ X_r \end{pmatrix}\right] = nCov(T_1) \tag{8.94}$$

One more step to go. Remember, T_1 is a vector, recording what happens on the first trial, e.g. the first roll of the die. Write it as

$$T_1 = \begin{pmatrix} U_1 \\ ... \\ U_r \end{pmatrix} \tag{8.95}$$

Then the covariance matrix of T_1 consists of elements of the form

$$Cov(U_i, U_j) \tag{8.96}$$

Let's evaluate them.

Case 1: $i = j$

$$\begin{aligned} Cov(U_i, U_j) &= Var(U_i) \quad (7.4) && (8.97)\\ &= p_i(1 - p_i) \quad (3.44) && (8.98) \end{aligned}$$

Case 2: $i \neq j$

$$\begin{aligned} Cov(U_i, U_j) &= E(U_iU_j) - EU_iEU_j \quad (7.5) && (8.99)\\ &= E(U_iU_j) - p_ip_j \quad (3.43) && (8.100)\\ &= -p_ip_j && (8.101) \end{aligned}$$

with that last step coming from the fact that U_i and U_j can never both be 1 (e.g. never on the same line of the our "notebook"). Thus the product U_iU_j is always 0, and thus so is is expected value. In the die example, for instance, if our roll resulted in the 2-dot face turned upward, then the 5-dot face definitely did NOT turn upward, so $U_2 = 1$ while $U_5 = 0$.

So, we've now found $Cov(T_1)$, and using this in (8.94), we see that

$$Cov\left[\begin{pmatrix} X_1 \\ ... \\ X_r \end{pmatrix}\right] = n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & ... & -p_1p_r \\ -p_1p_2 & p_2(1-p_2) & ... & -p_2p_r \\ ... & ... & ... & ... \\ ... & ... & ... & p_r(1-p_r) \end{pmatrix} \tag{8.102}$$

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 from (8.102) and (7.49), we have

$$Cov(R) = \frac{1}{n} \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & ... & -p_1p_r \\ -p_1p_2 & p_2(1-p_2) & ... & -p_2p_r \\ ... & ... & ... & ... \\ ... & ... & ... & p_r(1-p_r) \end{pmatrix} \tag{8.103}$$

Whew! That was a workout, but these formulas will become very useful later on, both in this chapter and subsequent ones.

8.4.1.4 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.

8.4.2 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.

8.4.2.1 Densities

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

$$f_{X,Y}(s,t) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(s-\mu_1)^2}{\sigma_1^2}+\frac{(t-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(s-\mu_1)(t-\mu 2)}{\sigma_1\sigma_2}\right]}, \quad -\infty < s,t < \infty \tag{8.104}$$

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 the 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.

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:

$$f_X(t) = ce^{-0.5(t-\mu)'\Sigma^{-1}(t-\mu)} \tag{8.105}$$

Here c is a constant, needed to make the density integrate to 1.0. It turns out that

$$c = \frac{1}{(2\pi)^{k/2}\sqrt{det(\Sigma)}} \tag{8.106}$$

but we'll never use this fact.

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.

8.4.2.2 Geometric Interpretation

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`.[4] 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 8.1 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 8.2, 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).

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

$$X_1 - \mu_1 = \frac{\sigma_1}{\sigma_2}(X_2 - \mu_2) \tag{8.107}$$

In this case, that would be

$$X_1 = \sqrt{\frac{10}{15}} X_2 = 0.82 X_2 \tag{8.108}$$

[4] 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.

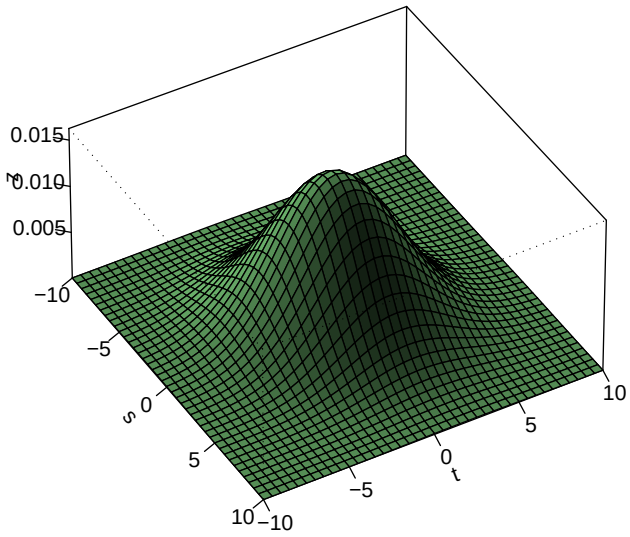


Figure 8.1: Bivariate Normal Density, $\rho = 0.2$

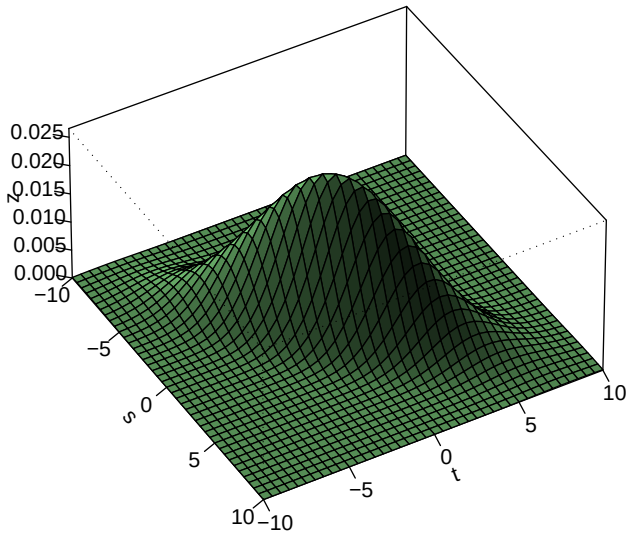


Figure 8.2: Bivariate Normal Density, $\rho = 0.8$

8.4.2.3 Properties of Multivariate Normal Distributions

Theorem 23 *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.[5]

The parameters of this new normal distribution must be $EY = A\mu$ and $Cov(Y) = A\Sigma A'$, by (7.48) and (7.50).

(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

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{8.109}$$

in (b) above.

(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 $a'\Sigma a$.

[5] Note that this is a generalization of the material on affine transformations on page 95.

(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

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \tag{8.110}$$

(iv) The r-component random vector X has a multivariate normal distribution if and only if $c'X$ has a univariate normal distribution for all constant r-component vectors c.

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**.

8.4.2.4 The Multivariate Central Limit Theorem

The multidimensional version of the Central Limit Theorem holds. A sum of independent identically distributed (iid) 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.

The proof of the multivariate CLT is easy, from Property (iv) above. Say we have a sum of iid random vectors:

$$S = X_1 + ... + X_n \tag{8.111}$$

Then

$$c'S = c'X_1 + ... + c'X_n \tag{8.112}$$

Now on the right side we have a sum of iid *scalars*, not vectors, so the univariate CLT applies! We thus know the right-hand side is a approximately normal for all c, which means $c'S$ is also approximately normal for all c, which then by (iv) above means that S itself is approximately multivariate normal.

8.4.2.5 Example: Finishing the Loose Ends from the Dice Game

Recall the game example in Section 7.3.5:

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.

Our analysis relied on the vector (X,Y,Z)' having an approximate multivariate normal distribution. Where does that come from? Well, first note that the exact distribution of (X,Y,Z)' is multinomial. Then recall (8.92). The latter makes (X,Y,Z)' a sum of iid vectors, so that the multivariate CLT applies.

8.4.2.6 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 15.4.1.

To see a bit of how this works, note that in Figure 8.2, 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 15.4.1 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]

[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.

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 $f_X(s)$.

(b) 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) To what parametric family in this book does the distribution of Y+Z belong?

3. Suppose X and Y are independent, each having an exponential distribution with means 1.0 and 2.0, respectively. Find $P(Y > X^2)$.

4. 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 7.3, find the covariance matrix of the pair U = (X+Y,X-2Y)'. Does U have a bivariate normal distribution?

5. 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.

- 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.

6. 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 $\rho(X, Y)$.

7. 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 7.3.5, 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 $Cov()$.

8. Consider the coin game in Section 3.14.1. Find $F_{X_3,Y_3}(0,0)$.

9. Suppose the random vector $X = (X_1, X_2, X_3)'$ has mean $(2.0, 3.0, 8.2)'$ and covariance matrix

$$\begin{pmatrix} 1 & 0.4 & -0.2 \\ & 1 & 0.25 \\ & & 3 \end{pmatrix} \tag{8.113}$$

(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.

10. Find the (approximate) output of this R code, by using the analytical techniques of this chapter:

```
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)
```

11. Use the convolution formula (8.52) to derive (4.78) for the case r = 2. Explain your steps carefully!

12. 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 _____.

13. Consider the computer worm example in Section 8.3.7. 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.)

14. 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, leaving your answers in integral form:

(a) $E(X^2 + XY^{0.5})$

(b) $P(Y > 0.5X)$

(c) $F_{X,Y}(0.6, 0.2)$

Chapter 10

Introduction to Confidence Intervals

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 14.

10.1 Sampling Distributions

We first will set up some infrastructure, which will be used heavily throughout the next few chapters.

10.1.1 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.

10.1.2 Example: Subpopulation Considerations

To get a better understanding of the fact that the X_i are random variables, consider an election poll in the following setting:

- The total population size is m.
- We sample n people at random.
- In the population, there are d Democrats, r Republicans and o people we'll refer to as Others.

Let D, R and O denote the number of people of the three types that we get in our sample. It would be nice if our sample contained Democrats, Republicans and Others in proportions roughly the same as in the population. In order to see how likely this is to occur, let's find the probability mass function of the random vector (D,R,O)',

$$p_{D,R,O}(i,j,k) = P(D = i, R = j, O = k) \tag{10.1}$$

Case I: Random Sample

Here the X_i are i.i.d., with each one being one of the three categories (Democrat, Republican, Other). Moreover, the random variables D, R and O are the total counts of the number of times each of the three categories occurs. In other words, this is exactly the setting of Section 8.4.1.1, and the random vector (D,R,O)' has a multinomial distribution!

So, we evaluate (10.1) by using (8.84) with

$$p_1 = d/m, \quad p_2 = r/m, \quad p_3 = o/m \tag{10.2}$$

Case I: Simple Random Sample

This is a combinatorial problem, from Section 2.13:

$$P(D = i, R = j, O = k) = \frac{\binom{m}{i} \cdot \binom{m-i}{j}}{\binom{m-i-j}{k}} \tag{10.3}$$

10.1.3 The Sample Mean—a Random Variable

A large part of this chapter will concern the **sample mean**,

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} \tag{10.4}$$

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:

$$\frac{69+69}{2} = 69, \frac{69+72}{2} = 70.5, \frac{69+70}{2} = 69.5, \frac{70+70}{2} = 70, \frac{70+72}{2} = 71, \frac{72+72}{2} = 72 \tag{10.5}$$

The probabilities of these values are 1/9, 2/9, 2/9, 1/9, 2/9 and 1/9, respectively. So,

$$p_{\overline{X}}(69) = \frac{1}{9}, \quad p_{\overline{X}}(70.5) = \frac{2}{9}, \quad p_{\overline{X}}(69.5) = \frac{2}{9}, \quad p_{\overline{X}}(70) = \frac{1}{9}, \quad p_{\overline{X}}(71) = \frac{2}{9}, \quad p_{\overline{X}}(72) = \frac{1}{9} \tag{10.6}$$

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:

$$
\begin{aligned}
E(\overline{X}) &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \quad \text{(def. of } \overline{X}\text{)} && (10.7)\\
&= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) \quad \text{(for const. c, } E(cU) = cEU) && (10.8)\\
&= \frac{1}{n}\sum_{i=1}^{n} EX_i \quad (E[U+V] = EU + EV) && (10.9)\\
&= \frac{1}{n}n\mu \quad (EX_i = \mu) && (10.10)\\
&= \mu && (10.11)
\end{aligned}
$$

$$
\begin{aligned}
Var(\overline{X}) &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] && (10.12)\\
&= \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) \quad \text{(for const. c, , } Var[cU] = c^2 Var[U]) && (10.13)\\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \quad \text{(for U,V indep., , } Var[U+V] = Var[U] + Var[V]) && (10.14)\\
&= \frac{1}{n^2}n\sigma^2 && (10.15)\\
&= \frac{1}{n}\sigma^2 && (10.16)
\end{aligned}
$$

10.1.4 Sample Means Are Approximately Normal–No Matter What the Population Distribution Is

The Central Limit Theorem tells us that the numerator in (10.4) has an approximate normal distribution. That means that affine transformations of that numerator are also approximately normally distributed (page 95). So:

Approximate distribution of (centered and scaled) $\overline{X}$:

The quantity

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{10.17}$$

has an approximately N(0,1) distribution, where σ^2 is the population variance.

Make sure you understand why it is the "N" that is approximate here, not the 0 or 1.

So even if the population distribution is very skewed, multimodal and so on, the sample mean will still have an approximate normal distribution. This will turn out to be the core of statistics.

10.1.5 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

$$Var(X) = \sigma^2 \quad (\sigma^2 \text{ is the population variance}) \tag{10.18}$$

Recall that by definition

$$Var(X) = E[(X - EX)^2] \tag{10.19}$$

Let's estimate $Var(X) = \sigma^2$ by taking sample analogs in (10.19). 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 (10.19) is

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{10.20}$$

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 (10.20).[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 (10.20). Though we will not take that approach here, it will be discussed in Section 12.2.2.

By the way, it can be shown that (10.20) is equal to

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \tag{10.21}$$

This is a handy way to calculate s^2, though it is subject to more roundoff error. Note that (10.21) is a sample analog of (3.29).

10.1.6 A Good Time to Stop and Review!

The material we've discussed in this section, that is since page 211, 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.

10.2 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:

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 2.12.3. Now we will finally address this issue.

As our example, recall from the Bus Paradox in Section 5.3: 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 μ 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

[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.

that we will be pretending that we don't know the mean wait is actually 10. Reminders of this will be brought up occasionally.)

We could write a program to do this:

```
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
```

Note that μ is a population mean, where our "population" here is the set of all possible bus wait times (some more frequent than others). Our simulation, then, drew a sample of size 1000 from that population. The expression **mean(waits)** was our sample mean.

Now, was 1000 iterations enough? How close is this value 9.653743 to the true expected value of waiting time?[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.

10.3 Confidence Intervals for Means

We are now set to make use of the infrastructure that we've built up in the preceding sections of this chapter. Everything will hinge on understand that the sample mean is a random variable, with a known approximate distribution.

The goal of this section (and several that follow) is to develop a notion of margin of error, just as you see in the election campaign polls. This raises two questions:

[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.

(a) What do we mean by "margin of error"?

(b) How can we calculate it?

10.3.1 Confidence Intervals for Population Means

So, suppose we have a random sample $W_1, ..., W_n$ from some population with mean μ and variance σ^2.

Recall that (10.17) has an approximate N(0,1) distribution. We will be interested in the central 95% of the distribution N(0,1). Due to symmetry, that distribution has 2.5% of its area in the left tail and 2.5% in the right one. Through the R call **qnorm(0.025)**, or by consulting a N(0,1) cdf table in a book, we find that the cutoff points are at -1.96 and 1.96. In other words, if some random variable T has a N(0,1) distribution, then $P(-1.96 < T < 1.96) = 0.95$.

Thus

$$0.95 \approx P\left(-1.96 < \frac{\overline{W} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \tag{10.22}$$

(Note the approximation sign.) Doing a bit of algebra on the inequalities yields

$$0.95 \approx P\left(\overline{W} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{W} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{10.23}$$

Now remember, not only do we not know μ, we also don't know σ. But we can estimate it, as we saw, via (10.20). One can show (the details will be given in Section 12.5) that (10.23) is still valid if we substitute s for σ, i.e.

$$0.95 \approx P\left(\overline{W} - 1.96\frac{s}{\sqrt{n}} < \mu < \overline{W} + 1.96\frac{s}{\sqrt{n}}\right) \tag{10.24}$$

In other words, we are about 95% sure that the interval

$$(\overline{W} - 1.96\frac{s}{\sqrt{n}}, \overline{W} + 1.96\frac{s}{\sqrt{n}}) \tag{10.25}$$

contains μ. This is called a 95% **confidence interval** for μ. The quantity $1.96\frac{s}{\sqrt{n}}$ is the margin of error.

10.3.2 Example: Simulation Output

We could add this feature to our program in Section 10.2:

```
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 10.4.

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 10.12.

10.4 Meaning of Confidence Intervals

10.4.1 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]

[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 10.3, 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 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 (10.25) turns out to be (142.6,158.8). We say that we are about 95% confident that the mean weight μ of all adults in Davis is contained in this interval. **What does this mean?**

Say we were to perform this experiment many, many times, recording the results in a notebook: We'd sample 1000 people at random, then record our interval $(\overline{W} - 1.96\frac{s}{\sqrt{n}}, \overline{W} + 1.96\frac{s}{\sqrt{n}})$ on the first line of the notebook. Then we'd sample another 1000 people at random, and record what interval we got that time on the second line of the notebook. This would be a different set of 1000 people (though possibly with some overlap), so we would get a different value of $\overline{W}$ and so, thus a different interval; it would have a different center and a different radius. Then we'd do this a third time, a fourth, a fifth and so on.

Again, each line of the notebook would contain the information for a different random sample of 1000 people. There would be two columns for the interval, one each for the lower and upper bounds. And though it's not immediately important here, note that there would also be columns for W_1 through W_{1000}, the weights of our 1000 people, and columns for $\overline{W}$ and s.

Now here is the point: Approximately 95% of all those intervals would contain μ, the mean weight in the entire adult population of Davis. The value of μ would be unknown to us—once again, that's why we'd be sampling 1000 people in the first place—but it does exist, and it would be contained in approximately 95% of the intervals.

As a variation on the notebook idea, think of what would happen if you and 99 friends each do this experiment. Each of you would sample 1000 people and form a confidence interval. Since each of you would get a different sample of people, you would each get a different confidence interval. What we mean when we say the confidence level is 95% is that of the 100 intervals formed—by you and 99 friends—about 95 of them will contain the true population mean weight. Of course, you hope you yourself will be one of the 95 lucky ones! But remember, you'll never know whose intervals are correct and whose aren't.

Now remember, in practice we only take *one* sample of 1000 people. Our notebook idea here is merely for the purpose of understanding what we mean when we say that we are about 95% confident that one interval we form does contain the true value of μ.

10.4.2 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.

10.5 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 (10.25),

$$\text{margin of error} = 1.96\times \text{ estimated standard deviation of } \overline{W} = 1.96 \times \frac{s}{\sqrt{n}}$$

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 (10.16), 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 (10.25) 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 the quantity
>
> $$\frac{\widehat{\theta} - \theta}{\text{s.e.}(\widehat{\theta})} \tag{10.26}$$
>
> has an approximate N(0,1) distribution.[7]
>
> That means we can mimic the derivation that led to (10.25), showing that an approximate 95% confidence interval for θ is
>
> $$\widehat{\theta} \pm 1.96 \cdot \text{s.e.}(\widehat{\theta}) \tag{10.27}$$
>
> 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.

10.6 Confidence Intervals for Proportions

So we know how to find confidence intervals for means. How about proportions?

[6] The quantity is pronounced "theta-hat." The "hat" symbol is traditional for "estimate of."

[7] This also presumes that $\widehat{\theta}$ is a **consistent** estimator of θ, meaning that $\widehat{\theta}$ converges to θ as $n \to \infty$.

10.6.1 Derivation

It turns out that we already have our answer, from Section 3.6. We found there that proportions are special cases of means: If Y is an indicator random variable with P(Y = 1) = p, then EY = p.

For example, in an election opinion poll, we might be interested in the proportion p of people in the entire population who plan to vote for candidate A. Each voter has a value of Y, 1 if he/she plans to vote for A, 0 otherwise. Then p is the population mean of Y.

We will estimate p by taking a random sample of n voters, and finding $\widehat{p}$, the *sample* proportion of voters who plan to vote for A. Let Y_i be the value of Y for the i^{th} person in our sample. Then

$$\widehat{p} = \overline{Y} \tag{10.28}$$

So, in order to get a confidence interval for p from $\widehat{p}$, we can use (10.25)! We have that an approximate 95% confidence interval for p is

$$\left(\widehat{p} - 1.96s/\sqrt{n}, \widehat{p} + 1.96s/\sqrt{n},\right) \tag{10.29}$$

where as before s^2 is the sample variance among the Y_i.

But there's more, because we can exploit the fact that in this special case, each Y_i is either 1 or 0. Recalling the convenient form of s^2, (10.21), we have

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2 && (10.30)\\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i - \overline{Y}^2 && (10.31)\\ &= \overline{Y} - \overline{Y}^2 && (10.32)\\ &= \widehat{p} - \widehat{p}^2 && (10.33) \end{aligned}$$

Then (10.29) becomes

$$\left(\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}, \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\right) \tag{10.34}$$

And note again that $\sqrt{\widehat{p}(1-\widehat{p})/n}$ is the standard error of $\widehat{p}$.

10.6.2 Simulation Example Again

In our bus example above, suppose we also want our simulation to print out the (estimated) probability that one must wait longer than 6.4 minutes. As before, we'd also like a margin of error for the output.

We incorporate (10.34) 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")
```

When I ran this, the value printed out for $\widehat{p}$ was 0.54, with a margin of error of 0.03, thus an interval of (0.51,0.57). We would say, "We don't know the exact value of $P(W > 6.4)$, so we ran a simulation. The latter estimates this probability to be 0.54, with a 95% margin of error of 0.03."

10.6.3 Examples

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 (10.34) 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 (10.34).

10.6.4 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 10.6.3, 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.

10.6.5 (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.

10.6.6 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.[8] 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.

10.7 Confidence Intervals for Differences of Means or Proportions

10.7.1 Independent Samples

Suppose in our sampling of people in Davis we are mainly interested in the difference in weights between men and women. Let $\overline{X}$ and n_1 denote the sample mean and sample size for men, and let $\overline{Y}$ and n_2 for the women. Denote the population means and variances by μ_i and σ_i^2, i = 1,2. We wish to find a confidence interval for $\mu_1 - \mu_2$. The natural estimator for that quantity is $\overline{X} - \overline{Y}$.

So, how can we form a confidence interval for $\mu_1 - \mu_2$ using $\overline{X} - \overline{Y}$? Since the latter quantity is composed of sums, we can use (10.27). 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

[8] Use calculus to find the maximum value of f(x) = x(1-x).

$$\begin{aligned}
\text{std.dev.}(\overline{X}-\overline{Y}) &= \sqrt{Var[\overline{X}-\overline{Y}]} \quad \text{(def.)} && (10.35)\\
&= \sqrt{Var[\overline{X}+(-1)\overline{Y}]} \quad \text{(algebra)} && (10.36)\\
&= \sqrt{Var(\overline{X})+Var[(-1)\overline{Y}]} \quad \text{(indep.)} && (10.37)\\
&= \sqrt{Var(\overline{X})+Var(\overline{Y})} \quad \text{(3.32.)} && (10.38)\\
&= \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \quad (10.16) && (10.39)
\end{aligned}$$

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,

$$s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2 \tag{10.40}$$

and

$$s_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2 \tag{10.41}$$

we finally have

$$\text{s.e.}(\overline{X}-\overline{Y}) = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \tag{10.42}$$

Thus (10.27) tells us that an approximate 95% confidence interval for $\mu_1-\mu_2$ is

$$\left(\overline{X}-\overline{Y}-1.96\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}, \overline{X}-\overline{Y}+1.96\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right) \tag{10.43}$$

What about confidence intervals for the difference in two population proportions p_1-p_2? Recalling that in Section 10.6 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 (10.43). 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, (10.43) reduces to

$$\left(\widehat{p}_1 - \widehat{p}_2 - 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \widehat{p}_1 - \widehat{p}_2 + 1.96\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_2}} \right) \tag{10.44}$$

10.7.2 Example: Network Security Application

In a network security application, C. Mano *et al*[9] 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

$$11.520 - 2.000 \pm 1.96\sqrt{\frac{9.939^2}{344} + \frac{6.299^2}{436}} = 9.52 \pm 1.22 \tag{10.45}$$

So you can see that there is a big difference between the two networks, even after allowing for sampling variation.

10.7.3 Dependent Samples

Note carefully, though, that a key point above was the independence of the two samples. By contrast, suppose we wish, for instance, to find a confidence interval for $\nu_1 - \nu_2$, the difference in

[9] 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.

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 (10.43).

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 (10.43), 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 (10.25) on these values, so that our approximate 95% confidence interval is

$$(\overline{T} - 1.96\frac{s}{\sqrt{n}}, \overline{T} + 1.96\frac{s}{\sqrt{n}}) \tag{10.46}$$

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 (10.44). 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 (7.10):

$$Var(\widehat{p}_A - \widehat{p}_B) = Var(\widehat{p}_A) + Var(\widehat{p}_B) - 2Cov(\widehat{p}_A, \widehat{p}_B) \tag{10.47}$$

Or, we could have taken a matrix approach, using (7.50) with A equal to the row vector (1,-1,0).

So, using (8.103), the standard error of $\widehat{p}_A - \widehat{p}_B$ is

$$\sqrt{0.001\widehat{p}_A(1-\widehat{p}_A) + 0.001\widehat{p}_B(1-\widehat{p}_B) + 0.002\widehat{p}_A\widehat{p}_B} \tag{10.48}$$

10.7.4 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 (10.43), we have that the interval is

$$223.8 - 226.3 \pm 1.96\sqrt{\frac{15.3^2}{226} + \frac{14.3^2}{585}} = -2.5 \pm 2.3 = (-4.8, -0.3) \tag{10.49}$$

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 11.8.

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

$$\widehat{p}_1 - \widehat{p}_2 = 0.226 - 0.585 = -0.359 \tag{10.50}$$

The standard error of this quantity, from (10.48), is

$$\sqrt{0.001 \cdot 0.226 \cdot 0.7740.001 \cdot 0.585 \cdot 0.415 + 002 \cdot 0.226 \cdot 0.585} = 0.019 \tag{10.51}$$

That gives us a confidence interval of

$$-0.359 \pm 1.96 \cdot 0.019 = (-0.397, -0.321) \tag{10.52}$$

10.8 R Computation

The R function **t.test()** forms confidence intervals for a single mean or for the difference of two means. In the latter case, the two samples must be independent; otherwise, do the single-mean CI on differences, as in Section 10.7.3.

This function uses the Student-t distribution, rather than the normal, but as discussed in Section 10.12, the difference is negligible except in small samples.

10.9 Example: Amazon Links

This example involves the Amazon product co-purchasing network, March 2 2003. The data set is large but simple. It stores a directed graph of what links to what: If a record show i then j, it means that i is often co-purchased with j (though not necessarily vice versa). Let's find a confidence interval for the mean number of inlinks, i.e. links into a node.

Actually, even the R manipulations are not so trivial, so here is the complete code (`http://snap.stanford.edu/data/amazon0302.html`):

```
1 mzn <- read.table("amazon0302.txt",header=F)
2 # cut down the data set for convenience
3 mzn1000 <- mzn[mzn[1.] <= 1000 & mzn[,2] <= 1000,]
4 # make an R list, one element per value of j
5 degrees1000 <- split(mzn1000,mzn1000[,2])
6 # by finding the number of rows in each matrix, we get the numbers of
7 # inlinks
8 indegrees1000 <- sapply(degrees1000,nrow)
```

Now run **t.test()**:

```
> t.test(indegrees1000)

        One Sample t-test

data:  indegrees1000
t = 35.0279, df = 1000, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3.728759 4.171340
sample estimates:
mean of x
  3.95005
```

10.10 The Multivariate Case

In the last few sections, the standard error has been key to finding confidence intervals for univariate quantities. Recalling that the standard error, squared, is the estimated variance of our estimator, one might guess that the analogous quantity in the multivariate case is the estimator covariance matrix of the estimator. This turns out to be correct.

10.10.1 Sample Mean and Sample Covariance Matrix

Say W is an r-element random vector, and we have a random sample $W_1, ..., W_n$ from the distribution of W.

In analogy with (10.20), we have

$$\widehat{Cov}(W) = \frac{1}{n}\sum_{i=1}^{n}(W_i - \overline{W})(W_i - \overline{W})' \tag{10.53}$$

where

$$\overline{W} = \frac{1}{n}\sum_{i=1}^{n} W_i \tag{10.54}$$

Note that (10.53) and (10.54) are of size r x r and r x 1, and are estimators of Cov(W) and EW.

Note too that (10.54) is a sum, thus reminding us of the Central Limit Theorem. In this case it's the Multivariate Central Limit Theorem, which implies that $\overline{W}$ has an approximate multivariate normal distribution. If you didn't read that chapter, the key content is the following:

Let

$$\begin{pmatrix} c_1 \\ ... \\ c_r \end{pmatrix} \tag{10.55}$$

denote any constant (i.e. nonrandom) r-element vector. Then the quantity

$$c'\overline{W} \tag{10.56}$$

has an approximate normal distribution with mean c EW and variance

$$c'\ Cov(\overline{W})c \tag{10.57}$$

An approximate 95% confidence interval for $c_1 EW_1 + ... + c_r EW_r$ is then

$$c'\overline{W} \pm 1.96\sqrt{c'\widehat{Cov}(\overline{W})c} \tag{10.58}$$

where the estimated covariance matrix is given in (10.53).

More generally, here is the extension of the material in Section 10.5:

> Suppose we are estimating some r-component vector θ, using an approximately r-variate normal estimator $\widehat{\theta}$. Let C denote the estimated covariance matrix of $\widehat{\theta}$. Then an approximate 95% confidence interval for ’*theta* is

$$c'\widehat{\theta} \pm 1.96\sqrt{c'Cc} \tag{10.59}$$

10.10.2 Growth Rate Example

Suppose we are studying children’s growth patterns, and have data on heights at ages 6, 10 and 18, denoted (X,Y,Z) = W. We’re interested in the growths between 6 and 10, and between 10 and 18, denoted by G_1 and G_2, respectively. Say we wish to form a confidence interval for $EG_2 - EG_1$, based on a random sample $W_i = (X_i, Y_i, Z_i), i = 1, ..., n$.

This fits right into the context of the previous section. We’re interested in

$$(EZ - EY) - (EY - EX) = EZ - 2EY + EX \tag{10.60}$$

So, we can set c = (1,-2,1) in (10.55), and then use (10.58).

10.11 Advanced Topics in Confidence Intervals

10.12 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

$$T = \frac{\overline{W} - \mu}{\tilde{s}/\sqrt{n}} \tag{10.61}$$

where $\tilde{s}^2$ is the version of the sample variance in which we divide by n-1 instead of by n, i.e. (10.21).

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 (10.22), 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:

$$(\overline{W} - 2.26\frac{\tilde{s}}{\sqrt{10}}, \overline{W} + 2.26\frac{\tilde{s}}{\sqrt{10}}) \tag{10.62}$$

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.

10.13 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 (10.25),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 $\alpha = 0.05$.

10.14 Real Populations and Conceptual Populations

In our example in Section 10.4.1, 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 14.2, 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.[10]

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.

10.15 One More Time: Why Do We Use Confidence Intervals?

After all the variations on a theme in the very long Section 10.2, 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.

Exercises

1. Consider Equation (10.24). In each of the entries in the table below, fill in either R for random, or NR for nonrandom:

[10] 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 14.

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

$$U = \frac{2W_1 + W_2 + ... + W_n}{n+1} \tag{10.63}$$

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 (10.24). 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 (10.25), 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 $p_2 - (p_1 + p_3)$.

10. 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 (10.25) 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.

11. In the example of the population of three people, page 214, 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)

12. In the derivation (10.11), 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.

13. Consider a toy example in which we take a random sample of size 2 (done with replacement) from a population of size 2. The two values in the population (say heights in some measure system) are 40 and 60. Find $p_{s^2}(100)$.

Chapter 11

Introduction to Significance Tests

Suppose (just for fun, but with the same pattern as in more serious examples) you have a coin that will be flipped at the Super Bowl to see who gets the first kickoff. (We'll assume slightly different rules here. The coin is not "called." Instead, it is agreed beforehand that if the coin comes up heads, Team A will get the kickoff, and otherwise it will be Team B.) You want to assess for "fairness." Let p be the probability of heads for the coin.

You could toss the coin, say, 100 times, and then form a confidence interval for p using (10.34). The width of the interval would tell you the margin of error, i.e. it tells you whether 100 tosses were enough for the accuracy you want, and the location of the interval would tell you whether the coin is "fair" enough.

For instance, if your interval were (0.49,0.54), you might feel satisfied that this coin is reasonably fair. In fact, **note carefully that even if the interval were, say, (0.502,0.506), you would still consider the coin to be reasonably fair**; the fact that the interval did not contain 0.5 is irrelevant, as the entire interval would be reasonably near 0.5.

However, this process would not be the way it's traditionally done. Most users of statistics would use the toss data to test the **null hypothesis**

$$H_0 : p = 0.5 \tag{11.1}$$

against the **alternate hypothesis**

$$H_A : p \neq 0.5 \tag{11.2}$$

For reasons that will be explained below, this procedure is called **significance testing**. It forms

the very core of statistical inference as practiced today. This, however, is unfortunate, as there are some serious problems that have been recognized with this procedure. We will first discuss the mechanics of the procedure, and then look closely at the problems with it in Section 11.8.

11.1 The Basics

Here's how significance testing works.

The approach is to consider H_0 "innocent until proven guilty," meaning that we assume H_0 is true unless the data give strong evidence to the contrary. **KEEP THIS IN MIND!**—we are continually asking, "What if...?"

The basic plan of attack is this:

> We will toss the coin n times. Then we will believe that the coin is fair unless the number of heads is "suspiciously" extreme, i.e. much less than n/2 or much more than n/2.

Let p denote the true probability of heads for our coin. As in Section 10.6.1, let $\widehat{p}$ denote the proportion of heads in our sample of n tosses. We observed in that section that $\widehat{p}$ is a special case of a sample mean (it's a mean of 1s and 0s). We also found that the standard deviation of $\widehat{p}$ is $\sqrt{p(1-p)/n}$.[1]

In other words,

$$\frac{\widehat{p}-p}{\sqrt{\frac{1}{n}\cdot p(1-p)}} \tag{11.3}$$

has an approximate N(0,1) distribution.

But remember, we are going to assume H_0 for now, until and unless we find strong evidence to the contrary. Thus we are assuming, for now, that the **test statistic**

$$Z=\frac{\widehat{p}-0.5}{\sqrt{\frac{1}{n}\cdot 0.5(1-0.5)}} \tag{11.4}$$

has an approximate N(0,1) distribution.

[1] This is the exact standard deviation. The estimated standard deviation is $\sqrt{\widehat{p}(1-\widehat{p})/n}$.

Now recall from the derivation of (10.25) that -1.96 and 1.96 are the lower- and upper-2.5% points of the N(0,1) distribution. Thus,

$$P(Z < -1.96 \text{ or } Z > 1.96) \approx 0.05 \tag{11.5}$$

Now here is the point: After we collect our data, in this case by tossing the coin n times, we compute $\widehat{p}$ from that data, and then compute Z from (11.4). If Z is smaller than -1.96 or larger than 1.96, we reason as follows:

> Hmmm, Z would stray that far from 0 only 5% of the time. So, either I have to believe that a rare event has occurred, or I must abandon my assumption that H_0 is true.

For instance, say n = 100 and we get 62 heads in our sample. That gives us Z = 2.4, in that "rare" range. We then **reject** H_0, and announce to the world that this is an unfair coin. We say, "The value of p is significantly different from 0.5."

The 5% "suspicion criterion" used above is called the **significance level**, typically denoted α. One common statement is "We rejected H_0 at the 5% level."

On the other hand, suppose we get 47 heads in our sample. Then Z = -0.60. Again, taking 5% as our significance level, this value of Z would not be deemed suspicious, as it occurs frequently. We would then say "We accept H_0 at the 5% level," or "We find that p is not significantly different from 0.5."

The word *significant* is misleading. It should NOT be confused with *important*. It simply is saying we don't believe the observed value of Z is a rare event, which it would be under H_0; we have instead decided to abandon our believe that H_0 is true.

11.2 General Testing Based on Normally Distributed Estimators

In Section 10.5, we developed a method of constructing confidence intervals for general approximately normally distributed estimators. Now we do the same for significance testing.

Suppose $\widehat{\theta}$ is an approximately normally distributed estimator of some population value θ. Then to test $H_0 : \theta = c$, form the test statistic

$$Z = \frac{\widehat{\theta} - c}{s.e.(\widehat{\theta})} \tag{11.6}$$

where $s.e.(\widehat{\theta})$ is the standard error of $\widehat{\theta}$,[2] and proceed as before:

> Reject $H_0 : \theta = c$ at the significance level of $\alpha = 0.05$ if $|Z| \geq 1.96$.

11.3 Example: Network Security

Let's look at the network security example in Section 10.7.1 again. Here $\widehat{\theta} = \overline{X} - \overline{Y}$, and c is presumably 0 (depending on the goals of Mano *et al*). From 10.42, the standard error works out to 0.61. So, our test statistic (11.6) is

$$Z = \frac{\overline{X} - \overline{Y} - 0}{0.61} = \frac{11.52 - 2.00}{0.61} = 15.61 \tag{11.7}$$

This is definitely larger in absolute value than 1.96, so we reject H_0, and conclude that the population mean round-trip times are different in the wired and wireless cases.

11.4 The Notion of "p-Values"

Recall the coin example in Section 11.1, in which we got 62 heads, i.e. Z = 2.4. Since 2.4 is considerably larger than 2.4, and our cutoff for rejection was only 1.96, we might say that in some sense we not only rejected H_0, we actually strongly rejected it.

To quantify that notion, we compute something called the **observed significance level**, more often called the **p-value**.

We ask,

> We rejected H_0 at the 5% level. Clearly, we would have rejected it even at some small—thus more stringent—levels. What is the smallest such level?

By checking a table of the N(0,1) distribution, or by calling **pnorm(2.40)** in R, we would find that the N(0,1) distribution has area 0.008 to the right of 2.40, and of course by symmetry there is an equal area to the left of -2.40. That's a total area of 0.016. In other words, we would have been able to reject H_0 even at the much more stringent significance level of 0.016 (the 1.6% level)instead of 0.05. So, Z = 2.40 would be considered even more significant than Z = 1.96. In the research

[2] See Section 10.5. Or, if we know the exact standard deviation of $\widehat{\theta}$ under H_0, which was the case in our coin example above, we could use that, for a better normal approximation.

community it is customary to say, "The p-value was 0.016." [3] The smaller the p-value, the more significant the results are considered.

In our network security example above in which Z was 15.61, the value is literally "off the chart"; **pnorm(15.61)** returns a value of 1. Of course, it's a tiny bit less than 1, but it is so far out in the right tail of the N(0,1) distribution that the area to the right is essentially 0. So the p-value would be essentially 0, and the result would be treated as very, very highly significant.

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.

11.5 R Computation

The R function **t.test()**, discussed in Section 10.8, does both confidence intervals and tests, including p-values in the latter case.

11.6 One-Sided H_A

Suppose that—somehow—we are sure that our coin in the example above is either fair or it is more heavily weighted towards heads. Then we would take our alternate hypothesis to be

$$H_A : p > 0.5 \tag{11.8}$$

A "rare event" which could make us abandon our belief in H_0 would now be if Z in (11.4) is very large in the positive direction. So, with $\alpha = 0.05$, we call **qnorm(0.95)**, and find that our rule would now be to reject H_0 if $Z > 1.65$.

One-sided tests are not common, as their assumptions are often difficult to justify.

11.7 Exact Tests

Remember, the tests we've seen so far are all approximate. In (11.4), for instance, $\widehat{p}$ had an approximate normal distribution, so that the distribution of Z was approximately N(0,1). Thus the

[3] 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.

significance level α was approximate, as were the p-values and so on.[4]

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 (11.8) . To keep things simple, let's suppose we toss the coin 10 times. We will make our decision based on X, the number of heads out of 10 tosses. Suppose we set our threshhold for "strong evidence" again H_0 to be 8 heads, i.e. we will reject H_0 if $X \geq 8$. What will α be?

$$\alpha = \sum_{i=8}^{10} P(X = i) = \sum_{i=8}^{10} \left(\begin{array}{c} 10 \\ i \end{array} \right) \left(\frac{1}{2} \right)^{10} = 0.055 \tag{11.9}$$

That's not the usual 0.05. Clearly we cannot get an exact significance level of 0.05,[5] 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

$$H_0 : \mu = 1000 \tag{11.10}$$

[4] 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.

[5] Actually, it could be done by introducing some randomization to our test.

vs.

$$H_A : \mu > 1000 \tag{11.11}$$

It is natural to have our test take the form in which we reject H_0 if

$$\overline{X} > w \tag{11.12}$$

for some constant w chosen so that

$$P(\overline{X} > w) = 0.05 \tag{11.13}$$

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.

11.8 What's Wrong with Significance Testing—and What to Do Instead

The first principle is that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists.—Richard Feynman, Nobel laureate in physics

"Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path"—Paul Meehl, professor of psychology and the philosophy of science

Significance testing is a time-honored approach, used by tens of thousands of people every day. But it is "wrong." I use the quotation marks here because, although significance testing is mathematically correct, it is at best noninformative and at worst seriously misleading.

11.8.1 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 physics Nobel laureate Richard Feynman and economics Nobelists Kenneth Arrow and Milton Friedman.

- See `http://www.indiana.edu/~stigtsts/quotsagn.html` for a nice collection of quotes from famous statisticians on this point.

- There is an entire chapter devoted to this issue in one of the best-selling elementary statistics textbooks in the nation.[6]

- 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.

11.8.2 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.

[6] *Statistics*, third edition, by David Freedman, Robert Pisani, Roger Purves, pub. by W.W. Norton, 1997.

Consider the coin example, for instance. No coin is absolutely perfectly balanced, and yet that is the question that significance testing is asking:

$$H_0 : p = 0.5000000000000000000000000000... \tag{11.14}$$

We know before even collecting any data that the hypothesis we are testing is false, and thus it's nonsense to test it.

But much worse is this word "significant." Say our coin actually has p = 0.502. From anyone's point of view, that's a fair coin! But look what happens in (11.4) as the sample size n grows. If we have a large enough sample, eventually the denominator in (11.4) will be small enough, and $\widehat{p}$ will be close enough to 0.502, that Z will be larger than 1.96 and we will declare that p is "significantly" different from 0.5. But it isn't! Yes, 0.502 is different from 0.5, but NOT in any significant sense in terms of our deciding whether to use this coin in the Super Bowl.

The same is true for government testing of new pharmaceuticals. We might be comparing a new drug to an old drug. Suppose the new drug works only, say, 0.4% (i.e. 0.004) better than the old one. Do we want to say that the new one is "signficantly" better? This wouldn't be right, especially if the new drug has much worse side effects and costs a lot more (a given, for a new drug).

Note that in our analysis above, in which we considered what would happen in (11.4) as the sample size increases, we found that eventually *everything* becomes "signficiant"—even if there is no practical difference. This is especially a problem in computer science applications of statistics, because they often use very large data sets. A data mining application, for instance, may consist of hundreds of thousands of retail purchases. The same is true for data on visits to a Web site, network traffic data and so on. In all of these, the standard use of significance testing can result in our pouncing on very small differences that are quite insignificant to us, yet will be declared "significant" by the test.

Conversely, if our sample is too small, we can miss a difference that actually *is* significant—i.e. important to us—and we would declare that p is NOT significantly different from 0.5. In the example of the new drug, this would mean that it would be declared as "not significantly better" than the old drug, even if the new one is much better but our sample size wasn't large enough to show it.

In summary, the basic problems with significance testing are

- H_0 is improperly specified. What we are really interested in here is whether p is *near* 0.5, not whether it is *exactly* 0.5 (which we know is not the case anyway).
- Use of the word *significant* is grossly improper (or, if you wish, grossly misinterpreted).

Significance testing forms the very core usage of statistics, yet you can now see that it is, as I said

above, "at best noninformative and at worst seriously misleading." This is widely recognized by thinking statisticians and prominent scientists, as noted above. But the practice of significance testing is too deeply entrenched for things to have any prospect of changing.

11.8.3 You Be the Judge!

This book has been written from the point of view that every educated person should understand statistics. It impacts many vital aspects of our daily lives, and many people with technical degrees find a need for it at some point in their careers.

In other words, statistics is something to be *used*, not just learned for a course. You should think about it critically, especially this material here on the problems of significance testing. You yourself should decide whether the latter's widespread usage is justified.

11.8.4 What to Do Instead

Note carefully that I am not saying that we should not make a decision. We *do* have to decide, e.g. decide whether a new hypertension drug is safe or in this case decide whether this coin is "fair" enough for practical purposes, say for determining which team gets the kickoff in the Super Bowl. But it should be an informed decision, and even testing the modified H_0 above would be much less informative than a confidence interval.

In fact, the real problem with significance tests is that they **take the decision out of our hands**. They make our decision mechanically for us, not allowing us to interject issues of importance to us, such possible side effects in the drug case.

So, what can we do instead?

In the coin example, we could set limits of fairness, say require that p be no more than 0.01 from 0.5 in order to consider it fair. We could then test the hypothesis

$$H_0 : 0.49 \leq p \leq 0.51 \tag{11.15}$$

Such an approach is almost never used in practice, as it is somewhat difficult to use and explain. But even more importantly, what if the true value of p were, say, 0.51001? Would we still really want to reject the coin in such a scenario?

Forming a confidence interval is the far superior approach. The width of the interval shows us whether n is large enough for $\widehat{p}$ to be reasonably accurate, and the location of the interval tells us whether the coin is fair enough for our purposes.

Note that in making such a decision, we do NOT simply check whether 0.5 is in the interval. That would make the confidence interval reduce to a significance test, which is what we are trying to avoid. If for example the interval is (0.502,0.505), we would probably be quite satisfied that the coin is fair enough for our purposes, even though 0.5 is not in the interval.

On the other hand, say the interval comparing the new drug to the old one is quite wide and more or less equal positive and negative territory. Then the interval is telling us that the sample size just isn't large enough to say much at all.

Significance testing is also used for model building, such as for predictor variable selection in regression analysis (a method to be covered in Chapter 14). 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 13.

11.8.5 Decide on the Basis of "the Preponderance of Evidence"

I was in search of a one-armed economist, so that the guy could never make a statement and then say: "on the other hand"—President Harry S Truman

If all economists were laid end to end, they would not reach a conclusion—Irish writer George Bernard Shaw

In the movies, you see stories of murder trials in which the accused must be "proven guilty beyond the shadow of a doubt." But in most noncriminal trials, the standard of proof is considerably lighter, **preponderance of evidence**. This is the standard you must use when making decisions based on statistical data. Such data cannot "prove" anything in a mathematical sense. Instead, it should be taken merely as evidence. The width of the confidence interval tells us the likely accuracy of that evidence. We must then weigh that evidence against other information we have about the subject being studied, and then ultimately make a decision on the basis of the preponderance of all the evidence.

Yes, juries must make a decision. But they don't base their verdict on some formula. Similarly, you the data analyst should not base your decision on the blind application of a method that is usually of little relevance to the problem at hand—significance testing.

11.8.6 Example: the Forest Cover Data

In Section 10.7.4, we found that an approximate 95% confidence interval for $\mu_1 - \mu_2$ was

$$223.8 - 226.3 \pm 2.3 = (-4.8, -0.3) \tag{11.16}$$

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

$$H_0 : \mu_1 = \mu_2 \tag{11.17}$$

and say that the two means are "significantly" different, which sounds like there is an important difference—which there is not.

11.8.7 Example: Assessing Your Candidate's Chances for Election

Imagine an election between Ms. Smith and Mr. Jones, with you serving as campaign manager for Smith. You've just gotten the results of a very small voter poll, and the confidence interval for p, the fraction of voters who say they'll vote for Smith, is (0.45,0.85). Most of the points in this interval are greater than 0.5, so you would be highly encouraged! You are certainly not sure of the final election result, as a small part of the interval is below 0.5, and anyway voters might change their minds between now and the election. But the results would be highly encouraging.

Yet a significance test would say "There is no significant difference between the two candidates. It's a dead heat." Clearly that is not telling the whole story. The point, once again, is that **the confidence interval is giving you much more information than is the significance test.**

Exercises

1. In the light bulb example on page 246, suppose the actual observed value of $\overline{X}$ turns out to be 15.88. Find the p-value.

Chapter 14

Relations Among Variables: Linear Regression

In many senses, this chapter and the next one form 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.

14.1 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.

[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."

- **Understanding:** Here we wish to determine which of several variables have a greater effect on (or relation to) a given variable. An important special case is that in which we are interested in determining the effect of one predictor variable, **after the effects of the other predictors are removed.**

Denote the **predictor variables** by, $X^{(1)}, ..., X^{(r)}$. 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.

14.2 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 Prasun Sinha, Danny Raz and Nidhan Choudhuri, 2006, the authors wanted to predict Y, the round-trip time (RTT) for packets in a network, using the predictor variables $X^{(1)}$ = geographical distance between the two nodes, $X^{(2)}$ = number of router-to-router hops, and other offline variables. The goal here was primarily Prediction.

Example: In a paper, Productivity Analysis of Object-Oriented Software Developed in a Commercial Environment, *Software—Practice and Experience*, by Thomas E. Potok, Mladen Vouk and Andy Rindos, 1999, the authors mainly had an Understanding goal: What impact, positive or

[2] Sometimes called a **dummy variable**.

negative, does the use of object-oriented programming have on programmer productivity? Here they predicted Y = number of person-months needed to complete the project, from $X^{(1)}$ = size of the project as measured in lines of code, $X^{(2)} = 1$ or 0 depending on whether an object-oriented or procedural approach was used, and other variables.

Example: Most **text mining** applications are classification problems. For example, the paper Untangling Text Data Mining, *Proceedings of ACL'99*, by Marti Hearst, 1999 cites, *inter alia*, an application in which the analysts wished to know what proportion of patents come from publicly funded research. They were using a patent database, which of course is far too huge to feasibly search by hand. That meant that they needed to be able to (reasonably reliably) predict Y = 1 or 0 according to whether the patent was publicly funded from a number of $X^{(i)}$, each of which was an indicator variable for a given key word, such as "NSF." They would then treat the predicted Y values as the real ones, and estimate their proportion from them.

14.3 Adjusting for Covariates

The first statistical consulting engagement I ever worked involved something called *adjusting for covariates*. I was retained by the Kaiser hospital chain to investigate how heart attack patients fared at the various hospitals—did patients have a better chance to survive in some hospitals than in others? There were four hospitals of particular interest.

I could have simply computed raw survival rates, say the proportion of patients who survive for a month following a heart attack, and then used the methods of Section 10.6, for instance. This could have been misleading, though, because one of the four hospitals served a largely elderly population. A straight comparison of survival rates might then unfairly paint that particular hospital as giving lower quality of care than the others.

So, we want to somehow adjust for the effects of age. I did this by setting Y to 1 or 0, for survival, $X^{(1)}$ to age, and X^{2+i} to be an indicator random variable for whether the patient was at hospital i, i = 1,2,3.[3]

14.4 What Does "Relationship" Really Mean?

Consider the Davis city population example again. In addition to the random variable W for weight, let H denote the person's height. Suppose we are interested in exploring the relationship between height and weight.

[3]Note that there is no i = 4 case, since if the first three hospital variables are all 0, that already tells us that this patient was at the fourth hospital.

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

$$m_{W;H}(t) = E(W|H = t) \tag{14.1}$$

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*!

14.5 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

$$(H_1, W_1), ..., (H_{1000}, W_{1000}) \tag{14.2}$$

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.[4] How do we do this?

[4]Of course, the population of Davis is finite, but there is the conceptual population of all people who *could* live in Davis.

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 15.3), 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 13.1.

Typically the parametric model chosen is linear, i.e. we assume that $m_{W;H}(t)$ is a linear function of t:

$$m_{W;H}(t) = ct + d \tag{14.3}$$

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 (14.3) 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. (14.3). Again, the quantities c and d in (14.3) are population values, and as such, we must estimate them from the data.

So, how can we estimate these population values c and d? We'll go into details in Section 14.9, but here is a preview:

Using the result on page 48, together with the Law of Total Expectation in Section 9.1.3, we have

that the minimum value of the quantity

$$E\left[(W - g(H))^2\right] \tag{14.4}$$

overall all possible functions $g(H)$, is attained by setting

$$g(H) = m_{W;H}(H) \tag{14.5}$$

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.[5]

Since we are assuming the model (14.3), this in turn means that:

> The quantity
>
> $$E\left[(W - (rH + s))^2\right] \tag{14.6}$$
>
> is minimized by setting r = c and s = d.

This then gives us a clue as to how to estimate c and d from our data, as follows.

If you recall, in earlier chapters we've often chosen estimators by using sample analogs, e.g. s^2 as an estimator of σ^2. Well, the sample analog of (14.6) is

$$\frac{1}{n}\sum_{i=1}^{n}[W_i - (r + sH_i)]^2 \tag{14.7}$$

Here (14.6) is the mean squared prediction error using r and s in the population, and (14.7) is the mean squared prediction error using r and s in our sample. Since r = d and s = c minimize (14.6), it is natural to estimate d and c by the r and s that minimize (14.7).

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 (14.3), estimating the coefficients by minimizing (14.7).

We will elaborate on this in Section 14.9.

[5] But if we wish to minimize the mean absolute prediction error, $E(|W - g(H)|)$, the best function turns out to be is $g(H) = \text{median}(W|H)$.

14.6 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

$$X = \left(X^{(1)}, X^{(2)}\right) = (H, A) = (\text{height, age}) \tag{14.8}$$

so as to study the relationship of weight with height and age. If we used a linear model, we would write for $t = (t_1, t_2)$,

$$m_{W;H,A}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 \tag{14.9}$$

In other words

$$\text{mean weight} = \beta_0 + \beta_1 \text{ height} + \beta_2 \text{ age} \tag{14.10}$$

(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.

In analogy with (14.7), we would estimate the β_i by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}[W_i - (u + vH_i + wA_i)]^2 \tag{14.11}$$

with respect to u, v and w. The minimizing values would be denoted $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

We might consider adding a third predictor, gender:

$$\text{mean weight} = \beta_0 + \beta_1 \text{ height} + \beta_2 \text{ age} + \beta_3 \text{ gender} \tag{14.12}$$

where **gender** is an indicator variable, 1 for male, 0 for female. Note that we would not have two gender variables, since knowledge of the value of one such variable would tell us for sure what the other one is. (It would also make a certain matrix noninvertible, as we'll discuss later.)

14.7 Interaction Terms

Equation (14.9) 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 (14.9), with the same number for height in both, and subtract; you get $10\beta_2$, an expression that has no height term.

That the assumption is not a good one, since people tend to get heavier as they age. If we don't like this assumption, we can add an **interaction term** to (14.9), 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

$$m_{W;H}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_1 t_2 \tag{14.13}$$

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 14.1. 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 14.10.1). And with so many interaction terms, the model would be difficult to interpret.

So, we may have a decision to make here, as to whether to introduce interaction terms. For that matter, it may be the case that age is actually not that important, so we even might consider dropping that variable altogether. These questions will be pursued in Section 14.10.

14.8 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 15.7.1) is that we predict his weight to be the mean weight for his height/age group,

$$m_{W;H,A}(70.4, 24.8) \tag{14.14}$$

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

$$\widehat{m}_{W;H,A}(70.4, 24.8) \tag{14.15}$$

If our model is (14.9), then (14.15) is

$$\widehat{m}_{W;H}(t) = \widehat{\beta}_0 + \widehat{\beta}_1 70.4 + \widehat{\beta}_2 24.8 \tag{14.16}$$

where the $\widehat{\beta}_i$ are estimated from our data by least-squares.

14.9 Parametric Estimation of Linear Regression Functions

14.9.1 Meaning of "Linear"

Here we model $m_{Y;X}$ as a linear function of $X^{(1)}, ..., X^{(r)}$:

$$m_{Y;X}(t) = \beta_0 + \beta_1 t^{(1)} + ... + \beta_r t^{(r)} \tag{14.17}$$

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 (14.31).

Instead, the word *linear* refers to the regression function being linear in the parameters. So, for instance, (14.31) 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 (14.22) below.

14.9.2 Point Estimates and Matrix Formulation

So, how do we estimate the β_i? Look for instance at (14.31). Keep in mind that in (14.31), the β_i are population values. We need to estimate them from our data. How do we do that? As previewed

in Section 14.5, 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 (14.7), 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:

$$\sum_{i=1}^{100}[A_i - (\hat{\beta}_0 + \hat{\beta}_1 b_i + \hat{\beta}_2 b_i^2)]^2 \tag{14.18}$$

Obviously, this is a calculus problem. We set the partial derivatives of (14.18) with respect to the $\hat{\beta}_i$ to 0, giving use three linear equations in three unknowns, and then solve.

For the general case (14.17), 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 (14.9.1), we have

$$E(Y_i|X_i^{(1)}, ..., X_i^{(r)}) = \beta_0 + \beta_1 X_i^{(1)} + ... + \beta_r X_i^{(r)},\ i = 1, ..., n \tag{14.19}$$

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:

$$\begin{pmatrix} 1 & H_1 & A_1 \\ 1 & H_2 & A_2 \\ ... & & \\ 1 & H_{100} & A_{100} \end{pmatrix} \tag{14.20}$$

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

$$V = (Y_1, ..., Y_n)', \tag{14.21}$$

Then the system (14.19) in matrix form is

$$E(V|Q) = Q\beta \tag{14.22}$$

where

$$\beta = (\beta_0, \beta_1, ..., \beta_r)' \tag{14.23}$$

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

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_r)' \tag{14.24}$$

with our goal now being to find $\hat{\beta}$. The matrix form of (14.18) (now for the general case, not just ALOHA) is

$$(V - Q\hat{\beta})'(V - Q\hat{\beta}) \tag{14.25}$$

Then it can be shown that, after all the partial derivatives are taken and set to 0, the solution is

$$\hat{\beta} = (Q'Q)^{-1}Q'V \tag{14.26}$$

By the way, recall in (14.12) we had only one indicator variable for gender, not two. If we were to have variables for both male and female, the corresponding columns in Q would add to the first column. The matrix would then not be of full rank, thus not invertible above.

It turns out that $\hat{\beta}$ is an unbiased estimate of β:[6]

$$\begin{aligned} E\hat{\beta} &= E[(Q'Q)^{-1}Q'V] \quad (14.26) && (14.27)\\ &= (Q'Q)^{-1}Q'EV \quad \text{(linearity of E())} && (14.28)\\ &= (Q'Q)^{-1}Q' \cdot Q\beta \quad (14.22) && (14.29)\\ &= \beta && (14.30) \end{aligned}$$

[6]Note that here we are taking the expected value of a vector. This is covered in Chapter 8.

In some applications, we assume there is no constant term β_0 in (14.17). This means that our Q matrix no longer has the column of 1s on the left end, but everything else above is valid.

14.9.3 Back to Our ALOHA Example

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)
   }
   return(md)
}

# 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 14.1, 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:

```
> md <- alohamain(4,0.1,1000,100)
> plot(md,cex=0.5,xlab="b",ylab="A")
```

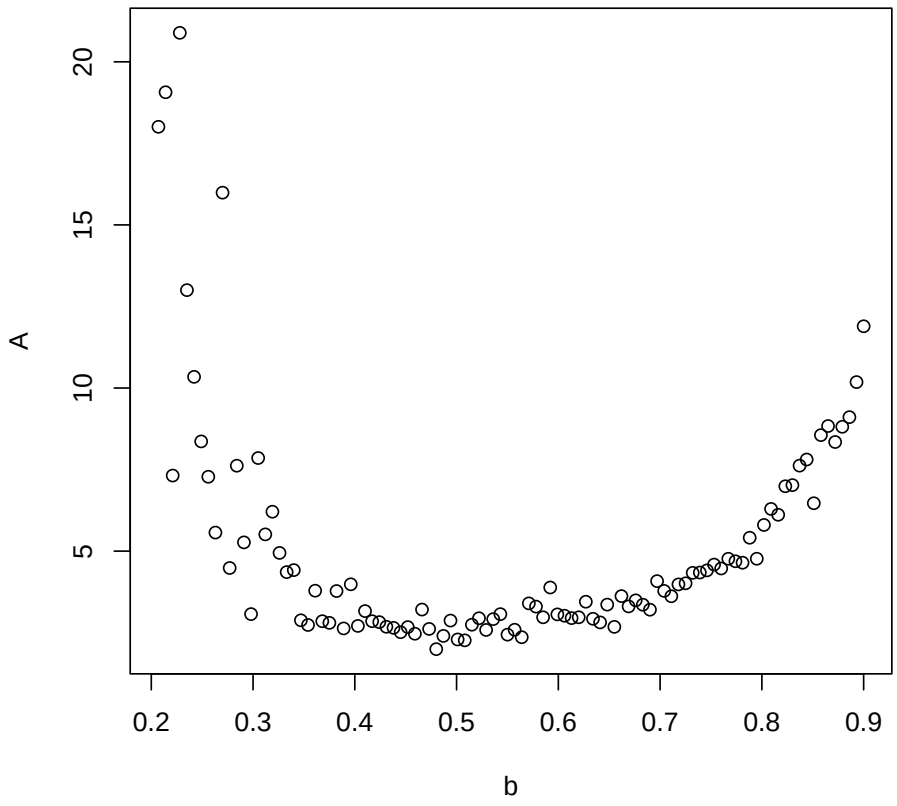


Figure 14.1: Scatter Plot

The plot is shown in Figure 14.1.

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)
```

The plot is shown in Figure 14.2.

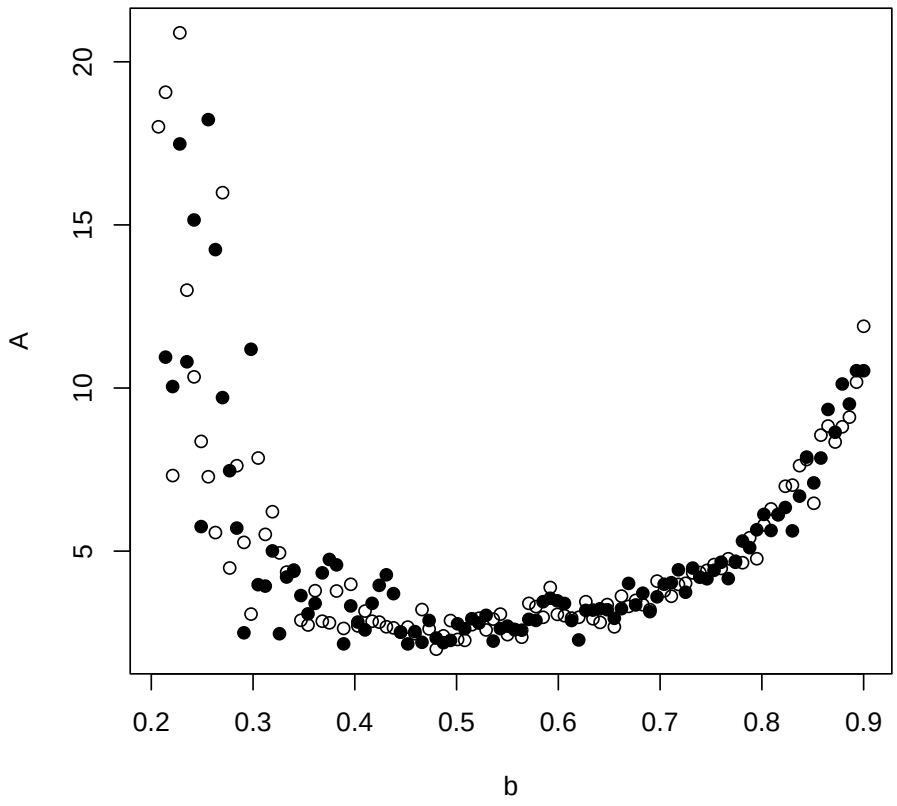


Figure 14.2: Scatter Plot, Two Data Sets

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.

So, a model that expresses mean A to be a linear function of b, as in our height-weight example, is clearly inappropriate. However, you may be surprised to know that we can still use a linear regression model! And this is common. Here are the details:

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

$$m_{A;b}(b) = \beta_0 + \beta_1 b + \beta_2 b^2 \tag{14.31}$$

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.

As mentioned, this is a *linear* model, in the sense that the β_i enter into (14.31) is a linear manner. The fact that that equation is quadratic in b is irrelevant. By the way, one way to look at the degree-2 term is to consider it to model the "interaction" of b with itself.

Again, we'll see how to estimate the β_i in Section 14.9.

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

$$m_{A;b,q}(u, v, w) = \beta_0 + \beta_1 u + \beta_2 u^2 + \beta_3 v + \beta_4 w \tag{14.32}$$

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.[7]

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 $\widehat{\beta}$:

[7] 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.

```
> lmout$coefficients
(Intercept)     md[, 1]     md[, 3]
   27.56852   -90.72585    79.98616
```

So, $\widehat{\beta}_0 = 27.57$ and so on.

The result is

$$\widehat{m}_{A,b}(t) = 27.57 - 90.73t + 79.99t^2 \tag{14.33}$$

(Do you understand why there is a hat about the m?)

Another member variable in the **lm** class is **fitted.values**. This is the "fitted curve," meaning the values of (14.33) at $b_1, ..., b_{100}$. In other words, this is (14.33). I plotted this curve on the same graph,

```
> lines(cbind(md[,1],lmout$fitted.values))
```

See Figure 14.3. 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
```

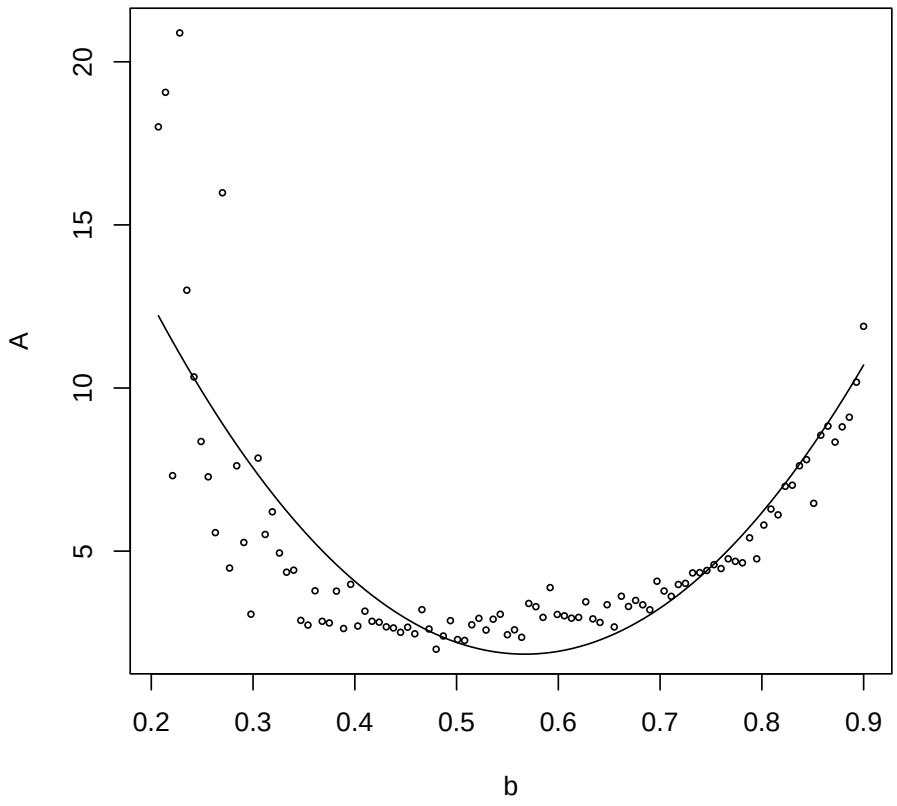


Figure 14.3: Quadratic Fit Superimposed

In other words, we have an estimated regression function of

$$\widehat{m}_{A,b}(t) = 95.98882 - 664.02780\ t + 1731.90848\ t^2 - 1973.00660\ t^3 + 835.89714\ t^4 \qquad (14.34)$$

The fit is shown in Figure 14.4. It looks much better. On the other hand, we have to worry about overfitting. We return to this issue in Section 14.10.1).

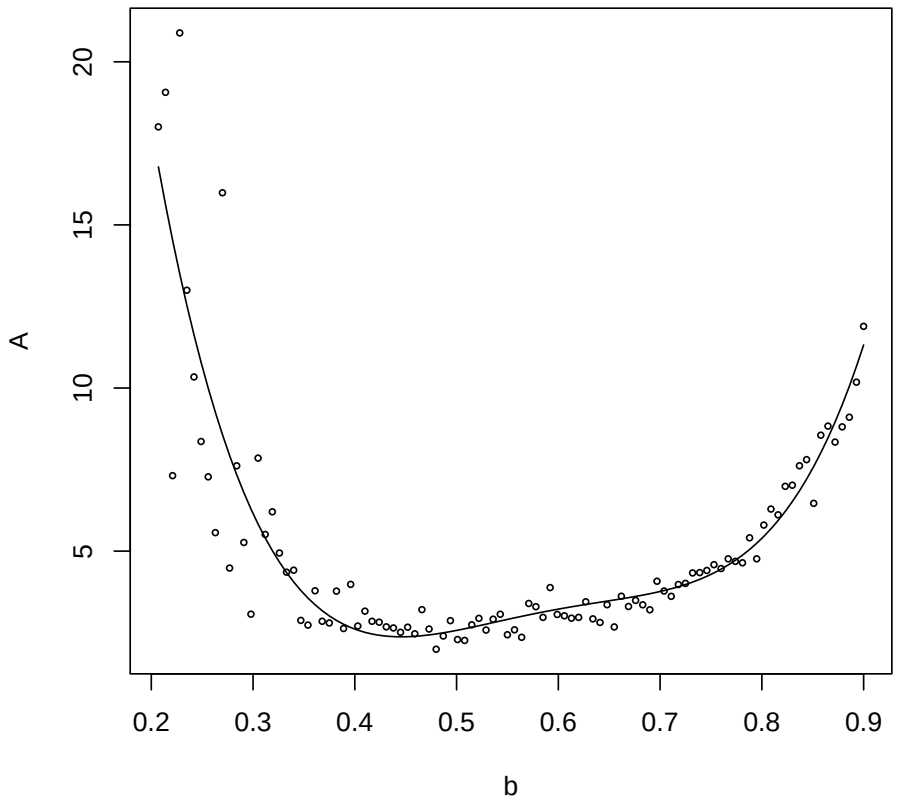


Figure 14.4: Fourth Degree Fit Superimposed

14.9.4 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 14.1. 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 (14.26) shows that the $\widehat{\beta}_i$ are sums of the components of V, i.e. the Y_j. So, the Central

Limit Theorem implies that the $\widehat{\beta}_i$ are approximately normally distributed. That in turn means that, in order to form confidence intervals, we need standard errors for the β_i. How will we get them?

Note carefully that so far we have made NO assumptions other than (14.17). Now, though, we need to add an assumption:[8]

$$Var(Y|X = t) = \sigma^2 \tag{14.35}$$

for all t. Note that this and the independence of the sample observations (e.g. the various people sampled in the Davis height/weight example are independent of each other) implies that

$$Cov(V|Q) = \sigma^2 I \tag{14.36}$$

where I is the usual identiy matrix (1s on the diagonal, 0s off diagonal).

Be sure you understand what this means. In the Davis weights example, for instance, it means that the variance of weight among 72-inch tall people is the same as that for 65-inch-tall people. That is not quite true—the taller group has larger variance—but research into this has found that as long as the discrepancy is not too bad, violations of this assumption won't affect things much.

We can derive the covariance matrix of $\hat{\beta}$ as follows. Again to avoid clutter, let $B = (Q'Q)^{-1}$. A theorem from linear algebra says that Q'Q is symmetric and thus B is too. Another theorem says that for any conformable matrices U and V, then (UV)' = V'U'. Armed with that knowledge, here we go:

$$\begin{aligned}
Cov(\hat{\beta}) &= Cov(BQ'V) \text{ ((14.26))} && (14.37)\\
&= BQ'Cov(V)(BQ')' \text{ (7.50)} && (14.38)\\
&= BQ'\sigma^2 I(BQ')' \text{ (14.36)} && (14.39)\\
&= \sigma^2 BQ'QB \text{ (lin. alg.)} && (14.40)\\
&= \sigma^2 (Q'Q)^{-1} \text{ (def. of B)} && (14.41)
\end{aligned}$$

Whew! That's a lot of work for you, if your linear algebra is rusty. But it's worth it, because (14.41) 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

[8] Actually, we could derive some usable, though messy,standard errors without this assumption.

$$\begin{aligned}
\sigma^2 &= Var(Y|X=t) && (14.42)\\
&= Var(Y|X^{(1)}=t_1,...,X^{(r)}=t_r) && (14.43)\\
&= E\left[\{Y-m_{Y;X}(t)\}^2\right] && (14.44)\\
&= E\left[(Y-\beta_0-\beta_1 t_1-...-\beta_r t_r)^2\right] && (14.45)
\end{aligned}$$

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:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1 X_i^{(1)}-...-\hat{\beta}_r X_i^{(r)})^2 \tag{14.46}$$

As in Chapter 12, 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

$$\widehat{Cov}(\hat{\beta}) = s^2(Q'Q)^{-1} \tag{14.47}$$

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 10.5, included factors like $1/\sqrt{n}$, you might wonder why you don't see such a factor in (14.47).

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.

14.9.5 Once Again, Our ALOHA Example

In R we can obtain (14.47) 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 (14.47) 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

$$835.89714 \pm 1.96 \cdot 142.6 = (556.4, 1115.4) \tag{14.48}$$

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), so we didn't really need to call **vcov()**. But that call can give us more:

Note that we can apply (7.50) to the estimated covariance matrix of $\widehat{\beta}$! Recall our old example of measuring the relation between people's weights and heights,

$$m_{W;H}(t) = \beta_0 + \beta_1 t \tag{14.49}$$

Suppose we estimate β from our data, and wish to find a confidence interval for the mean height of all people of height 70 inches, which is

$$\beta_0 + 70\beta_1 \tag{14.50}$$

Our estimate is

$$\widehat{\beta}_0 + 70\widehat{\beta}_1 \tag{14.51}$$

That latter quantity is

$$(1, 70)\widehat{\beta} \tag{14.52}$$

perfect for (10.59). Thus

$$\widehat{Var}(\widehat{\beta}_0 + 70\widehat{\beta}_1) = (1, 70)C \begin{pmatrix} 1 \\ 70 \end{pmatrix} \tag{14.53}$$

where C is the output from **vcov()**. The square root of this is then the standard error for (14.51). (Recall Section 10.5.)

14.9.6 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.

This issue is similar to that of Section 10.12. If we do make such a normality 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.

14.10 Model Selection

The issues raised in Chapter 13 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 fit the data well. Hence the Einstein quote in Chapter 13! Simpler models are often called **parsimonious**.

Here I use the term *model selection* to mean which predictor variables (including powers and interactions) 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:

14.10.1 The Overfitting Problem in Regression

Recall (14.31). There we assumed a second-degree polynomial for $m_{A;b}$. Later we extended it to a fourth-degree model. Why not a fifth-degree, or sixth, 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 13.1.4 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, because we often have so many variables we can use, thus often can make highly complex models.

In the regression context, the phrase "we have enough data to support the model" means (in the parametric model case) we have enough data so that the confidence intervals for the β_i will be reasonably narrow. For fixed n, the more complex the model, the wider the resulting confidence intervals will tend to be.

If we use too many predictor variables,[9], our data is "diluted," by being "shared" by so many β_i. As a result, $Var(\widehat{\beta}_i)$ will be large, with big implications: Whether our goal is Prediction or Understanding, our estimates will be so poor that neither goal is achieved.

On the other hand, if some predictor variable is really important (i.e. its β_i is far from 0), then it may pay to include it, even though the confidence intervals might get somewhat wider.

For example, look at our regression model for A against b in the ALOHA simulation in earlier sections. The relation between A and b was so far from a straight line that we should use at least a quadratic model, even if the sample size is pretty small.

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 14.10.3.

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.

14.10.2 Multicollinearity

In typical applications, the $X^{(i)}$ are correlated with each other, to various degrees. If the correlation is high—a condition termed **multicollinearity**—problems may occur.

Consider (14.26). Suppose one predictor variable were to be fully correlated with another. That would mean that the first is exactly equal to a linear function of the other, which would mean that in Q one column is an exact linear combination of the first column and another column. Then $Q'Q^{-1}$ would not exist.

Well, if one predictor is strongly (but not fully) correlated with another, $(Q'Q)^{-1}$ will exist, but it will be numerically unstable. Moreover, even without numeric roundoff errors, $(Q'Q)^{-1}$ would be very large, and thus (14.41) would be large, giving us large standard errors—not good!

[9] In the ALOHA example above, b, b^2, b^3 and b^4 are separate predictors, even though they are of course correlated.

Thus we have yet another reason to limit our set of predictor variables.

14.10.3 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 claims they have a foolproof way to do this, then 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.

Hypothesis testing:

The most commonly used methods for variable selection use hypothesis testing in one form or another. Typically this takes the form

$$H_0 : \beta_i = 0 \tag{14.54}$$

In the context of (14.10), for instance, a decision as to whether to include age as one of our predictor variables would mean testing

$$H_0 : \beta_2 = 0 \tag{14.55}$$

If we reject H_0, then we use the age variable; otherwise we discard it.

I hope I've convinced the reader, in Sections 11.8 and 13.2.1, 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 (14.54). 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.[10]

[10]They added so many predictors that r became greater than n. However, the problems they found would have

This problem has always been known by thinking statisticians, but the Wharton study certainly dramatized it.

Confidence intervals:

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 (14.48). Granted, the interval was 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 like b^4 is worth using as a predictor.

On the other hand, suppose in the weight/height/age example 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 15.2.3.2.

Predictive ability indicators:

Suppose you have several competing models, some using more predictors, some using fewer. If we had some measure of predictive power, we could decide to use whichever model has the maximum value of that measure. Here are some of the more commonly used methods of this type:

- One such measure is called *adjusted R-squared*. To explain it, we must discuss ordinary R^2 first.

  Let ρ denote the population correlation between actual Y and predicted Y, i.e. the correlation between Y and $m_{Y;X}(X)$, where X is the vector of predictor variables in our model. Then $|\rho|$ is a measure of the power of X to predict Y, but it is traditional to use ρ^2 instead.[11]

  R is then the *sample* correlation between the Y_i and the vectors X_i. The sample R^2 is then an estimate of ρ^2. However, the former is a **biased** estimate—over infinitely many samples, the long-run average value of R^2 is higher than ρ^2. And the worse the overfitting, the greater the bias. Indeed, if we have n-1 predictors and n observations, we get a perfect fit, with $R^2 = 1$, yet obviously that "perfection" is meaningless.

  Adjusted R^2 is a tweaked version of R^2 with less bias. So, in deciding which of several models to use, we might choose the one with maximal adjusted R^2. Both measures are reported when one calls **summary()** on the output of **lm()**.

- 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

been there to a large degree even if r were less than n but r/n was substantial.

[11] That quantity can be shown to be the proportion of variance of Y attributable to X.

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 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 $\hat{\beta}_i$. Here we choose the model with minimal AIC.

  The R statistical package includes a function **AIC()** for this, which is used by **step()** in the regression case.

14.10.4 A Rough Rule of Thumb

A rough rule of thumb is that one should have $r < \sqrt{n}$, where r is the number of predictors.[12]

14.11 Nominal Variables

Recall our example in Section 14.2 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. 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.

[12] Asymptotic Behavior of Likelihood Methods for Exponential Families When the Number of Parameters Tends to Infinity, Stephen Portnoy, *Annals of Statistics*, 1968.

14.12 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. (14.35)], 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.

14.13 Case Study: Prediction of Network RTT

Recall the paper by Raz *et al*, introduced in Section 14.2. They wished to predict network round-trip travel time (RTT) from offline variables. Now that we know how regression analysis works, let's look at some details of that paper.

First, they checked for multicollinearity. one measure of that is the ratio of largest to smallest eigenvalue of the matrix of correlations among the predictors. A rule of thumb is that there are problems if this value is greater than 15, but they found it was only 2.44, so they did not worry about multicollinearity.

They took a *backwards stepwise* approach to predictor variable selection, meaning that they started with all the variables, and removed them one-by-one while monitoring a goodness-of-fit criterion. They chose AIC for the latter.

Their initial predictors were DIST, the geographic distance between source and destination node, HOPS, the number of network hops (router processing) and an online variable, AS, the number of **autonomous systems**—large network routing regions—a message goes through. They measured the latter using the network tool **traceroute**.

But AS was the first variable they ended up eliminating. They found that removing it increased AIC only slightly, from about 12.6 million to 12.9 million, and reduced R^2 only a bit, from 0.785 to 0.778. They decided that AS was expendable, especially since they were hoping to use only offline variables.

Based on a scatter plot of RTT versus DIST, they then decided to try adding a quadratic term in that variable. This increased R^2 substantially, to 0.877. So, the final prediction equation they settled on predicts RTT from a quadratic function of DIST and a linear term for HOPS.

14.14 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

$$Y = \beta_0 + \beta_1 X + \epsilon,\ E\epsilon = 0 \tag{14.56}$$

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

$$m_{Y;X}(t) = E(Y|X = t) = \beta_0 + \beta_1 t \tag{14.57}$$

Note that in our context, we could define ϵ as

$$\epsilon = Y - m_{Y;X}(X) \tag{14.58}$$

Equation (14.56) is consistent with (14.57): The former has $E\epsilon = 0$, and so does the latter, since

$$E\epsilon = EY - E[m_{Y;X}(X)] = EY - E[E(Y|X)] = EY - EY = 0 \tag{14.59}$$

In order to produce confidence intervals, we later added the assumption (14.35), which you can see is consistent with (14.56) since the latter assumes that $Var(\epsilon) = \sigma^2$ no matter what value X has.

Now, what about the normality assumption in (14.56)? 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.

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 i^{th} row and j^{th} column. Then the determinant can be computed recursively across the k^{th} 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 quick 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, you may wish to consider ESS for Emacs, StatET for Eclipse or RStudio, all open source.

R is normally run in interactive mode, with > as the prompt. Among other things, that makes it easy to try little experiments to learn from; remember my slogan, "When in doubt, try it out!"

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. 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.[1]

```
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
14    }
```

[1] The source code for this file is at `http://heather.cs.ucdavis.edu/~matloff/MiscPLN/R5MinIntro.tex`.

```
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 (in the other window) 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 is 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)
54  $odds
```

```
55 [1]  5 13
56
57 $numodds
58 [1] 2
59
60 > ocy <- oddcount(y)
61 > ocy
62 $odds
63 [1]  5 13
64
65 $numodds
66 [1] 2
67
68 > ocy$odds
69 [1]  5 13
70 > ocy[[1]]
71 [1]  5 13
72 > ocy[[2]]
73 [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:

```
1 oddcount <- function(x)  {
2     x1 <- x[x %% 2 == 1]
3     return(list(odds=x1, numodds=length(x1)))
4 }
```

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:

```
1 > x <- runif(1000000)  # 1000000 random numbers from the interval (0,1)
2 > system.time(sum(x))
3    user  system elapsed
4   0.008   0.000   0.006
5 > system.time({s <- 0; for (i in 1:1000000) s <- s + x[i]})
6    user  system elapsed
7   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.