From Algorithms to Z-Scores:

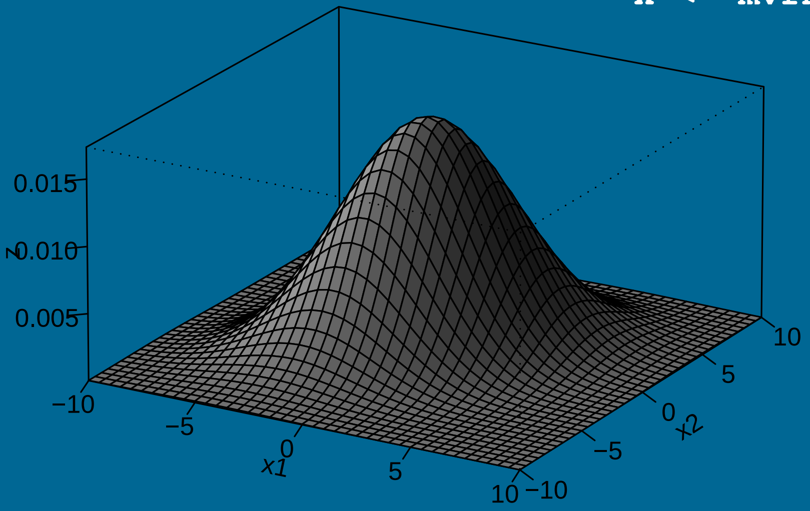
Probabilistic and Statistical Modeling in Computer Science

Norm Matloff

University of California, Davis

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

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



Contents

Preface

Why is this book different from all other books on 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

Discrete Probability Models

1.1 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, 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.

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

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

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

- both nodes try to send; this has probability p^2

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

Thus

$$P(X_1 = 2) = p^2 + (1-p)^2 = 0.52 \tag{1.1}$$

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

Table 1.1: Sample Space for the Dice Example

1.2 Basic Ideas of Probability

1.2.1 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 consists of the possible outcomes (X, Y), seen in Table 1.1. 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."

Though the notion of sample space is presented in every probability textbook, and is central to the advanced theory of probability, 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.

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

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

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 1.2: Notebook for the Dice Problem

The first 9 lines of the notebook might look like Table 1.2. 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.

1.2.2 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 proportion 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 proportion of lines with Yes entries in the new column labeled "A and B."[1]

- For any events A, B, imagine a new column in our notebook, labeled "A or B." In each line, this column will say Yes if and only if at least one of the entries for A and B says Yes.[2]

- For any events A, B, imagine a new column in our notebook, labeled "A | B" and pronounced "A given B." In each line:

  - This new column will say "NA" ("not applicable") if the B entry is No.
  - If it is a line in which the B column says Yes, then this new column will say Yes or No, depending on whether the A column says Yes or No.

Think of probabilities in this "notebook" context:

- P(A) means the long-run proportion of lines in the notebook in which the A column says Yes.

- P(A or B) means the long-run proportion of lines in the notebook in which the A-or-B column says Yes.

- P(A and B) means the long-run proportion of lines in the notebook in which the A-and-B column says Yes.

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

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

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

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

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

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

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

Here are some more very important definitions and properties:

- 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. Then

  $$P(A \text{ or } B) = P(A) + P(B) \tag{1.2}$$

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

  If A and B are not disjoint, then

  $$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{1.3}$$

  In the disjoint case, that subtracted term is 0, so (1.3) reduces to (1.2).

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

  $$P(A \text{ and } B) = P(A) \cdot P(B) \tag{1.4}$$

  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,

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

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

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 (1.4) generalizes to

$$P(A \text{ and } B) = P(A)P(B|A) \tag{1.5}$$

Note that if A and B actually are independent, then $P(B|A) = P(B)$, and (1.5) reduces to (1.4).

1.2.3 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. In Equation (1.1) we found that

$$P(X_1 = 2) = p^2 + (1-p)^2 = 0.52 \tag{1.6}$$

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

$$\begin{aligned}
P(X_1 = 2) &= P(C_1 \text{ and } C_2 \text{ or } \text{ not } C_1 \text{ and } \text{ not } C_2) && (1.7)\\
&= P(C_1 \text{ and } C_2) + P(\text{ not } C_1 \text{ and } \text{ not } C_2) \text{ (from (1.2)} && (1.8)\\
&= P(C_1)P(C_2) + P(\text{ not } C_1)P(\text{ not } C_2) \text{ (from (1.4)} && (1.9)\\
&= p^2 + (1-p)^2 && (1.10)
\end{aligned}$$

Here are the reasons for these steps:

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

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

(1.9): The two nodes act physically independently of each other. Thus the events C_1 and C_2 are stochastically independent, so we applied (1.4). Then we did the same for D_1 and D_2.

Note carefully that in Equation (1.7), our first step was to **"break big events down into small events,"** in this case breaking the event $\{X_1 = 2\}$ down into the events C_1 and C_2 and D_1 and D_2. This is a central part of most probability computations. In calculating a probability, ask yourself, **"How can it happen?"**

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

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

$$\begin{aligned} P(X_2 = 2) &= P(X_1 = 0 \text{ and } X_2 = 2 \text{ or } X_1 = 1 \text{ and } X_2 = 2 \text{ or } X_1 = 2 \text{ and } X_2 = 2) \\ &= P(X_1 = 0 \text{ and } X_2 = 2) \qquad (1.11) \\ &+ P(X_1 = 1 \text{ and } X_2 = 2) \\ &+ P(X_1 = 2 \text{ and } X_2 = 2) \end{aligned}$$

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 (1.5). 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

$$P(X_1 = 1 \text{ and } X_2 = 2) = P(X_1 = 1)P(X_2 = 2|X_1 = 1) \qquad (1.12)$$

To calculate $P(X_1 = 1)$, we use the same kind of reasoning as in Equation (1.1). 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

$$P(X_1 = 1) = 2p(1-p) = 0.48 \qquad (1.13)$$

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

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

Recalling the definitions of p and q in Section 1.1, we have that

$$P(X_2 = 2|X_1 = 1) = q[p^2 + (1-p)^2] = 0.41 \tag{1.14}$$

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

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

Let's do one more; let's find $P(X_1 = 1|X_2 = 2)$. [Pause a minute here to make sure you understand that this is quite different from $P(X_2 = 2|X_1 = 1)$.] From (1.5), we know that

$$P(X_1 = 1|X_2 = 2) = \frac{P(X_1 = 1 \; and \; X_2 = 2)}{P(X_2 = 2)} \tag{1.15}$$

We computed both numerator and denominator here before, in Equations (1.12) and (1.11), so we see that $P(X_1 = 1|X_2 = 2) = 0.20/0.47 = 0.43$.

1.2.4 Bayes' Rule

Following (1.15) above, we noted that the ingredients had already been computed, in (1.12) and (1.11). If we go back to the derivations in those two equations and substitute in (1.15), we have

$$\begin{aligned} P(X_1 = 1|X_2 = 2) &= \frac{P(X_1 = 1 \text{ and } X_2 = 2)}{P(X_2 = 2} && (1.16)\\ &= \frac{P(X_1 = 1 \text{ and } X_2 = 2)}{P(X_1 = 1 \text{ and } X_2 = 2) + P(X_1 = 2 \text{ and } X_2 = 2)} && (1.17)\\ &= \frac{P(X_1 = 1)P(X_2 = 2|X_1 = 1)}{P(X_1 = 1)P(X_2 = 2|X_1 = 1) + P(X_1 = 2)P(X_2 = 2|X_1 = 2)} && (1.18) \end{aligned}$$

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

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)} \tag{1.19}$$

This is known as **Bayes' Theorem** or **Bayes' Rule**.

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

Table 1.3: Top of Notebook for Two-Epoch ALOHA Experiment

1.2.5 ALOHA in the Notebook Context

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

- Run the network for two epochs, starting with both nodes active, the first time, and write the outcome on the first line of the notebook.
- Run the network for two epochs, starting with both nodes active, the second time, and write the outcome on the second line of the notebook.
- Run the network for two epochs, starting with both nodes active, the third time, and write the outcome on the third line of the notebook.
- Run the network for two epochs, starting with both nodes active, the fourth time, and write the outcome on the fourth line of the notebook.
- Imagine you keep doing this, thousands of times, filling thousands of lines in the notebook.

The first seven lines of the notebook might look like Table 1.3. We see that:

- Among those first seven lines in the notebook, 4/7 of them have $X_1 = 2$. After many, many lines, this proportion 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 proportion will be approximately 0.47.[5]
- Among those first seven lines in the notebook, 3/7 of them have $X_1 = 2 \text{ and } X_2 = 2$. After many, many lines, this proportion will be approximately 0.27.

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

- Among the first seven lines in the notebook, four of them do not say NA in the $X_2 = 2|X_1 = 2$ column. **Among these four lines**, two say Yes, a proportion of 2/4. After many, many lines, this proportion will be approximately 0.52.

1.2.6 Simulation

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

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

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

1.2.6.1 Simulation of the ALOHA Example

Following is a computation via simulation of the *approximate* value of $P(X_1 = 2)$, $P(X_2 = 2)$ and $P(X_2 = 2|X_1 = 1)$, using the R statistical language, the language of choice of professional statisticans. It is open source, it’s statistically correct (not all statistical packages are so), has dazzling graphics capabilities, etc. To learn about the syntax (e.g. $<-$ as the assignment operator), see my introduction to R for programmers at `http://heather.cs.ucdavis.edu/~matloff/R/RProg.pdf`.

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

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

1.2.6.2 Rolling Dice

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

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

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

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

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.

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

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

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

Here the code

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

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

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

1.2.7 Combinatorics-Based Probability Computation

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.

1.2.7.1 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. 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

$$P(1 \text{ king}) = \frac{4 \cdot \binom{48}{4}}{\binom{52}{5}} = 0.299 \tag{1.20}$$

The same reasoning gives us

$$P(2 \text{ hearts}) = \frac{\binom{13}{2} \cdot \binom{39}{3}}{\binom{52}{5}} = 0.274 \tag{1.21}$$

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

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)
}
```

1.2.7.2 “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.”[6] 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.[7]. Suppose the business has a total of 20 products available for sale. What percentage of potential rules have three or fewer antecedents?[8]

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

$$\frac{\sum_{k=1}^{3} \binom{20}{k} \cdot \binom{20-k}{1}}{\sum_{k=1}^{19} \binom{20}{k} \cdot \binom{20-k}{1}} = \frac{23180}{10485740} = 0.0022 \tag{1.22}$$

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.

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

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

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

1.3 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.[9]

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.

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.

1.4 Independence, Expected Value and Variance

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 variablescalculations, these apply to both discrete and continuous random variables

1.4.1 Independent Random Variables

We already have a definition for the independence of events; what about independence of random variables? Random variables U and V are said to be **independent** if for any sets I and J, the events $\{$X is in I$\}$ and $\{$Y is in J$\}$ are independent, i.e. P(X is in I and Y is in J) = P(X is in I) P(Y is in J).

1.4.2 Expected Value

1.4.2.1 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.

[9] 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.

Then the long-run average of X is

$$\lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} \tag{1.23}$$

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.

1.4.2.2 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

$$\begin{aligned}
E(X) &= \lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} && (1.24)\\
&= \lim_{n\to\infty} \frac{0 \cdot K_{0n} + 1 \cdot K_{1n} + 2 \cdot K_{2n} ... + 10 \cdot K_{10,n}}{n} && (1.25)\\
&= \sum_{i=0}^{10} i \cdot \lim_{n\to\infty} \frac{K_{in}}{n} && (1.26)
\end{aligned}$$

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

$$E(X) = \sum_{i=0}^{10} i \cdot P(X = i) \tag{1.27}$$

So in general, the expected value of a discrete random variable X which takes value in the set A is

$$E(X) = \sum_{c \in A} cP(X = c) \tag{1.28}$$

Note that (1.28) is the formula we'll use. The preceding equations were derivation, to motivate the formula. Note too that 1.28 is not the *definition* of expected value; that was in 1.23. It is quite important to distinguish between all of these, in terms of goals.

It will be shown in Section 1.5.2.2 that in our example above in which X is the number of heads we get in 10 tosses of a coin,

$$P(X = i) = \binom{10}{i} 0.5^i (1 - 0.5)^{10-i} \tag{1.29}$$

So

$$E(X) = \sum_{i=0}^{10} i \binom{10}{i} 0.5^i (1 - 0.5)^{10-i} \tag{1.30}$$

It turns out that E(X) = 5.

For X in our dice example,

$$E(X) = \sum_{c=1}^{6} c \cdot \frac{1}{6} = 3.5 \tag{1.31}$$

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

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

For S = X+Y in the dice example,

$$E(S) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + ...12 \cdot \frac{1}{36} = 7 \tag{1.32}$$

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

$$E(N) = \sum_{c=1}^{\infty} c \cdot \frac{1}{2^c} = 2 \tag{1.33}$$

(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

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 1.4: Expanded Notebook for the Dice Problem

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 1.4 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 (1.32); we should have realized beforehand that E(S) is $2 \times 3.5 = 7$.

In other words, 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:

$$E(U+V) = E(U) + E(V) \tag{1.34}$$

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, convince yourself that

$$E(aU+b) = aE(U) + b \tag{1.35}$$

for any constants *a* and *b*. 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 is expected from that of U by using (1.35) with $a = \frac{9}{5}$ and b = 32.

But if U and V *are* independent, then

$$E(UV) = EU \cdot EV \tag{1.36}$$

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

$$E(D) = 3.5^2 = 12.25 \tag{1.37}$$

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 (1.28). Then W takes on values in $B = \{g(c) : c \epsilon A\}$. Define

$$A_d = \{c : c \in A, g(c) = d\} \tag{1.38}$$

Then

$$P(W = d) = P(X \in A_d) \tag{1.39}$$

so

$$\begin{aligned} E(W) &= \sum_{d \in B} dP(W = d) && (1.40) \\ &= \sum_{d \in B} d \sum_{c \in A_d} P(X = c) && (1.41) \\ &= \sum_{c \in A} g(c)P(X = c) && (1.42) \end{aligned}$$

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 (1.34 right away.

1.4.2.3 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. It plays an absolutely central role in probability and statistics. Yet one should understand its limitations.

First, note that 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. Even without Gates, there is a question as to whether the mean has that much meaning.

More subtly than that, there is the basic question of what the mean means. What, for example, does Equation (1.23) mean in the context of people's incomes in Davis? We would sample a person at random and record his/her income as X_1. Then we'd sample another person, to get X_2, and so on. Fine, but in that context, what would (1.23) mean? The answer is, not much.

For a casino, though, (1.23) means plenty. Say X is the amount a gambler wins on a play of a roulette wheel, and suppose (1.23) 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.

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

By contrast, in describing how wealthy people of a town are, the total income of all the residents is not relevant. Similarly, in describing how well students did on an exam, the sum of the scores of all the students doesn't tell us much. A better description might involve percentiles, including the 50th percentile, the median.

Nevertheless, the mean has certain mathematical properties, such as (1.34), 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. So, the mean has become entrenched as a descriptive measure, and we will use it often.

1.4.3 Variance

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:

$$Var(U) = E[(U - EU)^2] \tag{1.43}$$

For X in the die example, this would be

$$Var(X) = E[(X - 3.5)^2] \tag{1.44}$$

To evaluate this, apply (1.42) with $g(c) = (c - 3.5)^2$:

$$Var(X) = \sum_{c=1}^{6} (c - 3.5)^2 \cdot \frac{1}{6} = 2.92 \tag{1.45}$$

You can see that variance does indeed give us a measure of dispersion. If the values of U are mostly clustered near its mean, the variance will be small; if there is wide variation in U, the variance will be large.

The properties of E() in (1.34) and (1.35) can be used to show that

$$Var(U) = E(U^2) - (EU)^2 \tag{1.46}$$

The term $E(U^2)$ is again evaluated using (1.42).

Thus for example, if X is the number of dots which come up when we roll a die, and $W = X^2$, then

$$E(W) = \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} = \frac{91}{6} \tag{1.47}$$

An important property of variance is that

$$Var(cU) = c^2 Var(U) \tag{1.48}$$

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. And shifting data over by a constant does not change the amount of variation in them, so

$$Var(cU + d) = c^2 Var(U) \tag{1.49}$$

for any constant d.

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

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 3.9.2 for details.)

As with expected values, the properties of variance discussed above, and also in Section 3.2.1 below, are key to the entire remainder of this book. You should notice immediately when you are in a setting in which they are applicable. For instance, if you see the variance of the sum of two random variables, you should instinctively think of (1.61 right away.

1.4.3.1 Is Var(X) Large or Small?

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.

1.4.3.2 Chebychev's Inequality

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

$$P(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2} \tag{1.50}$$

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.

To prove (1.50), let's first state and prove Markov's Inequality: For any nonnegative random variable Y,

$$P(Y \geq d) \leq \frac{EY}{d} \tag{1.51}$$

To prove (1.51), let Z equal 1 if $Y \geq d$, 0 otherwise. Then

$$Y \geq dZ \tag{1.52}$$

(think of the two cases), so

$$EY \geq dEZ \tag{1.53}$$

The right-hand side of (1.53) is $dP(Y \geq d)$, so (1.51) follows.

Now to prove (1.50), define

$$Y = (X - \mu)^2 \tag{1.54}$$

and set $d = c^2\sigma^2$. Then (1.51) says

$$P[(X - \mu)^2 \geq c^2\sigma^2] \leq \frac{E[(X - \mu)^2]}{c^2\sigma^2} \tag{1.55}$$

Since

$$(X - \mu)^2 \geq c^2\sigma^2 \text{ if and only if } |X - \mu| \geq c\sigma \tag{1.56}$$

the left-hand side of (1.55) is the same as the left-hand side of (1.50). The numerator of the right-hand size of (1.55) is simply Var(X), i.e. σ^2, so we are done.

1.4.3.3 The Coefficient of Variation

Continuing our discussion of the magnitude of a variance, look at our remark following (1.50):

> 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:

$$\text{coef. of var.} = \frac{\sqrt{Var(X)}}{EX} \tag{1.57}$$

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.

1.4.4 Covariance

This is a topic we'll cover fully in Chapter 3, but at least introduce here.

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

$$Cov(U,V) = E[(U-EU)(V-EV)] \tag{1.58}$$

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

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

$$Cov(U,V) = E(UV) - EU \cdot EV \tag{1.59}$$

Also

$$Var(U+V) = Var(U) + Var(V) + 2Cov(U,V) \tag{1.60}$$

If U and V are independent, then Cov(U,V) = 0 and

$$Var(U+V) = Var(U) + Var(V) \tag{1.61}$$

1.4.5 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).

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

$$ED = -2 \cdot P(D=-2) + 0 \cdot P(D=0) + 2 \cdot P(D=2) + 4 \cdot P(D=4) \tag{1.62}$$

Now, using reasoning along the lines in Section 1.2.7, we have

$$P(D = -2) = P(M = 1 \text{ and } W = 3) = \frac{\binom{6}{1}\binom{3}{3}}{\binom{9}{4}} \tag{1.63}$$

After similar calculations for the other probabilities in (1.62), we find the ED = 1.33. If we were to perform this experiment many times, i.e. choose committees again and again, on average we would have a bit more than one more man than women on the committee.

1.4.6 Expected Value, Etc. in the ALOHA Example

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

$$EX_1 = 0 \cdot P(X_1 = 0) + 1 \cdot P(X_1 = 1) + 2 \cdot P(X_1 = 2) = 1 \cdot 0.48 + 2 \cdot 0.52 = 1.52 \tag{1.64}$$

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

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

1.4.7 Reconciliation of Math and Intuition (optional section)

Here I have been promoting the notebook idea over the sterile, confusing mathematical definitions in the theory of probability. It is worth noting, though, that the theory actually does imply the notebook notion, through a theorem known as the Strong Law of Large Numbers:

Consider a random variable U, and a sequence of independent random variables $U_1, U_2, ...$ which all have the same distribution as U, i.e. they are "repetitions" of the experiment which generates U. Then

$$\lim_{n \rightarrow \infty} \frac{U_1 + ... + U_n}{n} = E(U) \text{ with probability 1} \tag{1.65}$$

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

1.5 Distributions

1.5.1 Basic Notions

For the type of random variables we've discussed so far, the **distribution** of a random variable U is simply a list of all the values it takes on, and their associated probabilities:

Example: For X in the dice example, the distribution of X is

$$\{(1,\frac{1}{6}),(2,\frac{1}{6}),(3,\frac{1}{6}),(4,\frac{1}{6}),(5,\frac{1}{6}),(6,\frac{1}{6})\} \tag{1.66}$$

Example: In the ALOHA example, distribution of X_1 is

$$\{(0,0.00),(1,0.48),(2,0.52)\} \tag{1.67}$$

Example: In our example in which N is the number of tosses of a coin needed to get the first head, the distribution is

$$\{(1, \frac{1}{2}), (2, \frac{1}{4}), (3, \frac{1}{8}), ...\} \tag{1.68}$$

It is common to express this in functional notation. We define the **probability mass function** (pmf) of a discrete random variable V, denoted p_V, as

$$p_V(k) = P(V = k) \tag{1.69}$$

for any value k which V can take on.

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

Example: In (1.68),

$$p_N(k) = \frac{1}{2^k}, k = 1, 2, ... \tag{1.70}$$

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

$$p_S(k) = \begin{cases} \frac{1}{36}, & k = 2 \\ \frac{3}{36}, & k = 3 \\ \frac{3}{36}, & k = 4 \\ ... \\ \frac{1}{36}, & k = 12 \end{cases} \tag{1.71}$$

It is important to note that there may not be some nice closed-form expression for p_V like that of (1.70). There was no such form in (1.71), nor is there in our ALOHA example for p_{X_1} and p_{X_2}.

1.5.2 Parameteric Families of pmfs

1.5.2.1 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

$$p_N(k) = (\frac{1}{2})^{k-1} \cdot \frac{1}{2}, k = 1, 2, ... \tag{1.72}$$

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

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

$$p_N(k) = \left(\frac{5}{6}\right)^{k-1} \frac{1}{6}, k = 1, 2, ... \tag{1.73}$$

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

The tosses of the coin and the rolls of the die are known as **Bernoulli trials**, which is a sequence of independent 1-0-valued random variables B_i, i = 1,2,3,... B_i is 1 for success, 0 for failure, with success probability p. For instance, p is 1/2 in the coin case, and 1/6 in the die example.

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

$$p_U(k) = (1-p)^{k-1} p, k = 1, 2, ... \tag{1.74}$$

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

It can be shown that

$$E(U) = \frac{1}{p} \tag{1.75}$$

(which should make good intuitive sense to you) and

$$Var(U) = \frac{1-p}{p^2} \tag{1.76}$$

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).

1.5.2.2 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).[10]

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

$$P(X=2) = \binom{5}{2} 0.5^2(1-0.5)^3 = \binom{5}{2}/32 = 5/16 \tag{1.77}$$

For general n and p,

$$P(X=k) = \binom{n}{k} p^k(1-p)^{n-k} \tag{1.78}$$

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:

$$X = \sum_{i=1}^{n} B_i \tag{1.79}$$

where B_i is 1 or 0, depending on whether there is success on the i^{th} trial or not. Then the reader should use our earlier properties of E() and Var() in Section 1.4 to fill in the details in the following derivations of the expected value and variance of a binomial random variable:

$$EX = E(B_1+...,+B_n) = EB_1+...+EB_n = np \tag{1.80}$$

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

and

$$Var(X) = Var(B_1 + ..., +B_n) = Var(B_1) + ... + Var(B_n) = np(1-p) \tag{1.81}$$

Again, (1.80) should make good intuitive sense to you.

1.5.2.3 The Poisson Family of Distributions

Another famous parametric family of distributions is the set of **Poisson Distributions**, which is used to model unbounded counts. The pmf is

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, ... \tag{1.82}$$

The parameter for the family, λ, turns out to be the value of E(X) and also Var(X).

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.

1.5.2.4 The Negative Binomial Family of Distributions

Recall that a typical example of the geometric distribution family (Section 1.5.2.1) 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:

$$\{N = 5\} = \{\text{2 heads in the first 4 tosses and head on the } 5^{th} \text{ toss}\} \tag{1.83}$$

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

$$P(\text{2 heads in the first 4 tosses}) = \binom{4}{2}\left(\frac{1}{2}\right)^4 \tag{1.84}$$

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

$$P(N=5) = \binom{4}{2} \left(\frac{1}{2}\right)^5 = \frac{3}{16} \tag{1.85}$$

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

$$P(N=k) = \binom{k-1}{r-1} (1-p)^{k-r} p^r, k = r, r+1, ... \tag{1.86}$$

We can write

$$N = G_1 + ... + G_r \tag{1.87}$$

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

$$E(N) = r \cdot \frac{1}{p} \tag{1.88}$$

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

$$Var(N) = r \cdot \frac{1-p}{p^2} \tag{1.89}$$

1.5.2.5 The Power Law Family of Distributions

Here

$$p_X(k) = ck^{-\gamma},\ k = 1, 2, 3, ... \tag{1.90}$$

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:

$$1.0 = \sum_{k=1}^{\infty} ck^{-\gamma} \approx c \int_1^{\infty} k^{-\gamma}\, dk = c \tag{1.91}$$

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. It turns out that many types of networks in the real world exhibit approximately power law behavior.

For instance, in a famous study of the Web (A. Barabasi and R. Albert, Emergence of Scaling in Random Networks, *Science*, 1999, 509-512), 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 $\gamma = 2.7$.

1.6 Recognizing Distributions When You See Them

Many random variables one encounters do not have a distribution in some famous parametric family. But many do, and it's important to be alert to this point, and recognize one when you see one.

1.6.1 A Coin Game

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 15th time, Jill has her 13th toss; etc.

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

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

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

$$P(Y_6 > X_6) = \sum_{i=1}^{4} \sum_{j=0}^{i-1} P(Y_6 = i \text{ and } X_6 = j) \tag{1.92}$$

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,

$$P(Y_6 = i \text{ and } X_6 = j) = P(Y_6 = i) \cdot P(X_6 = j) \tag{1.93}$$

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

$$P(Y_6 = i) = \binom{4}{i} 0.5^i (1-0.5)^{4-i} \tag{1.94}$$

and

$$P(X_6 = j) = \binom{6}{j} 0.5^j (1-0.5)^{6-j} \tag{1.95}$$

We would then substitute (1.94) and (1.95) in (1.92). We could then evaluate it by hand, but it would be more convenient to use R's **dbinom()** function:

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

1.6.2 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

$$p = \binom{4}{2} 0.5^4 = \frac{3}{8} \tag{1.96}$$

Thus using the fact that N has a geometric distribution,

$$P(N = 5) = (1-p)^4 p = 0.057 \tag{1.97}$$

1.6.3 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.[11] 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,

$$\binom{m}{1} p(1-p)^{m-1} = mp(1-p)^{m-1} \tag{1.98}$$

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 (1.98).

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.

1.7 A Cautionary Tale

1.7.1 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

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

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

$$P(A_1 \text{ and } A_2) = P(A_1)P(A_2) \tag{1.99}$$

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

$$P(A_1) = P(C_1)P(A_1|C_1) + P(C_2)P(A_1|C_2) = (0.5)(0.9) + (0.5)(0.1) = 0.5 \tag{1.100}$$

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

$$\begin{aligned}
P(A_1 \text{ and } A_2) &= P(A_1 \text{ and } A_2 \text{ and } C_1) + P(A_1 \text{ and } A_2 \text{ and } C_2) && (1.101)\\
&= P(A_1 \text{ and } A_2|C_1)P(C_1) + P(A_1 \text{ and} A_2|C_2)P(C_2) && (1.102)\\
&= (0.9)^2(0.5) + (0.1)^2(0.5) && (1.103)\\
&= 0.41 && (1.104)
\end{aligned}$$

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.

1.7.2 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.**

1.7.3 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.

1.8 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.

1.9 Tips on Finding Probabilities, Expected Values and So On

First, do not write/think nonsense. For example, the expression "P(A) or P(B)" is nonsense—do you see why?

Similarly, don't use "formulas" that you didn't learn and are in fact false. For example, in an expression involving a random variable X, one can NOT replace X by EX! (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"?)

As noted before, in calculating a probability, ask yourself, **"How can it happen?"** Then you will typically have a set of and/or terms, which you compute individually and add together. And until you get used to it, **write down every step, including reasons**, as you see in (1.7)-(1.9).

Another point is that you should define variables, e.g. "Let X denote the number of heads." *Write it down!* This makes it much easier to translate from words to math expressions and equations.

Exercises

1. This problem concerns the ALOHA network model of Section 1.1. 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. 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 $E(X)$.

3. Recall the committee example in Section 1.4.5. 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 $Var(D)$.

4. 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?

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

5. 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).

6. 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.

7. 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.

8. Consider the example of association rules in Section 1.2.7.2. How many two-antecedent, two-consequent rules are possible from 20 items? Express your answer in terms of combinatorial (“n choose k”) symbols.

9. 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?

10. 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) P(X = 2).

(b) P(X+Y = 3).

(c) Var(X+Y).

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

11. Assume the ALOHA network model as in Section 1.1, 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.

12. Consider the game in Section 1.6.1. Find $E(Z)$ and $Var(Z)$, where $Z = Y_6 - X_6$.

13. 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).

14. 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 $Var(X)$.

15. 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.

16. 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.

17. 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.

18. 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.

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

$$\binom{n}{k} = \binom{n-1}{k} + \binom{c}{d} \tag{1.105}$$

20. Prove Equation (1.46), and also show that $b = EU$ minimizes the quantity $E](U-b)^2]$.

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

$$EX = \sum_{i=1}^{\infty} P(X \geq i) \tag{1.106}$$

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

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

$$\lim_{n \to \infty} P(X = n/2) = 0 \tag{1.107}$$

Use Stirling's approximation,

$$k! \approx \sqrt{2\pi k} \left(\frac{k}{e}\right)^k \tag{1.108}$$

Chapter 2

Continuous Probability Models

2.1 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.

The first crucial point to note is that

$$P(D = c) = 0 \tag{2.1}$$

for any individual point c. That can be seen by the fact that c is in as tiny a subinterval as you wish, or by the fact that the interval (c,c), or even [c,c], has length 0. Or, 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.

That may 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.

But Equation (2.1) presents a problem for us in defining the term **distribution** for variables like this. We defined it 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, for any random variable W (including discrete ones), its

cumulative distribution function (cdf):

$$F_W(t) = P(W \leq t), -\infty < t < \infty \tag{2.2}$$

(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.

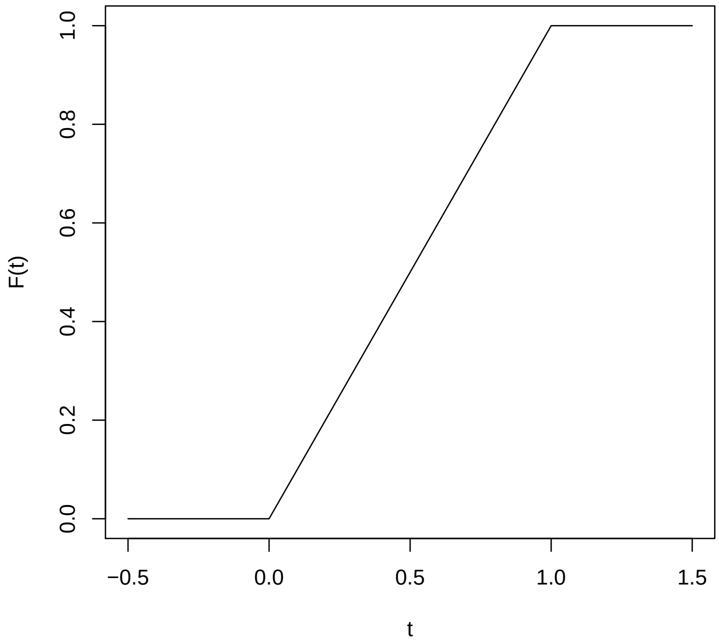
For instance, consider our "random dart" example above. We know that, for example

$$F_D(0.23) = P(D \leq 0.23) = 0.23 \tag{2.3}$$

In general for our dart,

$$F_D(t) = \begin{cases} 0, & \text{if } t \leq 0 \\ t, & \text{if } 0 < t < 1 \\ 1, & \text{if } t \geq 1 \end{cases} \tag{2.4}$$

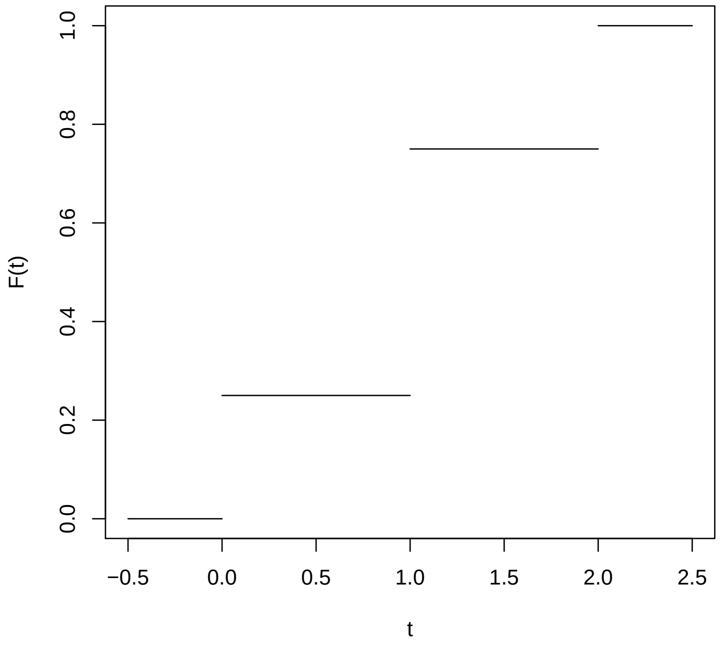
Here is the graph of F_D:



The cdf of a discrete random variable is defined as in Equation (2.2) too. For example, say Z is the number of heads we get from two tosses of a coin. Then

$$F_Z(t) = \begin{cases} 0, & \text{if } t < 0 \\ 0.25, & \text{if } 0 \leq t < 1 \\ 0.75, & \text{if } 1 \leq t < 2 \\ 1, & \text{if } t \geq 2 \end{cases} \tag{2.5}$$

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 (2.4) is continuous while F_Z in (2.5) has jumps. For this reason, we call random variables like D—ones which have 0 probability for individual points—**continuous random variables**.

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

2.2 Density Functions

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

2.2.1 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.

Think as follows. From (2.2) 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**. It is defined as

$$f_W(t) = \frac{d}{dt} F_W(t), -\infty < t < \infty \tag{2.6}$$

(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 2.1, 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:

$$2(0.1) f_X(1.3) \approx \int_{1.2}^{1.4} f_X(t) \, dt = P(1.2 < X < 1.4) \tag{2.7}$$

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

$$2c f_X(t) \approx P(t - c < X < t + c) \tag{2.8}$$

Thus we have:

Intrepetation of Density Functions

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

$$\frac{P(r - c < X < r + c)}{P(s - c < X < s + c)} \approx \frac{f_X(r)}{f_X(s)} \tag{2.9}$$

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.

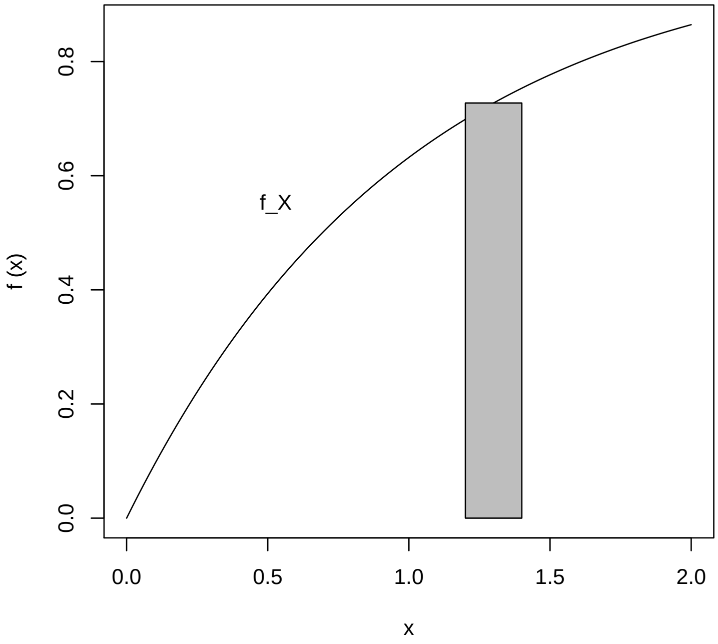


Figure 2.1: Approximation of Probability by a Rectangle

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.)

2.2.2 Use of Densities to Find Probabilities and Expected Values

Equation (2.6) implies that

$$P(a < W < b) = \int_a^b f_W(t) \, dt \tag{2.10}$$

This in turn implies that

$$\int_{-\infty}^{\infty} f_W(t) \, dt = 1 \tag{2.11}$$

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

$$E(W) = \sum_c c p_W(c) \tag{2.12}$$

So, the analog for continuous W is

$$E(W) = \int_t t f_W(t) \, dt \tag{2.13}$$

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

And of course,

$$E(W^2) = \int_t t^2 f_W(t) \, dt \tag{2.14}$$

and in general,

$$E[g(W)] = \int_t g(t) f_W(t) \, dt \tag{2.15}$$

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

$$EX = \int_1^4 t \cdot 2t/15 \, dt = 2.8 \tag{2.16}$$

$$P(X > 2.5) = \int_{2.5}^4 2t/15 \, dt = 0.65 \tag{2.17}$$

$$F_X(s) = \int_1^s 2t/15 \, dt = \frac{s^2-1}{15} \text{ for s in (1,4) (cdf is 0 for t < 1, and 1 for t > 4)} \tag{2.18}$$

2.3 Famous Parametric Families of Continuous Distributions

2.3.1 The Uniform Distributions

2.3.1.1 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 (2.11). So, that constant must be 1 divided by the length of the interval:

$$f_D(t) = \frac{1}{r-q} \tag{2.19}$$

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).

2.3.1.2 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**.

2.3.1.3 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.

2.3.2 The Normal (Gaussian) Family of Continuous Distributions

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

2.3.2.1 Density and Properties

Density and Parameters:

[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.

The density for a normal distribution is

$$f_W(t) = \frac{1}{\sqrt{2\pi}\sigma} \, e^{-0.5\left(\frac{t-\mu}{\sigma}\right)^2}, -\infty < t < \infty \tag{2.20}$$

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 $x\mu + d$ and variance $c^2\sigma^2$, which would follow from (1.49) *even if X did not have a normal distribution.* The key point is that this new variable Y is also a member of the normal family, i.e. its density is still given by (2.20), now with the new mean and variance.

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

$$\begin{aligned}
F_Y(t) &= P(Y \leq t) \;\; \text{(definition of } F_Y\text{)} && (2.21)\\
&= P(cX + d \leq t) \;\; \text{(definition of Y)} && (2.22)\\
&= P\left(X \leq \frac{t-d}{c}\right) \;\; \text{(algebra)} && (2.23)\\
&= F_X\left(\frac{t-d}{c}\right) \;\; \text{(definition of } F_X\text{)} && (2.24)
\end{aligned}$$

Therefore

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

$$
\begin{aligned}
f_Y(t) &= \frac{d}{dt} F_Y(t) \ \text{(definition of } f_Y\text{)} && (2.25)\\
&= \frac{d}{dt} F_X \left(\frac{t-d}{c} \right) \ \text{(from (2.24))} && (2.26)\\
&= f_X \left(\frac{t-d}{c} \right) \cdot \frac{d}{dt} \frac{t-d}{c} \ \text{(definition of } f_X \text{ and the Chain Rule)} && (2.27)\\
&= \frac{1}{c} \cdot \frac{1}{\sqrt{2\pi}\sigma} \ e^{-0.5\left(\frac{\frac{t-d}{c}-\mu}{\sigma}\right)^2} \ \text{(from (2.20)} && (2.28)\\
&= \frac{1}{\sqrt{2\pi}(c\sigma)} \ e^{-0.5\left(\frac{t-(c\mu+d)}{c\sigma}\right)^2} \ \text{(algebra)} && (2.29)
\end{aligned}
$$

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

Evaluating the Normal cdf

The function in (2.20) 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

$$\frac{X-\mu}{\sigma} \tag{2.30}$$

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()**.

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

2.3.2.2 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

$$P(X \geq 535) = P\left(Z \geq \frac{535 - 500}{15}\right) \approx 1 - \Phi(35/15) = 0.01 \tag{2.31}$$

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.

2.3.2.3 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:

Theorem 1 *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.

2.3.2.4 Example: Bug Counts

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

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

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
```

2.3.2.5 Example: Coin Tosses

Binomially distributed random variables, though discrete, also are approximately normally distributed. This comes from the fact that if say T has a binomial distribution with n trials, then we can write $T = T_1 + ... + T_n$, where T_i is 1 for a success and 0 for a failure. Since we have a sum, 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

$$P(X > 12) = P(Z > \frac{12 - 10}{\sqrt{5}}) \approx 1 - \Phi(0.894) = 0.186 \tag{2.32}$$

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 accounting for the fact that X is discrete, replacing 12 by 12.5 above. (Think of the number 13 “owning” the region between 12.5 and 13.5.) This is customary, and in this case gives us 0.1317762, while the exact answer to seven decimal places is 0.131588. This is called the **correction of continuity**. Of course, for larger n this adjustment is not necessary.

2.3.2.6 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.rand.org/statistics/applets/clt.html` and `http://www.jcu.edu/math/isep/Quincunx/Quincunx.html`. By collecting the balls in bins, the apparatus basically simulates a histogram for X, which will then be approximately bell-shaped.

2.3.2.7 Optional topic: Formal Statement of the CLT

Definition 2 *A sequence of random variables $L_1, L_2, L_3, ...$* **converges in distribution** *to a random variable M if*

$$\lim_{n \to \infty} P(L_n \leq t) = P(M \leq t), \textit{ for all t} \tag{2.33}$$

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

$$Z = \frac{X_1 + ...X_n - nm}{v\sqrt{n}} \tag{2.34}$$

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

2.3.2.8 Importance in Modeling

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. The latter implies that many things in nature do have approximate normal distributions.

2.3.3 The Chi-Square Family of Distributions

2.3.3.1 Density and Properties

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

$$Y = Z_1^2 + ... + Z_k^2 \tag{2.35}$$

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.

It turns out that chi-square is a special case of the gamma family in Section 2.3.5 below, with r = k/2 and $\lambda = 0.5$.

2.3.3.2 Importance in Modeling

This distribution is used widely in statistical applications. As will be seen in our chapters on statistics, many statistical methods involve a sum of squared normal random variables.[6]

2.3.4 The Exponential Family of Distributions

2.3.4.1 Density and Properties

The densities in this family have the form

$$f_W(t) = \lambda e^{-\lambda t}, 0 < t < \infty \tag{2.36}$$

This is a one-parameter family of distributions.[7]

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

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

[7] In the mathematical theory of statistics, the term *exponential family* has a broader meaning than this.

2.3.4.2 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

$$T_r = X_1 + ... + X_r \tag{2.37}$$

denote the time of the i^{th} replacement. Also, let N(t) denote the number of replacements up to and including time t.[8] 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 other words, N(t) will have a Poisson distribution if and only if the lifetimes are exponentially distributed. You can see the “only if” part quickly, by the following argument. First, note that

$$P(X_1 > t) = P[N(t) = 0] = e^{-\lambda t} \tag{2.38}$$

Then

$$f_{X_1}(t) = \frac{d}{dt}(1 - e^{-\lambda t}) = \lambda e^{-\lambda t} \tag{2.39}$$

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

The relation $E[N(t)] = \lambda t$ says is 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 (2.36).

2.3.4.3 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.

[8] Again, since N(t) is a continuous random variable, the phrase “and including” is unnecssary here.

2.3.5 The Gamma Family of Distributions

2.3.5.1 Density and Properties

Recall Equation (2.37), 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

$$f_{T_r}(t) = \frac{1}{(r-1)!} \lambda^r t^{r-1} e^{-\lambda t}, \ t > 0 \tag{2.40}$$

This is a two-parameter family.

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

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} \, dx \tag{2.41}$$

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

$$f_W(t) = \frac{1}{\Gamma(r)} \lambda^r t^{r-1} e^{-\lambda t}, \ t > 0 \tag{2.42}$$

(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 (2.37) 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 2.2 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 $\lambda = 0.5$.

2.3.5.2 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

$$P(Y > 552) = \int_{552}^{\infty} \frac{1}{4!} 0.01^5 t^4 e^{-0.01t} \, dt \tag{2.43}$$

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
```

2.3.5.3 Importance in Modeling

As seen in (2.37), 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 2.2.

2.4 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.

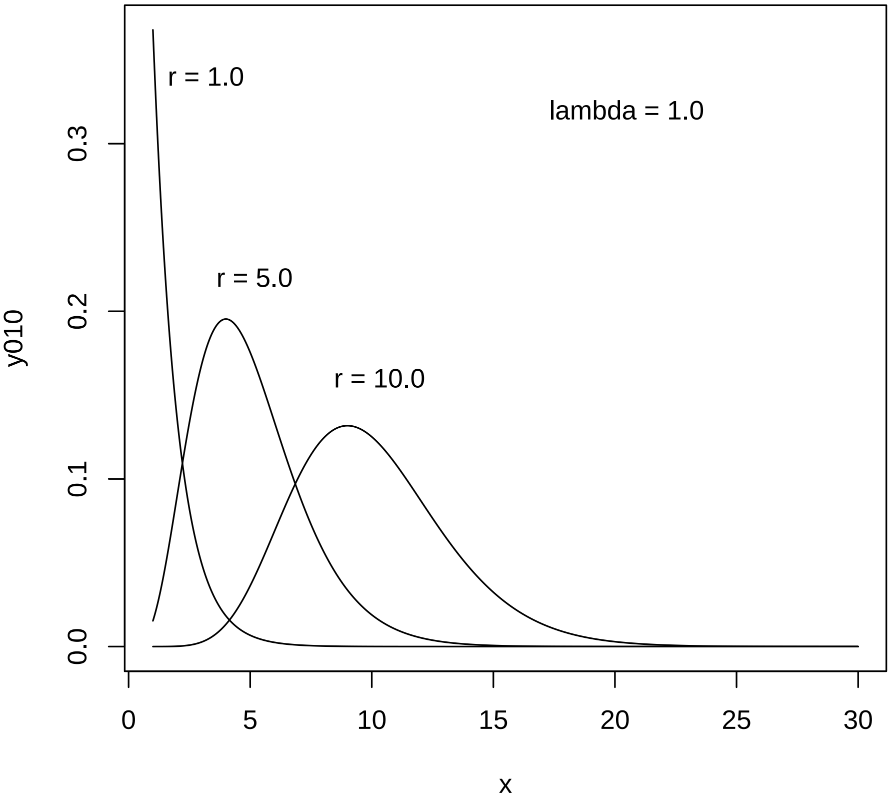


Figure 2.2: Various Gamma Densities

2.4.1 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**. What this means is that for positive t and u

$$P(W > t + u | W > t) = P(W > u) \tag{2.44}$$

Let's derive this:

$$\begin{aligned} P(W > t + u | W > t) &= \frac{P(W > t + u \text{ and } W > t)}{P(W > t)} && (2.45) \\ &= \frac{P(W > t + u)}{P(W > t)} && (2.46) \\ &= \frac{\int_{t+u}^{\infty} \lambda e^{-\lambda s} \, ds}{\int_{t}^{\infty} \lambda e^{-\lambda s} \, ds} && (2.47) \\ &= e^{-\lambda u} && (2.48) \\ &= P(W > u) && (2.49) \end{aligned}$$

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 continous even though it is discrete), Equation (2.44) 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.

The memorylessness of exponential distributions implies that a Poisson process N(t) also has a "time starts over" property (called the **Markov property**). Recall our example in Section 2.3.4.2 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 $t < z$.

By the way, the exponential distributions are the only continuous distributions which are memoryless. This too has implications for the theory.

2.4.2 Hazard Functions

2.4.2.1 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

$$P(L = 81|L > 80) = \frac{P(L = 81)}{P(L > 80)} = \frac{p_L(81)}{1 - F_L(80)} \tag{2.50}$$

By analogy, for continuous L we define

$$h_L(t) = \frac{f_L(t)}{1 - F_L(t)} \tag{2.51}$$

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 (2.51) for an exponential density with parameter λ; sure enough, the hazard function is constant, with value λ.

The reader should verify that in contrast to an exponential distribution's constant failure rate, a uniform distribution has an increasing failure rate (IFR). Some distributions have decreasing failure rates, while most have non-monotone rates.

Hazard function models have been used extensively in software testing. Here "failure" is the discovery of a bug, and with quantities of interest include the mean time until the next bug is discovered, and the total number of bugs.

People have what is called a “bathtub-shaped” hazard function. It is high near 0 (reflecting infant mortality) and after, say, 70, but is low and rather flat in between.

You may have noticed that the right-hand side of (2.51) is the derivative of $-ln[1 - F_L(t)]$. Therefore

$$\int_0^t h_L(s)\, ds = -\ln[1 - F_L(t)] \tag{2.52}$$

so that

$$1 - F_L(t) = e^{-\int_0^t h_L(s)\, ds} \tag{2.53}$$

and thus[9]

$$f_L(t) = h_L(t) e^{-\int_0^t h_L(s)\, ds} \tag{2.54}$$

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.

2.4.3 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.

See for example *Accurate Software Reliability Estimation*, by Jason Allen Denton, Dept. of Computer Science, Colorado State University, 1999, and the many references therein.

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

2.5 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.*

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.[10] (The formal name for this is **length-biased sampling**.)

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.

Once one recognizes this and carefully finds the density of that interval, we discover that that interval does indeed tend to be longer—so much so that the expected value of this interval is 20 minutes! 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.

Thus the halfway point comes at 10 minutes, consistent with the analysis which appealed to the memoryless property.

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. Suppose that,

[10] 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 the square of the diameter.

say, stick lengths 2 and 6 each comprise 10% of the sticks in the bag, i.e.

$$p_X(2) = p_X(6) = 0.1 \tag{2.55}$$

Intuitively, one would then reason that

$$p_Y(6) = 3p_Y(2) \tag{2.56}$$

In other words, the sticks of length 2 are just as numerous as those of length 6, but since the latter are three times as long, they should have triple the chance of being chosen.

Note that this is not some absolute physical law. Different people might draw sticks from the bag in different ways. But it is a reasonable model.

Now let X denote interarrival times between buses, and Y denote the interarrival time that we hit. The analog of (2.56) would be that $f_Y(t)$ is proportional to $tf_X(t)$, i.e.

$$f_Y(t) = ctf_X(t) \tag{2.57}$$

for some constant c. Recalling that f_Y must integrate to 1, we see that

$$c = \left(\int_0^\infty tf_X(t) \, dt \right)^{-1} \tag{2.58}$$

But that integral is just E(X)! The latter quantity is 10, and

$$f_X(t) = 0.1e^{-0.1t} \tag{2.59}$$

So,

$$f_Y(t) = 0.01te^{-0.1t} \tag{2.60}$$

You may recognize this as an Erlang density.

2.6 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 2.2, 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.

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**.

2.7 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

$$\begin{aligned} P[G(U) \leq t] & \\ &= P[U \leq G^{-1}(t)] \\ &= P[U \leq H(t)] \\ &= H(t) \end{aligned} \tag{2.61}$$

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.

Exercises

1. Fill in the blanks, in the following statements about continous 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 2.3.2.2, 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 continous 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 2.7, write an R function, named **oneminust()**, that generates a random variate sampled from this distribution. Then use this function to verify your answers in (b) above.

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

(a) The mean and variance of the error.

(b) The mean of the absolute error.

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

9. Suppose that computer roundoff error in computing the square roots of numbers in a certain range is distributed uniformly on (-0.5,0.5), and that we will be computing the sum of n such square roots.

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

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

(c) Find a number c such that the probability is approximately 95% that the sum is in error by no more than c.

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

$$f_X(t) = \frac{1}{\pi c} \frac{1}{1 + (\frac{t-b}{c})^2}, \ \infty < t < \infty \tag{2.62}$$

is bell-shaped but definitely not normal.

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

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

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

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

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

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

11. Use R to plot the hazard functions for the gamma distributions plotted in Figure 2.2, plus the case r = 0.5. Comment on the implications for trains at 8th and J Streets in Davis.

12. 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.

13. Consider a machine that places a pin in the middle of a flat, disk-shaped object. The placement is subject to error. Let X and Y be the placement errors in the horizontal and vertical directions, respectively, and let W denote the distance from the true center to the pin placement. Suppose X and Y are independent and have normal distributions with mean 0 and variance 0.04. Find $P(W > 0.7)$.

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

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

15. Consider the "random bucket" example in Footnote 10. 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.

16.

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

17. In Section 2.4.1, 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 (2.44).

Chapter 3

Multivariate Probability Models

3.1 Multivariate Distributions

3.1.1 Why Are They Needed?

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.

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

3.1.2 Discrete Case

Say we roll a blue die and a yellow one. Let X and Y denote the number of dots which appear on the blue and yellow dice, respectively, and let S denote the total number of dots appearing on the two dice. We will not discuss Y much here, focusing on X and S.

Recall that the *distribution* of X is defined to be a list of all the values X takes on, and their associated probabilities:

$$\{(1,\frac{1}{6}),(2,\frac{1}{6}),(3,\frac{1}{6}),(4,\frac{1}{6}),(5,\frac{1}{6}),(6,\frac{1}{6})\} \tag{3.1}$$

We can write this more compactly (but equivalently) by defining X's *probability mass function* (pmf):

$$p_X(i) = P(X = i) = \frac{1}{6}, i = 1, 2, ..., 6 \tag{3.2}$$

The distribution of S is defined similarly, either as a list,

$$\{(2, \frac{1}{36}), (3, \frac{2}{36}), (4, \frac{3}{36}), (5, \frac{4}{36}), (6, \frac{5}{36})(7, \frac{6}{36})(8, \frac{5}{36})(9, \frac{4}{36})(10, \frac{3}{36})(11, \frac{2}{36})(12, \frac{1}{36})\} \tag{3.3}$$

or via its pmf p_S.[1]

But it may also be important to describe how X and S vary jointly. For example, intuitively we would feel that X and S are positively correlated. How do we describe their joint variation?

To do this, we define the *bivariate probability mass function* (often called the *joint probability mass function*) of X and S. Just as the univariate pmf of X is defined to be $p_X(i) = P(X = i)$, we define the bivariate pmf as

$$p_{X,S}(i, j) = P(X = i, S = j) = \frac{1}{36}, i = 1, 2, ..., 6; j = i + 1, ..., i + 6 \tag{3.4}$$

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

$$E[g(X] = \sum_i g(i)p_X(i) \tag{3.5}$$

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

$$E[g(U, V)] = \sum_{i,j} g(i, j)p_{U,V}(i, j) \tag{3.6}$$

For instance:

$$E(XS) = \sum_{i=1}^{6}\sum_{j=2}^{12} ij\ p_{X,S}(i, j) = \sum_{i=1}^{6}\sum_{j=i+1}^{i+6} ij\frac{1}{36} \tag{3.7}$$

The univariate pmfs, called *marginal pmfs*, can of course be recovered from the bivariate pmf. To get $p_X()$ from $p_{X,S}()$, we sum over the values of S. For example, let's find $p_X(3)$, which is the probability that X =

[1] Recall that the convention for denoting pmfs is to use the letter 'p' with a subscript indicating the random variable.

3. How could the event X = 3 happen? Well, S could be anywhere from 4 to 9, each with probability 1/6. So,

$$p_X(3) = \sum_{j=4}^{9} p_{X,S}(3,j) = 6 \cdot \frac{1}{36} = \frac{1}{6} \tag{3.8}$$

That is consistent with our univariate calculation of $p_X(3)$, as of course it should be.

We get consistent results for expected values too. Treating X as a function of X and S, we have

$$E(X) = \sum_{i=1}^{6} \sum_{j=i+1}^{i+6} i p_{X,S}(i,j) \tag{3.9}$$

but the right-hand side (RHS) of (3.9) reduces to

$$E(X) = \sum_{i=1}^{6} i \sum_{j=i+1}^{i+6} p_{X,S}(i,j) = \sum_{i=1}^{6} i p_X(i) \tag{3.10}$$

from (3.8). The last expression in (3.10) is E(X) as defined in the univariate setting, so everything is indeed consistent.

3.1.3 Multivariate Densities

3.1.3.1 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

$$F_{X,Y}(u,v) = P(X \leq u \text{ and } Y \leq v) \tag{3.11}$$

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:

$$f_{X,Y}(u,v) = \frac{\partial}{\partial u} \frac{\partial}{\partial v} F_{X,Y}(u,v) \tag{3.12}$$

Densities in higher dimensions are defined similarly.

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

3.1.3.2 Use of Multivariate Densities in Finding Probabilities and Expected Values

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

$$P[(X,Y)\epsilon A] = \int \int_A f_{X,Y}(u,v) \; du \; dv \tag{3.13}$$

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),

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u,v) f_{X,Y}(u,v) \; du \; dv \tag{3.14}$$

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 (3.13). See (3.18) below for an example.

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

$$f_X(s) = \int_t f_{X,Y}(s,t) \; dt \tag{3.15}$$

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

3.1.3.3 Example: a Triangular Distribution

Suppose (X,Y) has the density

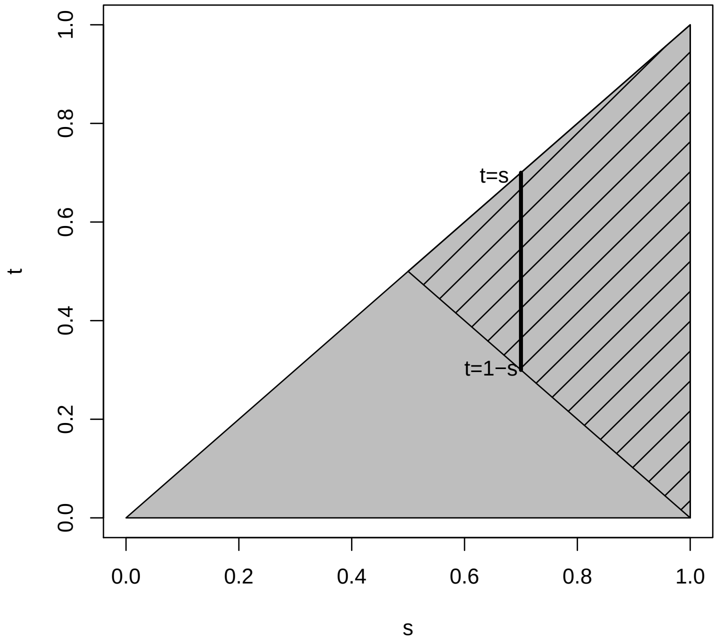
$$f_{X,Y}(s,t) = 8st, 0 < t < s < 1 \tag{3.16}$$

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 (3.13) 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.

The limits in the double integral are obtained through the following reasoning, 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 (3.13), 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.

$$P(X+Y>1) = \int_{0.5}^{1} \int_{1-s}^{s} 8st \; dt \; ds = \int_{0.5}^{1} 8s \cdot (s-0.5) \; ds = \frac{5}{6} \tag{3.17}$$

Following (3.14),

$$E[\sqrt{X+Y}] = \int_{0}^{1} \int_{0}^{s} \sqrt{s+t} \; 8st \; dt \; ds \tag{3.18}$$

Let's find the marginal density $f_Y(t)$. So we must "integrate out" the s in (3.16):

$$f_Y(t) = \int_{t}^{1} 8st \; ds = 4t - 4t^3 \tag{3.19}$$

3.2 More on Co-variation of Random Variables

3.2.1 Covariance

The *covariance* between random variables X and Y is defined a

$$Cov(X,Y) = E[(X-EX)(Y-EY)] \tag{3.20}$$

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 (3.20) 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* X is larger or smaller than its mean, etc.

Covariance is linear in both arguments:

$$Cov(aX+bY, cU+dV) = acCov(X,U) + adCov(X,V) + bcCov(Y,U) + bdCov(Y,V) \tag{3.21}$$

for any constants a, b, c and d. Also

$$Cov(X, Y+q) = Cov(X,Y) \tag{3.22}$$

for any constant q and so on.

Note that

$$Cov(X,X) = Var(X) \tag{3.23}$$

for any X with finite variance.

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

$$Cov(X,Y) = E(XY) - EX \cdot EY \tag{3.24}$$

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 (3.24).

$$\begin{aligned} Cov(X,Y) &= E[(X-EX)(Y-EY)] \text{ (definition)} && (3.25)\\ &= E\left[XY - EX\cdot Y - EY\cdot X + EX\cdot EY\right] \text{ (algebra)} && (3.26)\\ &= E(XY) + E[-EX\cdot Y] + E[-EY\cdot X] + E[EX\cdot EY] \text{ (E[U+V]=EU+EV)} && (3.27)\\ &= E(XY) - EX\cdot EY \text{ (E[cU] = cEU, Ec = c)} && (3.28) \end{aligned}$$

Another important property:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \tag{3.29}$$

This comes from (3.24), the relation $Var(X) = E(X^2) - EX^2$ and the corresponding one for Y. Just substitute and do the algebra.

3.2.2 Correlation

Covariance does measure how much or little X and Y vary together, but it is hard to decide whether a given value of covariance is “large” or not. For instance, if we are measuring lengths in feet and change to inches, then (3.21) 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

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \tag{3.30}$$

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

3.2.3 Example: Continuation of Section 3.1.3.3

Let's find the correlation between X and Y in the example in Section 3.1.3.3.

$$\begin{aligned} E(XY) &= \int_0^1 \int_0^s st \cdot 8st \; dt \; ds && (3.31) \\ &= \int_0^1 8s^2 \cdot s^3/3 \; ds && (3.32) \\ &= \frac{4}{9} && (3.33) \end{aligned}$$

$$\begin{aligned} f_X(s) &= \int_0^s 8st \; dt && (3.34) \\ &= 4st^2 \Big|_0^s && (3.35) \\ &= 4s^3 && (3.36) \end{aligned}$$

$$\begin{aligned} f_Y(t) &= \int_t^1 8st \; ds && (3.37) \\ &= 4t \cdot s^2 \Big|_t^1 && (3.38) \\ &= 4t(1-t^2) && (3.39) \end{aligned}$$

$$EX = \int_0^1 s \cdot 4s^3 \; ds = \frac{4}{5} \qquad (3.40)$$

$$E(X^2) = \int_0^1 s^2 \cdot 4s^3 \; ds = \frac{2}{3} \tag{3.41}$$

$$Var(X) = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = 0.027 \tag{3.42}$$

$$EY = \int_0^1 t \cdot (4t - 4t^3) \; ds = \frac{4}{3} - \frac{4}{5} = \frac{8}{15} \tag{3.43}$$

$$E(Y^2) = \int_0^1 t^2 \cdot (4t - 4t^3) \; dt = 1 - \frac{4}{6} = \frac{1}{3} \tag{3.44}$$

$$Var(Y) = \frac{1}{3} - \left(\frac{8}{15}\right)^2 = 0.049 \tag{3.45}$$

$$Cov(X,Y) = \frac{4}{9} - \frac{4}{5} \cdot \frac{8}{15} = 0.018 \tag{3.46}$$

$$\rho(X,Y) = \frac{0.018}{\sqrt{0.027 \cdot 0.049}} = 0.49 \tag{3.47}$$

3.2.4 Example: a Catchup Game

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

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

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

```
taketurn <- function(a,b) {
   win <- runif(1)
   if (a >= b) return(win)
   else return(win+0.2)
}

```

```
cdf2 <- function(xy,t1,t2) {  # 2-dim. cdf
   tmp <- xy[xy[,1] <= t1 & xy[,2] <= t2,]
   return(nrow(tmp)/nrow(xy))
}

nturns <- 10
xyvals <- matrix(nrow=nreps,ncol=2)
for (rep in 1:nreps) {
   x <- 0
   y <- 0
   for (turn in 1:nturns) {
      # x's turn
      x <- x + taketurn(x,y)
      # y's turn
      y <- y + taketurn(y,x)
   }
   xyvals[rep,] <- c(x,y)
}
print(cor(xyvals[,1],xyvals[,2]))
print(cdf2(xyvals,5.8,5.2))
```

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

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

3.3 Sets of Independent Random Variables

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.

3.3.1 Properties

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

3.3.1.1 Probability Mass Functions and Densities Factor

If X and Y are independent, then

$$p_{X,Y} = p_X p_Y \tag{3.48}$$

in the discrete case, and

$$f_{X,Y} = f_X f_Y \tag{3.49}$$

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:

$$\begin{aligned}
p_{X,Y}(i,j) &= P(X = i \text{ and } Y = j) \ \ \text{(definition)} &(3.50)\\
&= P(X = i)P(Y = j) \ \ \text{(independence)} &(3.51)\\
&= p_X(i)p_Y(j) \ \ \text{(definition))} &(3.52)
\end{aligned}$$

Here is the proof for the continuous case;

$$\begin{aligned}
f_{X,Y}(u,v) &= \frac{\partial}{\partial u}\frac{\partial}{\partial v}F_{X,Y}(u,v) &(3.53)\\
&= \frac{\partial}{\partial u}\frac{\partial}{\partial v}P(X \le u \text{ and } Y \le v) &(3.54)\\
&= \frac{\partial}{\partial u}\frac{\partial}{\partial v}P(X \le u) \cdot PY \le v) &(3.55)\\
&= \frac{\partial}{\partial u}\frac{\partial}{\partial v}F_X(u) \cdot F_Y(v) &(3.56)\\
&= f_X(u)f_Y(v) &(3.57)
\end{aligned}$$

3.3.1.2 Expected Values Factor

If X and Y are independent, then

$$E(XY) = E(X)E(Y) \tag{3.58}$$

To prove this, use (3.48) and (3.49) for the discrete and continuous cases.

3.3.1.3 Covariance Is 0

If X and Y are independent, then from (3.58) and (3.24), we have

$$Cov(X,Y) = 0 \tag{3.59}$$

and thus

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

However, the converse is false. A counterexample is the random pair (V,W) that is uniformly distributed on the unit disk, $\{(s,t) : s^2 + t^2 \leq 1\}$.

3.3.1.4 Variances Add

If X and Y are independent, then from (3.29) and (3.58), we have

$$Var(X+Y) = Var(X) + Var(Y). \tag{3.60}$$

3.3.1.5 Convolution

If X and Y are nonnegative, continuous random variables, and we set Z = X+Y, then the density of Z is the *convolution* of the densities of X and Y:

$$f_Z(t) = \int_0^t f_X(s) f_Y(t-s)\, ds \tag{3.61}$$

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;

$$\begin{aligned}
p_W(k) &= P(W=k) \text{ (by definition)} && (3.62)\\
&= P(U+V=k) \text{ (substitution)} && (3.63)\\
&= \sum_{i=0}^{k} P(U=i \text{ and } V=k-i) \text{ ("In what ways can it happen?")} && (3.64)\\
&= \sum_{i=0}^{k} p_{U,V}(i,k-i) \text{ (by definition)} && (3.65)\\
&= \sum_{i=0}^{k} p_U(i) p_V(k-i) \text{ (from Section 3.3.1.1)} && (3.66)
\end{aligned}$$

Review the analogy between densities and pmfs in our unit on continuous random variables, Section 2.2.1, and then see how (3.61) is analogous to (3.62) through (3.66):

- k in (3.62) is analogous to t in (3.61)
- the limits 0 to k in (3.66) are analogous to the limits 0 to t in (3.61)
- the expression k-i in (3.66) is analogous to t-s in (3.61)
- and so on

3.3.2 Examples

3.3.2.1 Example: Dice

In Section 3.2.1, we speculated that the correlation between X, the number on the blue die, and S, the total of the two dice, was positive. Let's compute it.

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

$$\begin{aligned} Cov(X,S) &= Cov(X,X+Y) \ \text{(by definition)} && (3.67)\\ &= Cov(X,X)+Cov(X,Y) \ \text{(from (3.21))} && (3.68)\\ &= Var(X)+0 \ \text{(from (3.23), (3.59)} && (3.69) \end{aligned}$$

Also, from (3.60),

$$Var(S) = Var(X+Y) = Var(X) + Var(Y) \tag{3.70}$$

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

$$\rho(X,S) = \frac{Var(X)}{\sqrt{Var(X)}\sqrt{2Var(X)}} = 0.707 \tag{3.71}$$

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.

3.3.2.2 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. 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 (3.49), we know that their joint density is the product of their marginal densities,

$$f_{X,Y}(s,t) = 25e^{-5(s+t)}, s,t > 0 \tag{3.72}$$

Now

$$P(|X-Y| \leq 0.1) = 1 - P(|X-Y| > 0.1) = 1 - P(X > Y + 0.1) - P(Y > X + 0.1) \tag{3.73}$$

Look at that first probability. Applying (3.13) with $A = \{(s,t) : s > t + 0.1, 0 < s, t\}$, we have

$$P(X > Y + 0.1) = \int_0^\infty \int_{t+0.1}^\infty 25e^{-5(s+t)} \, ds \, dt = 0.303 \tag{3.74}$$

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.

3.3.2.3 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 (3.14) to be $|s-t|$.

$$\int_0^1 \int_0^1 |s-t| \cdot 1 \, ds \, dt = \frac{1}{3} \tag{3.75}$$

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 51, 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.

3.3.2.4 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 (3.3.1.5), we have

$$f_W(t) = \int_0^t f_X(s) f_Y(t-s)\, ds = \int_0^t 0.5e^{-0.5s} e^{-(t-s)}\, ds = e^{-0.5t} - e^{-t},\ 0 < t < \infty \tag{3.76}$$

3.4 Matrix Formulations

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.

3.4.1 Properties of Mean Vectors

The expected value of W is defined to be the vector

$$EW = (EW_1, ..., EW_k)' \tag{3.77}$$

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

$$E(cV + dW) = cEV + dEW \tag{3.78}$$

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

3.4.2 Properties of Covariance Matrices

The covariance matrix Σ of W is the k x k matrix whose $(i,j)^{th}$ element is $Cov(W_i, W_j)$. Note that that means that the diagonal elements of the matrix are the variances of the W_i, and that the matrix is symmetric.

We write the covariance matrix of W as Cov(W).

As you can see, in the statistics world, the Cov() function is "overloaded." If it has two argument, 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, and define Q = c W. Then Q is a k-component random vector like W, and

$$Cov(Q) = c^2 Cov(W) \tag{3.79}$$

- If A is an r x k but nonrandom matrix, define Q = A W. Then Q is an r-component random vector, and

$$Cov(Q) = A\ Cov(W)\ A' \tag{3.80}$$

- 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

$$Cov(V + W) = Cov(V) + Cov(W) \tag{3.81}$$

- In analogy with (1.46), for any random vector Q,

$$Cov(Q) = E(QQ') - EQ\ (EQ)' \tag{3.82}$$

3.5 Conditional Distributions

The key to good probability modeling and statistical analysis is to understand conditional probability. The issue arises constantly.

3.5.1 Conditional Pmfs and Densities

First, let's review: In many repetitions of our "experiment," P(A) is the long-run proportion of the time that A occurs. By contrast, P(A|B) is the long-run proportion of the time that A occurs, *among those repetitions in which B occurs.* Keep this in your mind at all times.

Now we apply this to pmfs, densities, etc. We define the conditional pmf as follows for discrete random variables X and Y:

$$p_{Y|X}(j|i) = P(Y = j|X = i) = \frac{p_{X,Y}(i,j)}{p_X(i)} \tag{3.83}$$

By analogy, we define the conditional density for continuous X and Y:

$$f_{Y|X}(t|s) = \frac{f_{X,Y}(s,t)}{f_X(s)} \tag{3.84}$$

3.5.2 Conditional Expectation

Conditional expectations are defined as straightforward extensions of (3.83) and (3.84):

$$E(Y|X = i) = \sum_j j p_{Y|X}(j|i) \tag{3.85}$$

$$E(Y|X = s) = \int_t t f_{Y|X}(t|s)\ dt \tag{3.86}$$

3.5.3 The Law of Total Expectation (advanced topic)

3.5.3.1 Expected Value As a Random Variable

For a random variable Y and an event A, the quantity E(Y|A) is the long-run average of Y, among the times when A occurs. Note several things about the expression E(Y|A):

- The expression evaluates to a constant.
- The item to the left of the | symbol is a *random variable* (Y).
- The item on the right of the | symbol is an *event* (A).

By contrast, for the quantity E(Y|W) defined below, for a random variable W, it is the case that:

- The expression itself is a random variable, not a constant.
- The item to the left of the | symbol is again a random variable (Y).
- But the item to the right of the | symbol is also a random variable (W).

It will be very important to keep these differences in mind.

Consider the function g(t) defined as[2]

$$g(t) = E(Y|W = t) \tag{3.87}$$

In this case, the item to the right of the | is an event, and thus g(t) is a constant (for each value of t), not a random variable.

Now, define the random variable Q to be g(W). Since W is a random variable, then Q is too. The quantity E(Y|W) is then defined to be Q. (Before reading any further, re-read the two sets of bulleted items above, and make sure you understand the difference between E(Y|W=t) and E(Y|W).)

One can view E(Y|W) as a projection in an abstract vector space. This is very elegant, and actually aids the intuition. If (and only if) you are mathematically adventurous, read the details in Section 3.9.2.

[2] Of course, the t is just a placeholder, and any other letter could be used.

3.5.3.2 The Famous Formula (Theorem of Total Expectation)

An extremely useful formula, given only scant or no mention in most undergraduate probability courses, is

$$E(Y) = E[E(Y|W)] \tag{3.88}$$

for any random variables Y and W.

The RHS of (3.88) looks odd at first, but it's merely E[g(W)]; since Q = E(Y|W) is a random variable, we can certainly ask what its expected value is.

Equation (3.88) is a bit abstract. It's a very useful abstraction, enabling streamlined writing and thinking about the process. Still, you may find it helpful to consider the case of discrete W, in which (3.88) has the more concrete form

$$EY = \sum_i P(W = i) \cdot E(Y|W = i) \tag{3.89}$$

To see this intuitively, think of measuring the heights and weights of all the adults in Davis. Say we measure height to the nearest inch, so that height is discrete. We look at all the adults in Davis who are 72 inches tall, and write down their mean weight. Then we write down the mean weight of all adults of height 68. Then we write down the mean weight of all adults of height 75, and so on. Then (3.88) says that if we take the average of all the numbers we write down—the average of the averages—then we get the mean weight among *all* adults in Davis.

Note carefully, though, that this is a *weighted* average. If for instance people of height 69 inches are more numerous in the population, then their mean weight will receive greater emphasis in over average of all the means we've written down. This is seen in (3.89), with the weights being the quantities P(W=i).

The relation (3.88) is proved in the discrete case in Section 3.10.

3.5.4 What About the Variance?

By the way, one might guess that the analog of the Theorem of Total Expectation for variance is

$$Var(Y) = E[Var(Y|W)] \tag{3.90}$$

But this is false. Think for example of the extreme case in which Y = W. Then Var(Y|W) would be 0, but Var(Y) would be nonzero.

The correct formula, called the Law of Total Variance, is

$$Var(Y) = E[Var(Y|W)] + Var[E(Y|W)] \tag{3.91}$$

Deriving this formula is easy, by simply evaluating both sides, and using the relation $Var(X) = E(X^2) - (EX)^2$. This exercise is left to the reader.

3.5.5 Example: Trapped Miner

(Adapted from *Stochastic Processes,* by Sheldon Ross, Wiley, 1996.)

A miner is trapped in a mine, and has a choice of three doors. Though he doesn't realize it, if he chooses to exit the first door, it will take him to safety after 2 hours of travel. If he chooses the second one, it will lead back to the mine after 3 hours of travel. The third one leads back to the mine after 5 hours of travel. Suppose the doors look identical, and if he returns to the mine he does not remember which door(s) he tried earlier. What is the expected time until he reaches safety?

Let Y be the time it takes to reach safety, and let W denote the number of the door chosen (1, 2 or 3) on the first try. Then let us consider what values E(Y|W) can have. If W = 1, then Y = 2, so

$$E(Y|W = 1) = 2 \tag{3.92}$$

If W = 2, things are a bit more complicated. The miner will go on a 3-hour excursion, and then be back in its original situation, and thus have a further expected wait of EY, since "time starts over." In other words,

$$E(Y|W = 2) = 3 + EY \tag{3.93}$$

Similarly,

$$E(Y|W = 3) = 5 + EY \tag{3.94}$$

In summary, now considering the *random variable* E(Y|W), we have

$$Q = E(Y|W) = \begin{cases} 2, & w.p.\frac{1}{3} \\ 3 + EY, & w.p.\frac{1}{3} \\ 5 + EY, & w.p.\frac{1}{3} \end{cases} \tag{3.95}$$

where “w.p.” means “with probability.” So, using (3.88) or (3.89), we have

$$EY = EQ = 2 \times \frac{1}{3} + (3 + EY) \times \frac{1}{3} + (5 + EY) \times \frac{1}{3} = \frac{10}{3} + \frac{2}{3}EY \tag{3.96}$$

Equating the extreme left and extreme right ends of this series of equations, we can solve for EY, which we find to be 10.

It is left to the reader to see how this would change if we assume that the miner remembers which doors he has already hit.

3.5.6 Example: Analysis of Hash Tables

(Famous example, adapted from various sources.)

Consider a database table consisting of m cells, only some of which are currently occupied. Each time a new key must be inserted, it is used in a hash function to find an unoccupied cell. Since multiple keys map to the same table cell, we may have to probe multiple times before finding an unoccupied cell.

We wish to find E(Y), where Y is the number of probes needed to insert a new key. One approach to doing so would be to condition on W, the number of currently occupied cells at the time we do a search. After finding E(Y|W), we can use the Theorem of Total Expectation to find EY. We will make two assumptions (to be discussed later):

(a) Given that W = k, each probe will collide with an existing cell with probability k/m, with successive probes being independent.

(b) W is uniformly distributed on the set 1,2,...,m, i.e. P(W = k) = 1/m for each k.

To calculate E(Y|W=k), we note that given W = k, then Y is the number of independent trials until a “success” is reached, where “success” means that our probe turns out to be to an unoccupied cell. This is a **geometric** distribution, i.e.

$$P(Y = r|W = k) = (\frac{k}{m})^{r-1}(1 - \frac{k}{m}) \tag{3.97}$$

The mean of this geometric distribution is, from (1.75),

$$\frac{1}{1 - \frac{k}{m}} \tag{3.98}$$

Then

$$
\begin{aligned}
EY &= E[E(Y|W)] && (3.99)\\
&= \sum_{k=1}^{m-1} \frac{1}{m} E(Y|W=k) && (3.100)\\
&= \sum_{k=1}^{m-1} \frac{1}{m-k} && (3.101)\\
&= 1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{m-1} && (3.102)\\
&\approx \int_1^m \frac{1}{u} du && (3.103)\\
&= ln(m) && (3.104)
\end{aligned}
$$

where the approximation is something you might remember from calculus (you can picture it by drawing rectangles to approximate the area under the curve.).

Now, what about our assumptions, (a) and (b)? The assumption in (a) of each cell having probability k/m should be reasonably accurate if k is much smaller than m, because hash functions tend to distribute probes uniformly, and the assumption of independence of successive probes is all right too, since it is very unlikely that we would hit the same cell twice. However, if k is not much smaller than m, the accuracy will suffer.

Assumption (b) is more subtle, with differing interpretations. For example, the model may concern one specific database, in which case the assumption may be questionable. Presumably W grows over time, in which case the assumption would make no sense—it doesn't even *have* a distribution. We could instead think of a database which grows and shrinks as time progresses. However, even here, it would seem that W would probably oscillate around some value like m/2, rather than being uniformly distributed as assumed here. Thus, this model is probably not very realistic. However, even idealized models can sometimes provide important insights.

3.6 Parametric Families of Distributions

Since there are so many ways in which random variables can correlate with each other, there are rather few parametric families commonly used to model multivariate distributions (other than those arising from sets of independent random variables have a distribution in a common parametric univariate family). We will discuss two here.

3.6.1 The Multinomial Family of Distributions

3.6.1.1 Probability Mass Function

This is a generalization of the binomial family.

Suppose one tosses a die 8 times. What is the probability that the results consist of two 1s, one 2, one 4, three 5s and one 6? Well, if the tosses occur in that order, i.e. the two 1s come first, then the 2, etc., then the probability is

$$\left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^1 \tag{3.105}$$

But there are many different orderings, in fact

$$\frac{8!}{2!1!0!1!3!1!} \tag{3.106}$$

of them.

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. Then we say that $X_1, ..., X_r$ have a **multinomial distribution**, and the joint pmf of the $X_1, ..., X_r$ is

$$p_{X_1,...,X_r}(j_1, ..., j_r) = \frac{n!}{j1!...j_r!} p_1^{j_1} ... p_r^{j_r} \tag{3.107}$$

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

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

```
# n is the number of trial, p the vector of probabilities of the r
# categories
multinom <- function(n,p) {
   r <- length(p)
   outcome <- sample(x=1:r,size=n,replace=T,prob=p)
```

```
    counts <- vector(length=r)  # counts of the various categories
    # tabulate the counts (could be done more efficiently)
    for (i in 1:n) {
        j <- outcome[i]
        counts[j] <- counts[j] + 1
    }
    return(counts)
}
```

3.6.1.2 Means and Covariances

Now look at the vector $X = (X_1, ..., X_r)'$. Let's find its mean vector and covariance matrix.

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

$$EX_i = np_i \text{ and } Var(X_i) = np_i(1 - p_i) \tag{3.108}$$

So we know EX now:

$$EX = \begin{pmatrix} np_1 \\ ... \\ np_r \end{pmatrix} \tag{3.109}$$

We also know the diagonal elements of Cov(X); what about the rest? To this end, let T_{ki} equal 1 or 0, depending on whether the kth trial results in outcome i, k = 1,...,n and i = 1,...,r. We say that T_{ki} is the **indicator variable** for the event that kth trial results in outcome i. This is a simple concept, but it has powerful uses, as you'll see.

Make sure you understand that

$$X_i = \sum_{k=1}^{n} T_{ki} \tag{3.110}$$

From (3.110), you can see that

$$X = U_1 + ... + U_n \tag{3.111}$$

where

$$U_k = \begin{pmatrix} T_{k1} \\ ... \\ T_{kr} \end{pmatrix} \tag{3.112}$$

Now, here's where the power of the matrix operations in Section 3.4 will be seen:

$$\begin{aligned} Cov(X) &= Cov(U_1 + ... + U_n) \text{ (from (3.111))} && (3.113) \\ &= Cov(U_1) + ... + Cov(U_n) \text{ (from (3.81))} && (3.114) \\ &= nCov(U_1) \text{ (all have the same distribution)} && (3.115) \end{aligned}$$

Now, for $i \neq j$, we have from (3.24)

$$Cov(T_{1i}, T_{1j}) = E(T_{1i}T_{1j}) - ET_{1i} \cdot ET_{1j} \tag{3.116}$$

But $T_{1i} \cdot T_{1j} = 0$! And $ET_{1i} = p_i$ and the same for the j case. So,

$$Cov(T_{1i}, T_{1j}) = -p_i p_j \tag{3.117}$$

Of course, for i = j, $Cov(T_{1i}, T_{1j}) = Var(T_{1i} = p_i(1 - p_i)$, since T_{1i} has a binomial distribution with number of trials equal to 1.

Putting all this together, and recalling (3.115), we see that

$$Cov(X) = n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & ... & -p_1p_r \\ -p_1p_2 & p_2(1-p_2) & ... & -p_2p_r \\ ... & ... & ... & ... \\ ... & ... & ... & p_r(1-p_r) \end{pmatrix} \tag{3.118}$$

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 (3.118) and (3.79), we have

$$Cov(R) = \frac{1}{n} \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & ... & -p_1p_r \\ -p_1p_2 & p_2(1-p_2) & ... & -p_2p_r \\ ... & ... & ... & ... \\ ... & ... & ... & p_r(1-p_r) \end{pmatrix} \tag{3.119}$$

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

3.6.1.3 Application: Text Mining

One of the branches of computer science in which the multinomial family plays a prominent role is in text mining. One goal is automatic document classification. We want to write software that will make reasonably accurate guesses as to whether a document is about sports, the stock market, elections etc., based on the frequencies of various key words the program finds in the document.

Many of the simpler methods for this use the **bag of words model**. We have r key words we've decided are useful for the classification process, and the model assumes that statistically the frequencies of those words in a given document category, say sports, follow a multinomial distribution. Each category has its own set of probabilities $p_1, ..., p_r$. For instance, if "Barry Bonds" is considered one word, its probability will be much higher in the sports category than in the elections category, say. So, the observed frequencies of the words in a particular document will hopefully enable our software to make a fairly good guess as to the category the document belongs to.

Once again, this is a very simple model here, designed to just introduce the topic to you. Clearly the multinomial assumption of independence between trials is grossly incorrect here, most models are much more complex than this.

3.6.2 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.

3.6.2.1 Densities and Properties

Intuitively, this family has densities which are shaped like multidimensional bells, just like the univariate normal has the famous one-dimensional bell shape.

Let's look at the bivariate case first. The joint distribution of X_1 and X_2 is said to be **bivariate normal** if their density is

$$f_{X,Y}(s,t) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(s-\mu_1)^2}{\sigma_1^2}+\frac{(t-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(s-\mu_1)(t-\mu 2)}{\sigma_1\sigma_2}\right]}, \quad -\infty < s,t < \infty \tag{3.120}$$

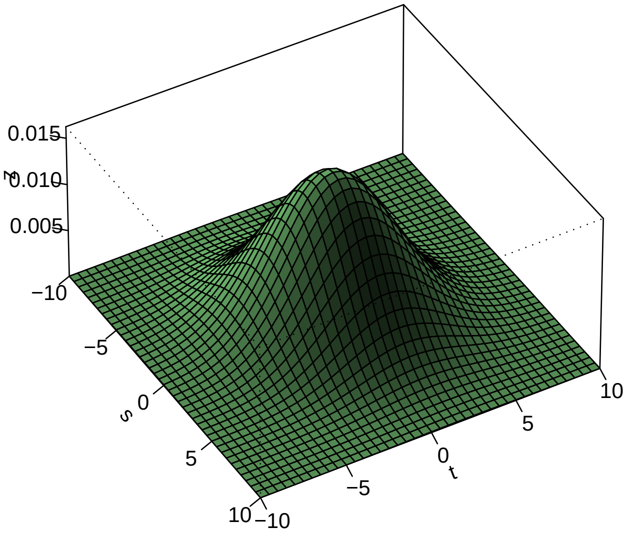


Figure 3.1: Bivariate Normal Density, $\rho = 0.2$

This looks horrible, and it is. But don't worry, as we won't work with this directly. It's important for conceptual reasons, as follows.

First, note parameters here: μ_1, μ_2, σ_1 and σ_2 are the means and standard deviations of X_1 and X_2, while ρ is the correlation between X_1 and X_2. So, we have a five-parameter family of distributions.

Now, let's look at some pictures, generated by R code which I've adapted from one of the entries in the R Graph Gallery, `http://addictedtor.free.fr/graphiques/graphcode.php?graph=42`.[3] 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 3.1 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

[3] 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.

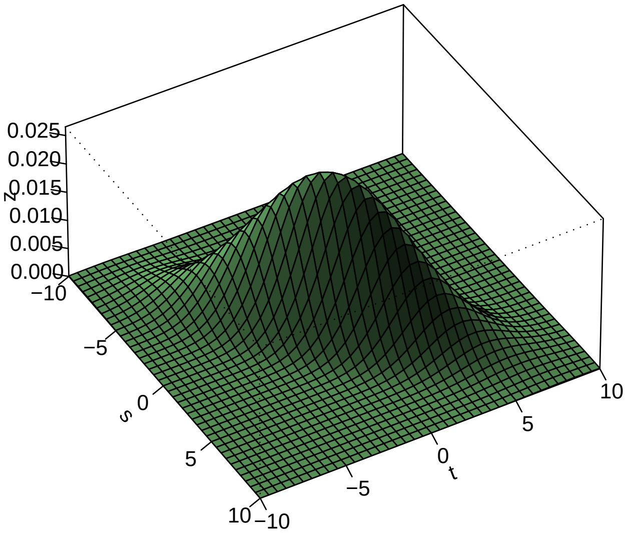


Figure 3.2: Bivariate Normal Density, $\rho = 0.8$

is rather low there, then there are rather few people whose height and weight are near those values.

Now compare that picture to Figure 3.2, 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

$$X_1 - \mu_1 = \frac{\sigma_1}{\sigma_2}(X_2 - \mu_2) \tag{3.121}$$

In this case, that would be

$$X_1 = \sqrt{\frac{10}{15}} X_2 = 0.82 X_2 \tag{3.122}$$

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:

$$f_X(t) = ce^{-0.5(t-\mu)'\Sigma^{-1}(t-\mu)} \tag{3.123}$$

where

$$c = \frac{1}{(2\pi)^{k/2}\sqrt{det(\Sigma)}} \tag{3.124}$$

Here again ' denotes matrix transpose, -1 denotes matrix inversion and det() means determinant. Again, note that t is a kx1 vector.

Since the matrix is symmetric, there are k(k+1)/2 distinct parameters there, and k parameters in the mean vector, for a total of k(k+3)/2 parameters for this family of distributions.

The family has the following important properties:

Theorem 4 (Properties of Multivariate Normal Distributions)

Suppose $X = (X_1, ..., X_k)$ has a multivariate normal distribution with mean vector μ and covariance matrix Σ. Then:

(a) The contours of f_X are k-dimensional ellipsoids. In the case k = 2 for instance, where we can visualize the density of X as a three-dimensional surface, the contours for points at which the bell has the same height (think of a topographical map) are elliptical in shape. The larger the correlation (in absolute value) between X_1 and X_2, the more elongated the ellipse. When the absolute correlation reaches 1, the ellipse degenerates into a straight line.

(b) Let A be a constant (i.e. nonrandom) matrix with k columns. Then the random vector Y = AX also has a multivariate normal distribution, with mean $A\mu$ and covariance matrix $A\Sigma A'$.

(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

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.125}$$

(ii) Scalar linear combinations of X are normal. In other words, for constant scalars $a_1, ..., a_k$, set $a = (a_1, ..., a_k)'$. Then the quantity $Y = a_1X_1 + ... + a_kX_k$ has a univariate normal distribution with mean $a'\mu$ and variance $a'\Sigma a$.

(iii) Vector linear combinations are multivariate normal. Again using the case k = 3 as our example, consider $(U, V)' = (X_1 - X_3, X_2 - X_3)$. Then set

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \tag{3.126}$$

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**.

3.6.2.2 The Multivariate Central Limit Theorem

The multidimensional version of the Central Limit Theorem holds. A sum of independent identically distributed random vectors has an approximate multivariate normal distribution.

For example, since a person's body consists of many different components, the CLT (a non-independent, non-identically version of it) explains intuitively why heights and weights are approximately bivariate normal. Histograms of heights will look approximately bell-shaped, and the same is true for weights. The multivariate CLT says that three-dimensional histograms—plotting frequency along the "Z" axis against height and weight along the "X" and "Y" axes—will be approximately three-dimensional bell-shaped.

3.6.2.3 Example: Dice Game

Suppose we roll a die 50 times. Let X denote the number of rolls in which we get one dot, and let Y be the number of times we get either two or three dots. For convenience, let's also define Z to be the number of times we get four or more dots, though our focus will be on X and Y. Suppose also that we win $5 for each roll of a one, and $2 for each roll of a two or three.

Let's find the approximate values of the following:

- $P(X \leq 12 \text{ and } Y \leq 16)$
- P(win more than $90)
- $P(X > Y > Z)$

The triple (X,Y,Z) has a multinomial distribution with n = 50 and three possible outcomes (1; 2 or 3; 4, 5 or 6), with $p_1 = 1/6$, $p_2 = 1/3$ and $p_3 = 1/2$. From (3.111), we see that (X,Y,Z) has an approximately multivariate normal distribution.

These probabilities of interest to us here would be quite difficult to find directly. For $P(X \leq 12 \text{ and } Y \leq 16)$, for instance, we would need to sum (3.107) over many, many different cases. So, the CLT will be very valuable here.

We'll of course need to know the mean vector and covariance matrix of the random vector (X,Y,Z)'. We have those from (3.108) and (3.118):

$$E[(X,Y,Z)] = (50/6, 50/3, 50/2) \tag{3.127}$$

and

$$Cov[(X,Y,Z)] = 50 \begin{pmatrix} 5/36 & -1/18 & -1/12 \\ -1/18 & 2/9 & -1/6 \\ -1/12 & -1/6 & 1/4 \end{pmatrix} \tag{3.128}$$

We use the R function **pmvnorm()** introduced in Section 3.6.2.1. To account for the integer nature of X and Y, we call the function with upper limits of 12.5 and 16.5, rather than 12 and 16, which is often used to get a better approximation. Our code is

```
p1 <- 1/6
p23 <- 1/3
meanvec <- 50*c(p1,p23)
var1 <- 50*p1*(1-p1)
var23 <- 50*p23*(1-p23)
covar123 <- -50*p1*p23
covarmat <- matrix(c(var1,covar123,covar123,var23),nrow=2)
print(pmvnorm(upper=c(12.5,16.5),mean=meanvec,sigma=covarmat))
```

We find that

$$P(X \leq 12 \text{ and } Y \leq 16) \approx 0.43 \tag{3.129}$$

Now, let's find the probability that our total winnings, W, is over \$90. We know that W = 5X + 2Y, and Theorem 4(b) tells us that linear combinations of a multivariate normal random vector are (univariate) normal. In other words, W has an approximate normal distribution!

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

$$EW = E(5X + 2Y) = 5EX + 2EY = 250/6 + 100/3 = 75 \tag{3.130}$$

For the variance, take the matrix A to be the row vector (5,2)] in the theorem, giving us Var(W) = 162.5. Then

$$P(W > 90) = 1 - \Phi\left(\frac{90-75}{162.5^{0.5}}\right) = 0.12 \tag{3.131}$$

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

$$A = \left(\begin{array}{rrr} 1 & -1 & 0 \\ 0 & 1 & -1 \end{array} \right) \tag{3.132}$$

and then proceed as before to find $P(U > 0, V > 0)$.

By the way, note that the fact that Z is an exact linear function of X and Y turns out to make the covariance matrix Σ **singular**, i.e. lacking an inverse. That would create problems in (3.123).

3.6.2.4 Application: Data Mining

The multivariate normal family plays a central role in multivariate statistical methods.

For instance, a major issue in data mining is **dimension reduction**, which means trying to reduce what may be hundreds or thousands of variables down to a manageable level. One of the tools for this, called **principle components analysis** (PCA), is based on multivariate normal distributions. Google uses this kind of thing quite heavily. We'll discuss PCA in Section 7.5.

To see a bit of how this works, note that in Figure 3.2, 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 7.5 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.[4]

3.7 Simulation of Random Vectors

Let $X = (X_1, ..., X_k)'$ be a random vector having a specified distribution. How can we write code to simulate it? It is not always easy to do this. We'll discuss a couple of easy cases here, and illustrate what one may do in other situations.

The easiest case (and a very frequenly-occurring one) is that in which the X_i are independent. One simply simulates them individually, and that simulates X!

Another easy case is that in which X has a multivariate normal distribution. We noted in Section 3.6.2.1 that R includes the function **mvrnorm()**, which we can use to simulate our X here. The way this function works is to use the notion of principle components mentioned in Section 3.6.2.4. We construct Y = AX for the matrix A discussed there. The Y_i are independent, thus easily simulated, and then we transform back to X via $X = A^{-1}Y$

In general, though, things may not be so easy. For instance, consider the distribution in (3.16). There is no formulaic solution here, but the following strategy works.

First we find the (marginal) density of X. As in the case for Y shown in (3.19), we compute

$$f_X(s) = \int_0^s 8st \, dt = 4s^3 \tag{3.133}$$

[4] 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.

Using the method shown in our unit on continuous probability, Section 2.7, we can simulate X as

$$X = F_X^{-1}(W) \tag{3.134}$$

where W is a U(0,1) random variable, generated as **runif(1)**. Since $F_X(u) = u^4$, $F_X^{-1}(v) = v^{0.25}$, and thus our code to simulate X is

```
runif(1)^0.25
```

Now that we have X, we can get Y. We know that

$$f_{Y|X}(t|S) = \frac{8st}{4s^3} = \frac{2}{s^2}t \tag{3.135}$$

Remember, s is considered constant. So again we use the "inverse-cdf" method here to find Y, given X, and then we have our pair (X,Y).

3.8 Transform Methods (advanced topic)

We often use the idea of **transform** functions. For example, you may have seen **Laplace transforms** in a math or engineering course. The functions we will see here differ from this by just a change of variable.

Though in the form used here they involve only univariate distributions, their applications are often multivariate, as will be the case here.

3.8.0.5 Generating Functions

Let's start with the **generating function**. For any nonnegative-integer valued random variable V, its generating function is defined by

$$g_V(s) = E(s^V) = \sum_{i=0}^{\infty} s^i p_V(i),\ 0 \leq s \leq 1 \tag{3.136}$$

For instance, suppose N has a geometric distribution with parameter p, so that $p_N(i) = (1-p)p^{i-1}$, i = 1,2,... Then

$$g_N(s) = \sum_{i=1}^{\infty} s^i \cdot (1-p)p^{i-1} = \frac{1-p}{p} \sum_{i=1}^{\infty} s^i \cdot p^i = \frac{1-p}{p} \frac{ps}{1-ps} = \frac{(1-p)s}{1-ps} \tag{3.137}$$

Why restrict s to the interval [0,1]? The answer is that for $s > 1$ the series in (3.136) may not converge. for $0 \leq s \leq 1$, the series does converge. To see this, note that if s = 1, we just get the sum of all probabilities, which is 1.0. If a nonnegative s is less than 1, then s^i will also be less than 1, so we still have convergence.

One use of the generating function is, as its name implies, to generate the probabilities of values for the random variable in question. In other words, if you have the generating function but not the probabilities, you can obtain the probabilities from the function. Here's why: For clarify, write (3.136) as

$$g_V(s) = P(V = 0) + sP(V = 1) + s^2P(V = 2) + ... \tag{3.138}$$

From this we see that

$$g_V(0) = P(V = 0) \tag{3.139}$$

So, we can obtain P(V = 0) from the generating function. Now differentiating (3.136) with respect to s, we have

$$\begin{aligned} g'_V(s) &= \frac{d}{ds}\left[P(V = 0) + sP(V = 1) + s^2P(V = 2) + ...\right] \\ &= P(V = 1) + 2sP(V = 2) + ... \end{aligned} \tag{3.140}$$

So, we can obtain P(V = 2) from $g'_V(0)$, and in a similar manner can calculate the other probabilities from the higher derivatives.

3.8.0.6 Moment Generating Functions

The generating function is handy, but it is limited to discrete random variables. More generally, we can use the **moment generating function**, defined for any random variable X as

$$m_X(t) = E[e^{tX}] \tag{3.141}$$

for any t for which the expected value exists.

That last restriction is anathema to mathematicians, so they use the characteristic function,

$$\phi_X(t) = E[e^{itX}] \tag{3.142}$$

which exists for any t. However, it makes use of pesky complex numbers, so we'll stay clear of it here.

Differentiating (3.141) with respect to t, we have

$$m'_X(t) = E[Xe^{tX}] \tag{3.143}$$

We see then that

$$m'_X(0) = EX \tag{3.144}$$

So, if we just know the moment-generating function of X, we can obtain EX from it. Also,

$$m''_X(t) = E(X^2 e^{tX}) \tag{3.145}$$

so

$$m''_X(0) = E(X^2) \tag{3.146}$$

In this manner, we can for various k obtain $E(X^k)$, the **k**th **moment** of X, hence the name.

3.8.1 Example: Network Packets

As an example, suppose say the number of packets N received on a network link in a given time period has a Poisson distribution with mean μ, i.e.

$$P(N = k) = \frac{e^{-\mu}\mu^k}{k!}, k = 0, 1, 2, 3, ... \tag{3.147}$$

3.8.1.1 Poisson Generating Function

Let's first find its generating function.

$$g_N(t) = \sum_{k=0}^{\infty} t^k \frac{e^{-\mu}\mu^k}{k!} = e^{-\mu} \sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} = e^{-\mu+\mu t} \tag{3.148}$$

where we made use of the Taylor series from calculus,

$$e^u = \sum_{k=0}^{\infty} u^k/k! \tag{3.149}$$

3.8.1.2 Sums of Independent Poisson Random Variables Are Poisson Distributed

Supposed packets come in to a network node from two independent links, with counts N_1 and N_2, Poisson distributed with means μ_1 and μ_2. Let's find the distribution of $N = N_1 + N_2$, using a transform approach.

$$g_N(t) = E[t^{N_1+N_2}] = E[t^{N_1}]E[t^{N_2}] = g_{N_1}(t)g_{N_2}(t) = e^{-\nu+\nu t} \tag{3.150}$$

where $\nu = \mu_1 + \mu_2$.

But the last expression in (3.150) is the generating function for a Poisson distribution too! And since there is a one-to-one correspondence between distributions and transforms, we can conclude that N has a Poisson distribution with parameter ν. We of course knew that N would have mean ν but did not know that N would have a Poisson distribution.

So: A sum of two independent Poisson variables itself has a Poisson distribution. By induction, this is also true for sums of k independent Poisson variables.

3.8.1.3 Random Number of Bits in Packets on One Link (advanced topic)

Consider just one of the two links now, and for convenience denote the number of packets on the link by N, and its mean as μ. Continue to assume that N has a Poisson distribution.

Let B denote the number of bits in a packet, with $B_1, ..., B_N$ denoting the bit counts in the N packets. We assume the B_i are independent and identically distributed. The total number of bits received during that time period is

$$T = B_1 + ... + B_N \tag{3.151}$$

Suppose the generating function of B is known to be h(s). Then what is the generating function of T?

$$\begin{aligned}
g_T(s) &= E(s^T) && (3.152)\\
&= E[E(s^T|N)] && (3.153)\\
&= E[E(s^{B_1+...+B_N}|N)] && (3.154)\\
&= E[E(s^{B_1}|N)...E(s^{B_N}|N)] && (3.155)\\
&= E[h(s)^N] && (3.156)\\
&= g_N[h(s)] && (3.157)\\
&= e^{-\mu+\mu h(s)} && (3.158)
\end{aligned}$$

Here is how these steps were made:

- From the first line to the second, we used the Theorem of Total Expectation.
- From the second to the third, we just used the definition of T.
- From the third to the fourth lines, we have used algebra plus the fact that the expected value of a product of independent random variables is the product of their individual expected values.
- From the fourth to the fifth, we used the definition of h(s).
- From the fifth to the sixth, we used the definition of g_N.
- From the sixth to the last we used the formula for the generating function for a Poisson distribution with mean μ.

We can then get all the information about T we need from this formula, such as its mean, variance, probabilities and so on, as seen previously.

3.8.2 Other Uses of Transforms

Transform techniques are used heavily in queuing analysis, including for models of computer networks. The techniques are also used extensively in modeling of hardware and software reliability.

3.9 Vector Space Interpretations (for the mathematically adventurous only)

3.9.1 Properties of Correlation

Let $\mathcal{V}$ be the set of all random variables with finite variance and mean 0. Treat this as a vector space, with the sum of two vectors X and Y taken to be the random variable X+Y, for a constant c, the vector cX being the random variable cX. Note that $\mathcal{V}$ is closed under these operations, as it must be.

Define an inner product on this space:

$$(X,Y) = E(XY) = Cov(X,Y) \tag{3.159}$$

(Recall that Cov(X,Y) = E(XY) - EX EY, and that we are working with random variables that have mean 0.) Thus the norm of a vector X is

$$||X|| = (X,X)^{0.5} = \sqrt{E(X^2)} = \sqrt{Var(X)} \tag{3.160}$$

again since E(X) = 0.

The famous Cauchy-Schwarz Inequality for inner products says,

$$|(X,Y)| \leq ||X||\ ||Y|| \tag{3.161}$$

i.e.

$$|\rho(X,Y)| \leq 1 \tag{3.162}$$

Also, the Cauchy-Schwarz Inequality yields equality if and only if one vector is a scalar multiple of the other, i.e. Y = cX for some c. When we then translate this to random variables of nonzero means, we get Y = cX + d.

In other words, the correlation between two random variables is between -1 and 1, with equality if and only if one is an exact linear function of the other.

3.9.2 Conditional Expectation As a Projection

Continue to consider the vector space in Section 3.9.1.

For a random variable X, let $\mathcal{W}$ denote the subspace of $\mathcal{V}$ consisting of all functions h(X) with mean 0 and finite variance. (Again, note that this subspace is indeed closed under vector addition and scalar multiplication.)

Now consider any Y in $\mathcal{V}$. Recall that the *projection* of Y onto $\mathcal{W}$ is the closest vector T in $\mathcal{W}$ to Y, i.e. T minimizes

$$||Y-T|| = \left(E[(Y-T)^2]\right)^{0.5} \tag{3.163}$$

To find the minimizing T, consider first the minimization of

$$E[(S-c)^2] \tag{3.164}$$

with respect to constants c for some random variable S. Expanding the square, we have

$$E[(S-c)^2] = E(S^2) - 2cES + (ES)^2 \tag{3.165}$$

Taking $\frac{d}{dc}$ and setting the result to 0, we find that the minimizing c is c = ES.

Getting back to (3.163), use the Law of Total Expectation to write

$$E[(Y-T)^2] = E\left(E[(Y-T)^2|X]\right) \tag{3.166}$$

From what we learned with (3.164), applied to the conditional (i.e. inner) expectation in (3.166), we see that the T which minimizes (3.166) is T = E(Y|X).

In other words, the conditional mean is a projection! Nice, but is this useful in any way? The answer is yes, in the sense that it guides the intuition. All this is related to issues of statistical prediction—here we would be predicting Y from X—and the geometry here can really guide our insight. This is not very evident without getting deeply into the prediction issue, but let's explore some of the implications of the geometry.

For example, a projection is perpendicular to the line connecting the projection to the original vector. So

$$0 = (E(Y|X), Y - E(Y|X)) = Cov[E(Y|X),\ Y - E(Y|X)] \tag{3.167}$$

This says that the prediction E(Y|X) is uncorrelated with the prediction error, Y-E(Y|X). This in turn has statistical importance. Of course, (3.167) could have been derived directly, but the geometry of the vector space intepretation is what suggested we look at the quantity in the first place. Again, the point is that the vector space view can guide our intuition.

Simlarly, the Pythagorean Theorem holds, so

$$||Y||^2 = ||E(Y|X)||^2 + ||Y - E(Y|X)||^2 \tag{3.168}$$

which means that

$$Var(Y) = Var[E(Y|X)] + Var[Y - E(Y|X)] \tag{3.169}$$

Equation (3.169) is a common theme in linear models in statistics, the decomposition of variance.

3.10 Proof of the Law of Total Expectation

Let's prove (3.88) for the case in which W and Y take values only in the set {1,2,3,...}. Recall that if T is an integer-value random variable and we have some function h(), then L = h(T) is another random variable, and its expected value can be calculated as[5]

$$E(L) = \sum_k h(k)P(T = k) \tag{3.170}$$

In our case here, Q is a function of W, so we find its expectation from the distribution of W:

$$\begin{aligned}
E(Q) &= \sum_{i=1}^{\infty} g(i)P(W = i) \\
&= \sum_{i=1}^{\infty} E(Y|W = i)P(W = i) \\
&= \sum_{i=1}^{\infty} \left[\sum_{j=1}^{\infty} jP(Y = j|W = i) \right] P(W = i) \\
&= \sum_{j=1}^{\infty} j \sum_{i=1}^{\infty} P(Y = j|W = i)P(W = i) \\
&= \sum_{j=1}^{\infty} jP(Y = j) \\
&= E(Y)
\end{aligned}$$

In other words,

$$E(Y) = E[E(Y|W)] \tag{3.171}$$

Exercises

1. Suppose the random pair (X, Y) has the density $f_{X,Y}(s, t) = 8st$ on the triangle $\{(s, t) : 0 < t < s < 1\}$.

(a) Find $\rho(X, Y)$ and $f_X(s)$.

(b) Consider the bivariate density (3.1.3.3). Find $P(X < Y/2)$.

[5] This is sometimes called The Law of the Unconscious Statistician, by nasty probability theorists who look down on statisticians. Their point is that technically $EL = \sum_k kP(L = k)$, and that (3.170) must be proven, whereas the statisticians supposedly think it's a definition.

2. Suppose packets on a network are of three types. In general, 40% of the packets are of type A, 40% have type B and 20% have type C. We observe six packets, and denote the numbers of packets of types A, B and C by X, Y and Z, respectively.

(a) Find P(X = Y = Z = 2).

(b) Find Cov(X,Y+Z).

(c) Which one of the following is the distribution of Y+Z?

(i) Binomial.
(ii) Multinomial.
(iii) Negative binomial.
(iv) Poisson.
(v) Uniform.
(vi) Exponential.
(vii) A distribution not listed above.
(viii) None; Y+Z doesn't have a distribution.

3. In the catchup game in Section 3.2.4, let V and W denote the winnings of the two players after only one turn. Find $P(V > 0.4)$.

4. Suppose X and Y are independent, each having an exponential distribution with means 1.0 and 2.0, respectively.

(a) Find $P(Y > X^2)$.

(b) Fill in the subscripts for f and the upper limit in the second integral:

(i) $f_{\boxed{}}(t) = \int_0^t e^{-s} \cdot 0.5e^{-0.5(t-s)} \, ds$

(ii) $f_{\boxed{}}(t) = \int_0^{\boxed{}} e^{-s} \cdot 0.5e^{-0.5(t/s)} \, ds$

5. Bus lines A and B intersect at a certain transfer point, with the schedule stating that buses from both lines will arrive there at 3:00 p.m. However, they are often late, by amounts X and Y, measured in hours, for the two buses. The bivariate density is

$$f_{X,Y}(s,t) = 2 - s - t, \; 0 < s,t < 1 \tag{3.172}$$

Two friends agree to meet at the transfer point, one taking line A and the other B. Let W denote the time in minutes the person arriving on line B must wait for the friend.

(a) Show that X and Y are not independent, by evaluating $P(X \in A \text{ and } Y \in B)$, $P(X \in A)$ and $P(Y \in B)$ for some sets A and B.

(b) Find $P(W > 6)$.

(c) Find EW. Note that W is neither fully discrete nor fully continuous. We have not developed machinery for this, but the Law of Total Expectation will give you what you need here.

6. Suppose the pair (X,Y)' has a bivariate normal distribution with mean vector (0,2) and covariance matrix $\begin{pmatrix} 1 & 2 \\ 2 & 6 \end{pmatrix}$

(a) Set up (but do not evaluate) the double integral for the exact value of $P(X^2 + Y^2 \leq 2.8)$.

(b) Using the matrix methods of Section 3.4, find the covariance matrix of the pair U = (X+Y,X-2Y)'. Does U have a bivariate normal distribution?

7. Show that

$$\rho(aX + b, cY + d) = \rho(X, Y) \tag{3.173}$$

for any constants a, b, c and d.

8. Suppose X are Y independent, and each has a U(0,1) distribution. Let V = X + Y.

(a) Find f_V. (Advice: It will be a “two-part function,” i.e. the type we have to describe by saying something like, “The function has value 2z for z < 6 and 1/z for z > 6.”)

(b) Verify your answer in (a) by finding EV from your answer in (a) and then using the fact that EX = EY = 0.5.

9. Suppose the following:

- In the general population of parents who have 10-year-old kids, the parent/kid weight pairs have an exact bivariate normal distribution.
- Parents’ weights have mean 152.6 and standard deviation 25.0.

- Weights of kids have mean 62 and standard deviation 6.4.
- The correlation between the parents' and kids' weights is 0.4.

Use R functions (not simulation) in the following:

(a) Find the fraction of parents who weigh more than 160.

(b) Find the fraction of kids who weigh less than 56.

(c) Find the fraction of parent/child pairs in which the parent weighs more than 160 and the child weighs less than 56.

(d) Suppose a ride at an amusement park charges by weight, one cent for each pound of weight in the parent and child. State the exact distribution of the fee, and find the fraction of parent/child pairs who are charged less than $2.00.

10. Suppose X, Y and Z are "i.i.d." (independent, identically distributed) random variables, with $E(X^k)$ being denoted by ν_k, i = 1,2,3. Find Cov(XY,XZ) in terms of the ν_k.

11. Using the properties of covariance in Section 3.2.1, show that for any random variables X and Y, Cov(X+Y,X-Y) = Var(X) - Var(Y).

12. Newspapers at a certain vending machine cost 25 cents. Suppose 60% of the customers pay with quarters, 20% use two dimes and a nickel, 15% insert a dime and three nickels, and 5% deposit five nickels. When the vendor collects the money, five coins fall to the ground. Let X, Y amd Z denote the numbers of quarters, dimes and nickels among these five coins.

(a) Is the joint distribution of (X, Y, Z) a member of a parametric family presented in this chapter? If so, which one?

(b) Find $P(X = 2, Y = 2, Z = 1)$.

(c) Find $\rho(X, Y)$.

Hint: First find the proportion of quarters, among all coins deposited in this machine generally.

13. Suppose we wish to predict a random variable Y by using another random variable, X. We may consider predictors of the form $cX + d$ for constants c and d. Show that the values of c and d that minimize the mean squared prediction error, $E[(Y - cX - d)^2$ are

$$c = \frac{E(XY) - EX \cdot EY}{Var(X)} \tag{3.174}$$

$$d = \frac{E(X^2) \cdot EY - EX \cdot E(XY)}{Var(X)} \tag{3.175}$$

14. Programs A and B consist of r and s modules, respectively, of which c modules are common to both. As a simple model, assume that each module has probability p of being correct, with the modules acting independently. Let X and Y denote the numbers of correct modules in A and B, respectively. Find the correlation (X, Y) as a function of r, s, c and p.

Hint: Write $X = X_1 + ...X_{r-c}$, where X_i is 1 or 0, depending on whether module i of A is correct, for the nonoverlapping modules of A. Do the same for B, and for the set of common modules.

15. Show that if random variables U and V are independent,

$$Var(UV) = E(U^2) \cdot Var(V) + Var(U) \cdot (EV)^2 \tag{3.176}$$

16. Use transform methods to derive some properties of the Poisson family:

(a) Show that for any Poisson random variable, its mean and variance are equal.

(b) Suppose X and Y are independent random variables, each having a Poisson distribution. Show that $Z = X + Y$ again has a Poisson distribution.

17. Suppose one keeps rolling a die. Let S_n denote the total number of dots after n rolls, mod 8, and let T be the number of rolls needed for the event $S_n = 0$ to occur. Find $E(T)$, using an approach like that in the "trapped miner" example in Section 3.5.5.

18. In our ordinary coins which we use every day, each one has a slightly different probability of heads, which we'll call H. Say H has the distribution $N(0.5, 0.03^2)$. We choose a coin from a batch at random, then toss it 10 times. Let N be the number of heads we get. Find $Var(N)$.

19. Jack and Jill play a dice game, in which one wins \$1 per dot. There are three dice, die A, die B and die C. Jill always rolls dice A and B. Jack always rolls just die C, but he also gets credit for 90% of die B. For instance, say in a particular roll A, B and C are 3, 1 and 6, respectively. Then Jill would win \$4 and Jack would get \$6.90. Let X and Y be Jill's and Jack's total winnings after 100 rolls. Use the Central Limit Theorem to find the approximate values of $P(X > 650, Y < 660)$ and $P(Y > 1.06X)$.

Hints: This will follow a similar pattern to the dice game in Section 3.6.2.3, which we win \$5 for one dot, and \$2 for two or three dots. Remember, in that example, the key was that we noticed that the pair (X, Y) was a sum of random pairs. That meant that (X, Y) had an approximate bivariate normal distribution, so we could find probabilities if we had the mean vector and covariance matrix of (X, Y). Thus we needed to find $EX, EY, Var(X), Var(Y)$ and $Cov(X, Y)$. We used the various properties of $E(), Var()$ and $Cov()$ to get those quantities.

You will do the same thing here. Write $X = U_1 + ... + U_{100}$, where U_i is Jill's winnings on the ith roll. Write Y as a similar sum of V_i. You probably will find it helpful to define A_i, B_i and C_i as the numbers of dots appearing on dice A, B and C on the ith roll. Then find EX etc. Again, make sure to utilize the various properties for $E(), Var()$ and $Cov()$.

20. Suppose the number N of bugs in a certain number of lines of code has a Poisson distribution, with parameter L, where L varies from one programmer to another. Show that Var(N) = EL + Var(L).

21. This problem arises from the analysis of random graphs, which for concreteness we will treat here as social networks such as Facebook.

In the model here, each vertex in the graph has N friends, N being a random variable with the same distribution at every vertex. One thinks of each vertex as generating its links, unterminated, i.e. not tied yet to a second vertex. Then the unterminated links of a vertex pair off at random with those of other vertices. (Those that fail will just pair in self loops, but we'll ignore that.)

Let M denote the number of friends a friend of mine has. That is, start at a vertex A, and follow a link from A to another vertex, say B. M is the number of friends B has (we'll include A in this number).

(a) Since an unterminated link from A is more likely to pair up with a vertex that has a lot of links, a key assumption is that P(M = k) = ck P(N = k) for some constant c. Fill in the blank: This is an example of the setting we studied called __________________.

(b) Show the following relation of generating functions: $g_M(s) = g'_N(s)/EN$.

22. Suppose Type 1 batteries have exponentially distributed lifetimes with mean 2.0 hours, while Type 2 battery lifetimes are exponentially distributed with mean 1.5. We have a large box containing a mixture of the two types of batteries, in proportions q and 1-q. We reach into the box, choose a battery at random, then use it. Let Y be the lifetime of the battery we choose. Use the Law of Total Variance, (3.91), to find $Var(Y)$.

Chapter 4

Introduction to Statistical Inference

4.1 What Statistics Is All About

Consider these problems:

- Suppose you buy a ticket for a lottery, 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 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 use data on forest fires to estimate the probability that this is a fire?

If you thought 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 7.

4.2 The “Margin of Error”: Introduction to Confidence Intervals

4.2.1 How Long Should We Run a Simulation?

In our simulations in previous units, it was never quite clear how long the simulation should be run, i.e. what value to set for **nreps** in Section 1.2.6.1. Now we will finally address this issue.

As our example, recall from the Bus Paradox in Section 2.5: Buses arrive at a certain bus stop at random times, with interarrival times being independent exponentially distributed random variables with mean 10 minutes. You arrive at the bus stop every day at a certain time, say four hours (240 minutes) after the buses start their morning run. What is your mean wait for the next bus?

We 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. We could write a program:

```
doexpt <- function(opt) {
   lastarrival <- 0.0
   while (lastarrival < opt)
      lastarrival <- lastarrival + rexp(1,0.1)
   return(lastarrival-opt)
}

observationpt <- 240
nreps <- 1000
waits <- vector(length=nreps)
for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
cat("approx. mean wait = ",mean(waits),"\n")
```

Running the program yields

```
approx. mean wait = 9.653743
```

Was 1000 iterations enough? How close is this value 9.653743 to the true expected value of waiting time?[1]

What we would like to do is something like what the pollsters do during presidential elections, when they say “Ms. X is supported by 62% of the voters, with a margin of error of 4%.” In fact, we will do exactly this, in the next section.

4.2.2 Confidence Intervals for Means

The goal of this section (and several that follow) is to develop a notion of margin of error, just as you

[1] Of course, continue to ignore the fact that we know that this value is 10.0. What we’re trying to do here is figure out how to answer “how close is it” questions in general, when we don’t know the true mean.

see in the election campaign polls. This raises two questions:

(a) What do we mean by “margin of error”?

(b) How can we calculate it?

The answer to (a) is that we would like to make a statement, e.g. in the simulation output above, like “We estimate the mean wait to be 9.65, and we are 95% confident that the true population mean wait is between 9.65-0.22 and 9.65+22.”

Our answer to (b) will now be developed, in the next two subsections.

4.2.2.1 Sampling Distributions

In our example in Section 4.2.1, let W denote the random wait time one experiences in general in this situation. We are using the program to estimate $E(W)$, which we will denote by μ. While we're at it, let's denote $Var(W)$ by σ^2. (Remember, this does NOT mean that W has a normal distribution. The use of μ and σ^2 to denote mean and variance is standard, regardless of distribution, which in fact is exponential here, not normal.)

So, what if estimate EW and Var(W) by sampling for n days, either by actually waiting at the bus stop, or by simulating n days as we did in the above program (where our n was **nreps**)? How accurate are our estimates?

Let W_i denote the i^{th} waiting time, i = 1,2,...,n and let $\overline{W}$ denote the sample mean,

$$\overline{W} = \frac{W_1 + ...W_n}{n} \tag{4.1}$$

$\overline{W}$ is what the program prints out.

At this point we focus on the distribution of $\overline{W}$. **Note carefully** what this means in the notebook sense. Our simulation program outputs one $\overline{W}$, so now one line of the notebook corresponds to one run of the program. So for example E($\overline{W}$) is the long-run average of $\overline{W}$ through infinitely many lines of the notebook.

Say Michael runs the program, with **nreps** (again, this is n in (4.1) above) equal to 1000. His program will generate $W_1, ..., W_{1000}$, as well as $\overline{W}$. There will thus be 1001 entries on the first line of our notebook. Next Eric runs the program, resulting in 1001 entries in the second line of the notebook. Then Ying-Chi then runs the program, then Ben and so on, each recording the results in his/her own line in the notebook.

The notebook has 1001 columns, labeled $W_1, ..., W_{1000}, \overline{W}$. Then for instance $P(\overline{W} > 15.1)$ is the long-run proportion of entries in the $\overline{W}$ column that are greater than 15.1.

Note that the long-run fraction of entries in the W_1 column that are greater than 12.2 is still $P(W > 12.2) = 0.295$, etc. The same is true of the W_2 column, the W_3 column and so on.

So, $\overline{W}$ is a random variable. Thus in order to achieve our goal of finding a margin of error, we need to know the distribution of that random variable, i.e. the long-run behavior in the $\overline{W}$ column of the notebook. Since $\overline{W}$ is a constant multiple of a sum, it is approximately normally distributed for large n, and we'll make use of that fact.

The key points are that

- As seen above, The random variables W_i each have the distribution F_W, and thus each have mean $EW = \mu$ and variance $Var(W) = \sigma^2$.
- The random variables W_i are independent.
- The mean of $\overline{W}$ is also μ:

$$
\begin{aligned}
E(\overline{W}) &= \frac{1}{n} E\left(\sum_{i=1}^{n} W_i\right) \ \text{(for const. c, } E(cU) = cEU) && (4.2)\\
&= \frac{1}{n} \sum_{i=1}^{n} EW_i \ \ (E[U+V] = EU + EV) && (4.3)\\
&= \frac{1}{n} n\mu \ \ (EW_i = \mu) && (4.4)\\
&= \mu && (4.5)
\end{aligned}
$$

- The variance of $\overline{W}$ is 1/n of the population variance:

$$
\begin{aligned}
Var(\overline{W}) &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} W_i\right) \ \text{(for const. c, , } Var[cU] = c^2 Var[U]) && (4.6)\\
&= \frac{1}{n^2} \sum_{i=1}^{n} Var(W_i) \ \ \text{(for U,V indep., , } Var[U+V] = Var[U] + Var[V]) && (4.7)\\
&= \frac{1}{n^2} n\sigma^2 && (4.8)\\
&= \frac{1}{n} \sigma^2 && (4.9)
\end{aligned}
$$

Recall that W here has an exponential distribution with mean 10.0. If you recall, any exponential distributions has its variance equal to the square of its mean, in this case 100.0. Then (4.5) says that the mean of all the numbers in the $\overline{W}$ column of our notebook will be 10.0, and the variance of all those numbers will be $100.0/1000^2 = 0.001$.

These points are absolutely key, forming the very basis of statistics. You should spend extra time pondering them.

4.2.2.2 Our First Confidence Interval

The Central Limit Theorem then tells us that

$$Z = \frac{\overline{W} - \mu}{\sigma/\sqrt{n}} \tag{4.10}$$

has an approximately N(0,1) distribution.[2] We will be interested in the central 95% of that distribution, which due to symmetry have 2.5% of the 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 there cuttoff points are at -1.96 and 1.96. Thus

$$0.95 \approx P\left(-1.96 < \frac{\overline{W} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \tag{4.11}$$

Doing a bit of algebra on the inequalities yields

$$0.95 \approx P\left(\overline{W} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{W} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{4.12}$$

Now remember, not only do we not know μ, we also don’t know σ. But we can estimate it, as follows:

Recall that by definition

$$\sigma^2 = E[(W - \mu)^2] \tag{4.13}$$

Let’s estimate σ^2 by taking sample analogs. The sample analog of μ is $\overline{W}$. What about the sample analog of the “E()”? Well, since E() averaging over the whole population of Ws, the sample analog is to average over the sample. So, we get

$$\frac{1}{n}\sum_{i=1}^{n}(W_i - \overline{W})^2 \tag{4.14}$$

[2] Note that it is the “N” here tht is approximate, not the 0 or 1.

In other words, just as it is natural to estimate the population mean of W by its sample mean, the same holds for Var(W):

> The population variance of W is the mean squared distance from W to its population mean. Therefore it is natural to estimate Var(W) by the average squared distance of W from its sample mean, among our sample values W_i, shown in (4.14).[3]

We use s^2 as our symbol for this estimate of population variance.[4] We thus take our estimate of σ to be s, the square root of that quantity.

By the way, (4.14) is equal to

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} W_i^2 - \overline{W}^2 \tag{4.15}$$

(Caution: This way of computing s^2 is subject to more roundoff error.)

One can show (the details will be given later in this section) that (4.12) is still valid if we substitute s for σ, i.e.

$$0.95 \approx P \left(\overline{W} - 1.96 \frac{s}{\sqrt{n}} < \mu < \overline{W} + 1.96 \frac{s}{\sqrt{n}} \right) \tag{4.16}$$

In other words, we are about 95% sure that the interval

$$(\overline{W} - 1.96 \frac{s}{\sqrt{n}}, \overline{W} + 1.96 \frac{s}{\sqrt{n}}) \tag{4.17}$$

contains μ. This is called a 95% **confidence interval** for μ. The quantity $1.96 \frac{s}{\sqrt{n}}$ is the margin of error.

We could add this feature to our program:

```
doexpt <- function(opt) {
   lastarrival <- 0.0
   while (lastarrival < opt)
      lastarrival <- lastarrival + rexp(1,0.1)
   return(lastarrival-opt)
}

```

[3] Note the similarity to (1.46).

[4] 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.

```
observationpt <- 240
nreps <- 10000
waits <- vector(length=nreps)
for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
wbar <- mean(waits)
cat("approx. mean wait =",wbar,"\n")
s2 <- mean(waits^2) - wbar^2
s <- sqrt(s2)
radius <- 1.96*s/sqrt(nreps)
cat("approx. CI for EW =",wbar-radius,"to",wbar+radius,"\n")
```

When I ran this, I got 10.02565 for the estimate of EW, and got an interval of (9.382715, 10.66859). We would then say, “We are about 95% confident that the true mean wait time until the next bus is between 9.38 and 10.67.”

What does this really mean? This question is of the utmost importance. We will devote Section 4.2.3 to it.

4.2.3 Meaning of Confidence Intervals

4.2.3.1 A Weight Survey in Davis

Consider the question of measuring the mean weight, denoted by μ, of all adults in the city of Davis. Say we sample 1000 people at random, and record their weights, with W_i being the weight of the i^{th} person in our sample.[5]

Now remember, we don’t know the true value of that population mean, μ—once again, that’s why we are collecting the sample data, to estimate μ! Our estimate will be our sample mean, $\overline{\mathbf{W}}$. But we don’t know how accurate that estimate might be. That’s the reason we form the confidence interval, as a gauge of the accuracy of $\overline{\mathbf{W}}$ as an estimate of μ.

Say our interval (4.17) 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

[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 4.2.2, i.e. the letter W, so we’ll have our example involve people’s weights instead of heights. It works out neatly, because the word *weight* has the same sound as *wait*.

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.

4.2.3.2 Back to Our Bus Simulation

Well, in our simulation case, it is *exactly the same situation*. Simulation is a sampling process. Our μ is the mean in the "population" of all bus waits, while $\overline{W}$ is the mean in our sample of 1000 waits. This is not mere analogy; mathematically the two situations are completely identical, two instances of the same principle.

Let's use the "you and your 99 friends" idea again. Supposed each of you 100 people run the R program at the end of Section 4.2.2.2. Each of you will get a different confidence interval printed out at the end of your run.[6] Well, when we say that the program prints out a 95% confidence interval, we mean that about 95 of you 100 people will have an interval that contains the true value of EW.

In the Davis weight example above, I stressed that we don't know μ—after all, that's the reason we are taking a sample of people, so as to estimate μ!

Similarly, the whole point of doing a simulation to find some quantity E(R) is that we don't know the value of E(R)! We will simulate many values of R, forming $\overline{R}$, and use that quantity as an estimate of ER.

But our bus example was just that—an *example*, set up to illustrate the notion of adding a confidence interval to the output of a simulation. We actually do know the value of EW here; it's 10. That makes this a rather artificial example, but that's good, because it will allow us to really see the "you and 99 friends" idea in action, as follows.

We'll expand the code to simulate 1000 people running the original program. In other words, we'll add an

[6] Recall that R will generate a different stream of random numbers each time you run your program, unless you call **set.seed()**.

extra outer loop to do 1000 runs of the program. Each run will compute the confidence interval, and then we’ll see in the end how many of the 1000 runs have a confidence interval that includes the true EW, 10.0:

```
doexpt <- function(opt) {
   lastarrival <- 0.0
   while (lastarrival < opt)
      lastarrival <- lastarrival + rexp(1,0.1)
   return(lastarrival-opt)
}

observationpt <- 240
nreps <- 1000
numruns <- 1000
waits <- vector(length=nreps)
numcorrectcis <- 0  # number of conf. ints. that contain 10.0
for (run in 1:numruns) {
   for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
   wbar <- mean(waits)
   s2 <- mean(waits^2) - wbar^2
   s <- sqrt(s2)
   radius <- 1.96*s/sqrt(nreps)
   if (abs(wbar - 10.0) <= radius) numcorrectcis <- numcorrectcis + 1
}
cat("approx. true confidence level=",numcorrectcis/numruns,"\n")
```

In fact, the output of that program was 0.958, sure enough about 95%.

Why is it not exactly 0.95?

- We only simulated 1000 runs of the program; ideally it should be an infinite number, to get the exact probability that an interval contains μ.
- The Central Limit Theorem is only approximate.
- Ideally we would use (4.12), but due to lack of knowledge of the true value of σ (we don’t know μ, so why would we know σ?), we resorted to using s instead, in (4.17).

Again remember that in practice we only do *one* run of simulating 1000 waits for the bus. Our simulation code above 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.

4.2.3.3 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.”[7]

[7] See for example the Wikipedia entry, “Confidence Intervals,” `http://en.wikipedia.org/wiki/Confidence_interval#Meaning_and_interpretation`.

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, 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.

4.2.4 Sampling With and Without Replacement

Implicit in our analyses so far in our assumption that the W_i are independent is that we are sampling **with replacement**, which means it's possible in the Davis weights example that our random sampling process might choose the same person twice.

If we sample with replacement, we say that we have a **random sample**. If it is done without replacement, it's called a **simple random sample**. In the latter case, (4.9) does not hold, because the W_i are not independent (though they are still identically distributed). To see this, suppose that Davis were a tiny town consisting of just three adults, with weights 120, 161 and 190. Then if for example $W_1 = 190$, then $E(W_2|W_1) = (120+161)/2 = 140.5$, while $E(W_1) = (120+161+190)/3 = 157$. Thus W_1 and W_2 are not independent, and (4.9) would fail.[8]

But except for cases in which our sample size is a substantial fraction of the population size, the probability of getting the same person twice would be very low, so it doesn't matter. Thus we can safely use analyses which assume with-replacement sampling even if we are using without-replacement sampling.

4.2.5 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 (4.17),we multiply by 1.65 instead of by 1.96,

[8] Note, though, that (4.5) *does* hold, because expected values of sums equal sums of expected values even for dependent random variables.

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 $\alpha = 0.05$.

4.2.6 General Margins of Error: the Standard Error of the Estimate

Recall that the idea of a confidence interval is really simple: We report our estimate, plus or minus a margin of error. In (4.17),

margin of error = $1.96\times$ estimated standard deviation of $\overline{W}$

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 squart root of that quantity, i.e. the standard deviation of $\overline{W}$. In (4.9), 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** (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.

We can see from (4.17) what to do in general, if we are estimating some number θ by $\widehat{\theta}$,[9] and if (note this qualifier) the latter has an approximately normal distribution. Let $s.e.(\widehat{\theta})$ denote our estimate for the standard deviation of that distribution, i.e. the standard error of $\widehat{\theta}$, obtained somehow. Then an approximate 95% confidence interval for θ is

$$\widehat{\theta} \pm 1.96 \cdot \text{s.e.}(\widehat{\theta}) \tag{4.18}$$

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. Make sure you understand what it means and how it is used.

4.2.7 Why Not Divide by n-1? The Notion of Bias

It should be noted that it is customary in (4.14) to divide by n-1 instead of n, for reasons that are largely historical. Here's the issue:

[9] The quantity is pronounced “theta-hat.” The “hat” symbol is traditional for “estimate of.”

If we divide by n, as we have been doing, then it turns out that

$$E(s^2) = \frac{n-1}{n} \cdot \sigma^2 \tag{4.19}$$

Think about this in the Davis people example, once again in the notebook context. Remember, here n is 1000, and each line of the notebook represents our taking a different random sample of 1000 people. Within each line, there will be entries for W_1 through W_{1000}, the weights of our 1000 people, and for $\overline{W}$ and s. For convenience, let's suppose we record that last column as s^2 instead of s.

Now, say we want to estimate the population variance σ^2. As discussed earlier, the natural estimator for it would be the sample variance, s^2. What (4.19) says is that after looking at an infinite number of lines in the notebook, the average value of s^2 would be just...a...little...bit...too...small. All the s^2 values would average out to $0.999\sigma^2$, rather than to σ^2. We might say that s^2 has a little bit more tendency to underestimate σ^2 than to overestimate it.

We say that s^2 is a **biased** estimator of σ^2, with the amount of bias being

$$E(s^2) = \frac{1}{n} \cdot \sigma^2 \tag{4.20}$$

Let's prove (4.19). We'll use (4.15). As before, let W be a random variable distributed as the population. Write the first term as

$$\begin{aligned}
E\left(\frac{1}{n}\sum_{i=1}^{n} W_i^2\right) &= \frac{1}{n} E \sum_{i=1}^{n} W_i^2 \quad \text{(constants factor out of E())} && (4.21)\\
&= \frac{1}{n} \cdot nE(W^2) \quad \text{(each } W_i \text{ has distr. of W)} && (4.22)\\
&= E(W^2) && (4.23)\\
&= Var(W) + (EW)^2 && (4.24)\\
&= \sigma^2 + \mu^2 && (4.25)
\end{aligned}$$

Continuing to work from (4.15) and using (4.9), write

$$E[\overline{W}^2] = Var(\overline{W}) + [E(\overline{W})]^2 = \frac{1}{n}\sigma^2 + \mu^2 \tag{4.26}$$

Now using all this in (4.15), we get

$$E(s^2) = \frac{n-1}{n}\sigma^2 \tag{4.27}$$

The earlier developers of statistics were bothered by this bias, so they introduced a "fudge factor" by dividing by n-1 instead of n in (4.14). We will call that $\tilde{s}^2$:

$$\tilde{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (W_i - \overline{W})^2 \tag{4.28}$$

This is the "classical" definition of sample variance, in which we divide by n-1 instead of n.

But we will use n. After all, when n is large—which is what we are assuming by using the Central Limit Theorem in the entire development so far—it doesn't make any appreciable difference. Clearly it is not important in our Davis example, or our bus simulation example.

Moreover, speaking generally now rather than necessarily for the case of s^2 there is no particular reason to insist that an estimator be unbiased anyway. An alternative estimator may have a little bias but much smaller variance, and thus might be preferable. And anyway, even though the classical version of s^2, i.e. $\tilde{s}^2$, is an unbiased estimator for σ^2, s is not an unbiased estimator for σ, the population standard deviation. In other words, unbiasedness is not such an important property.

The R functions **var()** and **sd()** calculate the versions of s^2 and s, respectively, that have a divisor of n-1.

4.2.8 And What About the Student-t Distribution?

Another thing we are not doing here is to use the **Student t-distribution**. That is the name of the distribution of the quantity

$$T = \frac{\overline{W} - \mu}{\tilde{s}/\sqrt{n}} \tag{4.29}$$

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 (4.11), 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:

$$(\overline{W} - 2.26\frac{\tilde{s}}{\sqrt{10}}, \overline{W} + 2.26\frac{\tilde{s}}{\sqrt{10}}) \tag{4.30}$$

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.

4.2.9 Confidence Intervals for Proportions

In our bus example above, suppose we also want our simulation to print out the (estimated) probability that one must wait longer than 6.4 minutes:

```
doexpt <- function(opt) {
   lastarrival <- 0.0
   while (lastarrival < opt)
      lastarrival <- lastarrival + rexp(1,0.1)
   return(lastarrival-opt)
}

observationpt <- 240
nreps <- 1000
waits <- vector(length=nreps)
for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
wbar <- mean(waits)
cat("approx. mean wait =",wbar,"\n")
s2 <- (mean(waits^2) - wbar^2)
s <- sqrt(s2)
radius <- 1.96*s/sqrt(nreps)
cat("approx. CI for EW =",wbar-radius,"to",wbar+radius,"\n")
prop <- length(waits[waits > 6.4]) / nreps
cat("approx. P(W > 6.4) =",prop,"\n")
```

The value printed out for the probability is 0.516. We again ask the question, how can we gauge the accuracy of this number as an estimator of the true probability $P(W > 6.4)$?

4.2.9.1 Derivation

It turns out that we already have our answer, because a probability is just a special case of a mean. To see this, let

$$Y = \begin{cases} 1, & \text{if } W > 6.4 \\ 0, & \text{otherwise} \end{cases} \tag{4.31}$$

Then

$$E(Y) = 1 \cdot P(Y = 1) + 0 \cdot P(Y = 0) = P(W > 6.4) \tag{4.32}$$

Let p denote this probability, and let $\widehat{p}$ denote our estimate of it; $\widehat{p}$ is our **prop** in the program. In (4.15), take W_i to be our Y_i here, and note that $Y_i^2 = Y_i$. That means that

$$s^2 = \widehat{p} - \widehat{p}^2 = \widehat{p}(1 - \widehat{p}) \tag{4.33}$$

Equation (4.17) becomes

$$\left(\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}, \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n}\right) \tag{4.34}$$

4.2.9.2 Examples

We incorporate that into our program:

```
doexpt <- function(opt) {
   lastarrival <- 0.0
   while (lastarrival < opt)
      lastarrival <- lastarrival + rexp(1,0.1)
   return(lastarrival-opt)
}

observationpt <- 240
nreps <- 1000
waits <- vector(length=nreps)
for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
wbar <- mean(waits)
cat("approx. mean wait =",wbar,"\n")
s2 <- (mean(waits^2) - mean(wbar)^2)
s <- sqrt(s2)
radius <- 1.96*s/sqrt(nreps)
cat("approx. CI for EW =",wbar-radius,"to",wbar+radius,"\n")
```

```
prop <- length(waits[waits > 6.4]) / nreps
s2 <- prop*(1-prop)
s <- sqrt(s2)
radius <- 1.96*s/sqrt(nreps)
cat("approx. P(W > 6.4) =",prop,", with a margin of error of",radius,"\n")
```

In this case, we get margin of error of 0.03, thus an interval of (0.51,0.57). We would say, "We don't know the exact value of $P(W > 6.4)$, so we ran a simulation. The latter estimates this probability to be 0.54, with a 95% margin of error of 0.03."

Note again that this uses the same principles as our Davis weights example. Suppose we were interested in estimating the proportion of adults in Davis who weigh more than 150 pounds. Suppose that proportion is 0.45 in our sample of 1000 people. This would be our estimate $\widehat{p}$ for the population proportion p, and an approximate 95% confidence interval (4.34) 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 (4.34).

4.2.9.3 Interpretation

The same interpretation holds as before. Consider the examples in the last section:

- If each of you and 99 friends were to run the R program at the beginning of Section 4.2.9.2, you 100 people would get 100 confidence intervals for $P(W > 6.4)$. About 95 of you would have intervals that do contain that number.
- If each of you and 99 friends were to sample 1000 people in Davis and come up with confidence intervals for the true population proportion of people who weight more than 150 pounds, about 95 of you would have intervals that do contain that true population proportion.
- If each of you and 99 friends were to sample 1200 people in an election campaign, to estimate the true population proportion of people who will vote for candidate X, about 95 of you will have intervals that do contain this population proportion.

4.2.9.4 (Non-)Effect of the Population Size

Note that in both the Davis and election examples, it doesn't matter what the size of the population is. The approximate distribution of $\widehat{p}$, N(p,p(1-p)/n), and thus 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.

4.2.9.5 Planning Ahead

Now, why do the pollsters sample 1200 people?

First, note that the maximum possible value of $\widehat{p}(1-\widehat{p})$ is 0.25.[10] 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.

4.2.10 One-Sided Confidence Intervals

Confidence intervals as discussed so far give one both an upper and lower bound for the parameter of interest. (From here on, the word *parameter* is used in a broader context than just parametric families of distributions. The term will refer to any population quantity.)

In some applications, we are interested in having only an upper bound, or only a lower bound. One can go through the same kind of reasoning as in Section 4.2 above to obtain approximate 95% <u>one-sided</u> confidence intervals:

$$(\overline{W} - 1.65\frac{s}{\sqrt{n}}, \infty) \tag{4.35}$$

$$(-\infty, \overline{W} + 1.65\frac{s}{\sqrt{n}}) \tag{4.36}$$

Note the constant 1.65, which is the 0.95 quantile of the N(0,1) distr, compared to 1.96, the 0.975 quantile.

[10] Use calculus to find the maximum value of f(x) = x(1-x).

One-sided intervals might be used simply out of preferance, or from the dictates of the application. For example, in the **market basket** problems in data mining, one is interested in finding proportions that are larger than a specified value, so one-sided intervals are natural there.

4.2.11 Confidence Intervals for Differences of Means or Proportions

4.2.11.1 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 m_1 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 $\overline{X} - \overline{Y}$.

In order to form a confidence interval for $\mu_1 - \mu_2$ using $\overline{X} - \overline{Y}$, we need to know the distribution of that latter quantity. To see this, recall that this is how we eventually got (4.17); we started by noting the distribution of $\overline{W}$, or more precisely the distribution of $(\overline{W} - \mu)/(\sigma/\sqrt{n})$ in (4.10), and then used that to derive our confidence interval. So, here we need to know the distribution of $\overline{X} - \overline{Y}$.

Note first that $\overline{X}$ and $\overline{Y}$ are independent. They come from separate people. Also, as noted before, they are approximately normally distributed. So, they jointly have an approximately bivariate normal distribution. Then from our earlier unit on multivariate distributions, Theorem 4, we know that the linear combination

$$\overline{X} - \overline{Y} = 1 \cdot \overline{X} + (-1) \cdot \overline{Y} \tag{4.37}$$

will also have an approximately normal distribution, with mean $\mu_1 + (-1)\mu_2$ and variance $\sigma_1^2/n_1 + (-1)^2 \sigma_2^2/n_2$. If we then let s_i^2, i = 1,2 denote the two sample variances, we have that

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.38}$$

has an approximate N(0,1) distribution, and working as before, we have that an approximate 95% confidence interval for $\mu_1 - \mu_2$ is

$$\left(\overline{X} - \overline{Y} - 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \overline{X} - \overline{Y} + 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) \tag{4.39}$$

A similar derivation gives us an approximate 95% confidence interval for the difference in two population proportions $p_1 - p_2$:

$$\left(\widehat{p_1} - \widehat{p_2} - 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \widehat{p_1} - \widehat{p_2} + 1.96\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_2}} \right) \tag{4.40}$$

where

$$s_i^2 = \widehat{p_i}(1 - \widehat{p_i}) \tag{4.41}$$

Example: In a network security application, C. Mano *et al*[11] 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

$$11.520 - 2.000 \pm 1.96\sqrt{\frac{9.939^2}{344} + \frac{6.299^2}{436}} = 9.52 \pm 1.22 \tag{4.42}$$

So you can see that there is a big difference between the two networks, even after allowing for sampling variation.

4.2.11.2 Random Sample Size

In our Davis weights example in Section 4.2.3.1, we were implicitly assuming that the samples sizes of the two groups, n_1 and n_2, were nonrandom. For instance, we might sample 500 men and 500 women.

On the other hand, we might simply sample 1000 people without regard to gender. Then the number of men and women in the sample would be random. Think once again of our notebook view. In our first sample of 1000 people, we might have 492 men and 508 women. In our second sample, the gender breakdown might be 505 and 495, and so on. In keeping with the convention to denote random quantities by capital letters, we might write the numbers of men and women in our sample as N_1 and N_2.

[11] 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.

However, in most cases it should not matter. As long as there is not some odd property of our sampling method, e.g. in which there would be tendency for larger samples to have shorter men, we can simply do our inference conditionally on N_1 and N_2, thus treating them as constants.

4.2.11.3 Dependent Samples

Note carefully, though, that a key point above was the independence of the two samples. By contrast, suppose we wish, for instance, to find a confidence interval for $\nu_1 - \nu_2$, the difference in mean weights in Davis of 15-year-old and 10-year-old children, and suppose our data consist of pairs of weight 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 weight U_i at age 15 and V_i at age 10. Let $\overline{V}$ and $\overline{U}$ 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 (4.39).

However, the random variables $T_i = V_i - U_i$, i = 1,2,...,n are still independent. Thus we can use (4.17), so that our approximate 95% confidence interval is

$$(\overline{T} - 1.96\frac{s}{\sqrt{n}}, \overline{T} + 1.96\frac{s}{\sqrt{n}}) \tag{4.43}$$

where s^2 is the 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. To deal with this, we could set up 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.

That would get a bit confusing. 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 (3.29):

$$Var(\widehat{p}_A - \widehat{p}_B) = Var(\widehat{p}_A) + Var(\widehat{p}_B) - 2Cov(\widehat{p}_A, \widehat{p}_B) \tag{4.44}$$

So, using (3.119), the standard error of $\widehat{p}_A - \widehat{p}_B$ is

$$\sqrt{0.001\widehat{p}_A(1-\widehat{p}_A) + 0.001\widehat{p}_B(1-\widehat{p}_B) + 0.002\widehat{p}_A\widehat{p}_B} \tag{4.45}$$

4.2.12 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, researchers want to predict forest cover type for a given location (there are seven different types), from known geographic data, as direct observation is too expensive and may suffer from land access permission issues. (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.)

There were over 50,000 observations, but for simplicity we'll just use the first 1,000 here.

One of the variables was the amount of hillside shade at noon, which we'll call HS12. Let's find an approximate 95% confidence interval for the difference in population mean HS12 values in cover type 1 and type 2 locations. 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. So our confidence interval is

$$223.8 - 226.3 \pm 1.96\sqrt{\frac{15.3^2}{226} + \frac{14.3^2}{585}} = -2.5 \pm 2.3 = (-4.8, -0.3) \tag{4.46}$$

Now let's find a confidence interval for the difference in population proportions of sites that have cover types 1 and 2. Our sample estimate is

$$\widehat{p}_1 - \widehat{p}_2 = 0.226 - 0.585 = -0.359 \tag{4.47}$$

The standard error of this quantity, from (4.45), is

$$\sqrt{0.001 \cdot 0.226 \cdot 0.7740.001 \cdot 0.585 \cdot 0.415 + 002 \cdot 0.226 \cdot 0.585} = 0.019 \tag{4.48}$$

That gives us a confidence interval of

$$-0.359 \pm 1.96 \cdot 0.019 = (-0.397, -0.321) \tag{4.49}$$

4.2.13 Exact Confidence Intervals

Recall how we derived our previous confidence intervals. We began with a probability statement involving our estimator, and then did some algebra to turn it around into a formula for a confidence interval. Those operations had nothing to do with the approximate nature of the distributions involved. We can do the same thing if we have exact distributions.

For example, suppose we have a random sample $X_1, ..., X_{10}$ from an exponential distribution with parameter λ. Let's find an exact 95% confidence interval for λ.

Let

$$T = X_1 + ... + X_{10} \tag{4.50}$$

Recall that T has a gamma distribution with parameters 10 (the "shape," in R's terminology) and λ. Let $q(\lambda)$ denote the 0.95 quantile of this distribution, i.e. the point to the right of which there is only 5% of the area under the density. Note carefully that this is indeed a function of λ; it has different values for different λ. Then:

$$0.95 = P[T \leq q(\lambda)] = P[q^{-1}(T) \geq \lambda] \tag{4.51}$$

(Here we have used the fact that q() is a decreasing function.)

So, an EXACT 95% one-sided confidence interval for λ is

$$(0, q^{-1}(T)) \tag{4.52}$$

Now, what IS q^{-1}? Recall what q() is, the 0.95 quantile of the gamma distribution with shape 10. It always helps intuition to look at some specific numbers:

```
> qgamma(0.95,10,2.5)
[1] 6.282087
> qgamma(0.95,10,4)
[1] 3.926304
```

So, q(2.5) = 6.28 and q(4) = 3.92. That means $q^{-1}(6.28) = 2.5$ and $q^{-1}(3.92) = 4$.

You can now see how we can form the interval. Say T = 16.4. Then we do some trial-and-error until we find a number w such that **qgamma(0.95,10,w) = 16**. Our confidence interval is then (0,w).

4.3 One More Time: Why Do We Use Confidence Intervals?

After all the variations on a theme in the very long Section 4.2, 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.

4.4 Significance Testing

4.4.1 The Basics

Suppose you have a coin which 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 (4.34). The width of the interval would tell you the margin of error, i.e. it tells you whether 100 tosses was 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.**

Unfortunately, this entire process would be counter to the traditional usage of statistics. Most users of statistics would use the toss data to test the **null hypothesis**

$$H_0 : p = 0.5 \tag{4.53}$$

against the **alternate hypothesis**

$$H_A : p \neq 0.5 \tag{4.54}$$

The approach is to consider H_0 "innocent until proven guilty." We form the **test statistic**

$$Z = \frac{\widehat{p} - 0.5}{\sqrt{\frac{1}{n}\widehat{p}(1-\widehat{p})}} \tag{4.55}$$

Under H_0 the random variable Z would have an approximate N(0,1) distribution. The basic idea is that if Z turns out to have a value which is rare for that distribution, we say, "Rather than believe we've observed a rare event, we choose instead to abandon our assumption that H_0 is true."

So, what do we take for our cutoff value for "rareness"? This probability is called the **significance level**, denoted by α. The classical value for α is 0.05. If H_0 were true, Z would have an approximate N(0,1) distribution, and thus would be less than -1.96 or greater than 1.96 only 5% of the time, a "rare event."

So, if Z does stray that far (i.e. 1.96 or more in either direction) from 0, we reject H_0, and decide that $p \neq 0.5$. We say, "The value of p is significantly different from 0.5"; more on this below, as it is NOT what it sounds like.

Let X be the number of heads we get from our 100 tosses. Note that our rule for decision making formulated above is equivalent (do the algebra to see this for yourself) to saying that we will accept H_0 if $40 \leq X \leq 60$, and reject it otherwise.

4.4.2 General Testing Based on Normally Distributed Estimators

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

$$Z = \frac{\widehat{\theta} - c}{s.e.(\widehat{\theta})} \tag{4.56}$$

where $s.e.(\widehat{\theta})$ is the standard error of $\widehat{\theta}$ (Section 4.2.6), and proceed as before:

Reject $H_0 : \theta = c$ at the significance level of $\alpha = 0.05$ if $|Z| \geq 1.96$.

4.4.3 Example: Network Security

Let's look at the network security example in Section 4.2.11.1 again. Here $\widehat{\theta} = \overline{X} - \overline{Y}$, and c is presumably 0 (depending on the goals of Mano *et al*). If you review the material leading up to (4.38), you'll see that

$$s.e.(\overline{X} - \overline{Y}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{4.57}$$

In that example, we found that the standard error was 0.61. So, our test statistic (4.56) is

$$Z = \frac{\overline{X} - \overline{Y} - 0}{0.61} = \frac{11.52 - 2.00}{0.61} = 15.61 \tag{4.58}$$

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.

4.4.4 The Notion of "p-Values"

In that example above, the Z value, 15.61, was far larger than the cutoff for rejection of H_0, 1.96. You might say that we "resoundingly" rejected H_0. When data analysts encounter such a situation, they want to indicate it in their reports. This is done through something called the **observed significance level**, more often called the **p-value**.

To illustrate this, let's look at a somewhat milder case, in which Z = 2.14. By checking the a table of the N(0,1) distribution, or say by calling **pnorm(2.14)** in R, we would find that the N(0,1) distribution has area 0.016 to the right of 2.14, and of course an equal area to the left of -2.14. In other words, in the general formulation in Section 4.4.2, we would be able to reject H_0 even at the much more stringent significance level of 0.032 instead of 0.05. This would be a stronger statement, and in the research community it is customary to say, "The p-value was 0.032."

In our example above in which Z was 15.61, the value is literally "off the chart"; **pnorm(15.61)** returns a value of 1. Of course, it's a tiny bit less than 1, but it is so far out in the right tail of the N(0,1) distribution that the area to the right is essentially 0. So, this would be treated as very, very highly significant.

If many tests are performed and are summarized in a table, it is customary to denote the ones with 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. Well, that's a common interpretation, but careful analysts know it to be misleading, as we will now discuss.

4.4.5 What's Random and What Is Not

It is crucial to keep in mind that H_0 is not an event or any other kind of random entity. This coin either has p = 0.5 or it doesn't. If we repeat the experiment, we will get a different value of X, but p doesn't change. So for example, it would be wrong and meaningless to speak of the "probability that H_0 is true."

Similarly, it would be wrong and meaningless to write $0.05 = P(|Z| > 1.96|H_0)$, again because H_0 is not an event and this kind of conditional probability would not make sense. What is customarily written is something like

$$0.05 = P_{H_0}(|Z| > 1.96) \tag{4.59}$$

This is read aloud as "the probability that $|Z|$ is larger than 1.96 under H_0," with the phrase *under* H_0 referring to the probability measure in the case in which H_0 is true.

4.4.6 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

$$H_A : p > 0.5 \tag{4.60}$$

A "rare event" which could make us abandon our belief in H_0 would now be if Z in (4.55) is very large in the positive direction. So, with $\alpha = 0.05$, our rule would now be to reject H_0 if $Z > 1.65$.

The same would be the case if our null hypothesis were

$$H_A : p \leq 0.5 \tag{4.61}$$

instead of

$$H_A : p = 0.5 \tag{4.62}$$

Then (4.59) would change to

$$0.05 \geq P_{H_0}(|Z| > 1.65) \tag{4.63}$$

4.4.7 Exact Tests

Remember, the tests we've seen so far are all approximate. In (4.55), for instance, $\widehat{p}$ had an approximate normal distribution, so that the distribution of Z was approximately N(0,1). Thus the significance level α was approximate, as were the p-values and so on.[12]

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.

Let's consider the coin example again. 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?

$$\alpha = \sum_{i=8}^{10} P(X = i) = \sum_{i=8}^{10} \left(\begin{array}{c} 10 \\ i \end{array} \right) \left(\frac{1}{2}\right)^{10} = 0.055 \tag{4.64}$$

That's not 0.05. Clearly we cannot get an exact significance level of 0.05,[13] but our α is exactly 0.055.

Of course, 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.

As another 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

$$H_0 : \mu = 1000 \tag{4.65}$$

vs.

$$H_A : \mu > 1000 \tag{4.66}$$

It is natural to have our test take the form in which we reject H_0 if

$$\overline{X} > w \tag{4.67}$$

[12] 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.

[13] Actually, it could be done by introducing some randomization to our test.

for some constant w chosen so that

$$P(\overline{X} > w) = 0.05 \tag{4.68}$$

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.

4.5 What's Wrong with Significance Testing

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.

We'll see why this is the case shortly, but first a bit of history. When the concept of significance testing 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. Interesting, on page 2, they note the prominent people who have criticized testing. Their list is a virtual "who's who" of statistics, as well as Nobel laureates Richard Feynman (physics) and Kenneth Arrow (economics).
- See `http://www.indiana.edu/~stigtsts/quotsagn.html` for a nice collection of quotes from famous statisticians on this point.
- There is an entire chapter devoted to this issue in one of the best-selling elementary statistics textbooks in the nation.[14]

[14] *Statistics*, third edition, by David Freedman, Robert Pisani, Roger Purves, pub. by W.W. Norton, 1997.

4.5.1 The Basic Fallacy

To begin with, it's absurd to test H_0 in the first place. No coin is absolutely perfectly balanced, with p = 0.5000000000000000000000000000... We know that before even collecting any data.

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 (4.55) as the sample size n grows. if we have a large enough sample, eventually the denominator in (4.55) 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, p is different from 0.5, but NOT in any significant sense.

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 summary, the two 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.

4.5.2 What to Do Instead

In the coin example, we could set limits of fairness, say require that p be no more than 0.01 from 0.5 in order to consider it fair. We could then test the hypothesis

$$H_0 : 0.49 \leq p \leq 0.51 \tag{4.69}$$

Such an approach is almost never used in practice, as it is somewhat difficult to use and explain. But even more importantly, what if the true value of p were, say, 0.51001? Would we still really want to reject the coin in such a scenario?

Note carefully that I am not saying that we should not make a decision. We *do* have to decide, e.g. decide whether a new hypertension drug is safe or in this case decide whether this coin is "fair" enough for practical purposes, say for determining which team gets the kickoff in the Super Bowl. But it should be an informed decision, and even testing the modified H_0 above would be much less informative than a confidence interval.

Forming a confidence interval is the far superior approach. The width of the interval shows us whether n is large enough for $\widehat{p}$ to be reasonably accurate, and the location of the interval tells us whether the coin is fair enough for our purposes.

Note that in making such a decision, we do NOT simply check whether 0.5 is in the interval. That would make the confidence interval reduce to a significance test, which is what we are trying to avoid. If for example the interval is (0.502,0.505), we would probably be quite satisfied that the coin is fair enough for our purposes, even though 0.5 is not in the interval.

Significance testing is also used for model building, such as for predictor variable selection in regression analysis (a method to be covered in a later unit). 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.

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 6.

4.5.3 Decide on the Basis of "the Preponderance of Evidence"

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.

4.5.4 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." 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.**

Another way to describe this is that the preponderance of evidence is that Smith is winning the election. This can be formalized by noting that the point at the center of the interval has the highest likelihood, while the further a point is from the center, the lower its likelihood. (I mean this in the sense of Section 4.6.3, NOT in a Bayesian sense, which I disapprove of.)

4.6 General Methods of Estimation

In the preceding sections, we often referred to certain estimators as being "natural." For example, if we are estimating a population mean, an obvious choice of estimator would be the sample mean. But in many applications, it is less clear what a "natural" estimate for a parameter of interest would be.[15] We will present general methods for estimation in this section.

4.6.1 Example: Guessing the Number of Raffle Tickets Sold

You've just bought a raffle ticket, and find that you have ticket number 68. You check with a couple of friends, and find that their numbers are 46 and 79. Let c be the total number of tickets. How should we estimate c, using our data 68, 46 and 79?

It is reasonable to assume that each of the three of you is equally likely to get assigned any of the numbers 1,2,...,c. In other words, the numbers we get, X_i, i = 1,2,3 are uniformly distributed on the set $\{$1,2,...,c$\}$. We can also assume that they are independent; that's not exactly true, since we are sampling without replacement, but for large c—or better stated, for n/c small—it's close enough.

So, we are assuming that the X_i are independent and identically distributed—famously written as **i.i.d.** in the statistics world—on the set $\{$1,2,...,c$\}$. How do we use the X_i to estimate c?

[15] Recall from Section 4.2.10 that we are now using the term *parameter* to mean any population quantity, rather an an index into a parametric family of distributions.

4.6.2 Method of Moments

One approach, an intuitive one, would be to reason as follows. Note first that

$$E(X) = \frac{c+1}{2} \tag{4.70}$$

Let's solve for c:

$$c = 2EX - 1 \tag{4.71}$$

We know that we can use

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{4.72}$$

to estimate EX, so by (4.71), $2\overline{X} - 1$ is an estimate of c. Thus we take our estimator for c to be

$$\widehat{c} = 2\overline{X} - 1 \tag{4.73}$$

This estimator is called the Method of Moments estimator of c.

Let's step back and review what we did:

- We wrote our parameter as a function of the population mean EX of our data item X. Here, that resulted in (4.71).
- In that function, we substituted our sample mean $\overline{X}$ for EX, and substituted our estimator $\widehat{c}$ for the parameter c, yielding (4.73). We then solved for our estimator.

We say that an estimator $\widehat{\theta}$ of some parameter θ is **consistent** if

$$\lim_{n \to \infty} \widehat{\theta} = \theta \tag{4.74}$$

where n is the sample size. In other words, as the sample size grows, the estimator eventually converges to the true population value.

Of course here $\overline{X}$ is a consistent estimator of EX . Thus you can see from (4.71) and (4.73) that $\widehat{c}$ is a consistent estimator of c. In other words, the Method of Moments generally gives us consistent estimators.

What if we have more than one parameter to estimate, say $\theta_1, ..., \theta_r$? We generalize what we did above. To see how, recall that $E(X^i)$ is called the i^{th} **moment** of X;[16] let's denote it by η_i. Also, note that although we derived (4.71) by solving (4.70) for c, we did start with (4.70). So we do the following:

- For i = 1,...,r we write η_i as a function g_i of all the θ_k.
- For i = 1,...,r set

$$\widehat{\eta_i} = \frac{1}{n} \sum_{j=1}^{n} X_j^i \tag{4.75}$$

- Substitute the $\widehat{\theta_k}$ in the g_i and then solve for them.

In the above example with the raffle, we had r = 1, $\theta_1 = c$, $g_1(c) = (c+1)/2$ and so on. A two-parameter example will be given below.

4.6.3 Method of Maximum Likelihood

Another method, much more commonly used, is called the **Method of Maximum Likelihood**. In our example above, it means asking the question, "What value of c would have made our data—68, 46, 79—most likely to happen?" Well, let's find what is called the **likelihood**, i.e. the probably of our particular data values occurring:

$$L = P(X_1 = 68, X_2 = 46, X_3 = 79) = \begin{cases} (\frac{1}{c})^3, & \text{if } c \geq 79 \\ 0, & \text{otherwise} \end{cases} \tag{4.76}$$

Now keep in mind that c is a fixed, though unknown constant. It is not a random variable. What we are doing here is just asking "What if" questions, e.g. "If c were 85, how likely would our data be? What about c = 91?"

Well then, what value of c maximizes (4.76)? Clearly, it is c = 79. Any smaller value of c gives us a likelihood of 0. And for c larger than 79, the larger c is, the smaller (4.76) is. So, our maximum likelihood estimator (MLE) is 79. In general, if our sample size in this problem were n, our MLE for c would be

$$\check{c} = \max_i X_i \tag{4.77}$$

[16] Hence the name, Method of Moments.

4.6.4 Example: Estimation the Parameters of a Gamma Distribution

As another example, suppose we have a random sample $X_1, ..., X_n$ from a gamma distribution.

$$f_X(t) = \frac{1}{\Gamma(c)} \lambda^c t^{c-1} e^{-\lambda t}, \ t > 0 \tag{4.78}$$

for some unknown c and λ. How do we estimate c and λ from the X_i?

4.6.4.1 Method of Moments

Let's try the Method of Moments, as follows. We have two population parameters to estimate, c and λ, so we need to involve two moments of X. That could be EX and $E(X^2)$, but here it would more conveniently be EX and Var(X). We know from our previous unit on continuous random variables, Chapter 2, that

$$EX = \frac{c}{\lambda} \tag{4.79}$$

$$Var(X) = \frac{c}{\lambda^2} \tag{4.80}$$

In our earlier notation, this would be r = 2, $\theta_1 = c$, $\theta_2 = \lambda$ and $g_1(c, \lambda) = c/\lambda$ and $g_2(c, \lambda) = c/\lambda^2$.

Switching to sample analogs and estimates, we have

$$\frac{\widehat{c}}{\widehat{\lambda}} = \overline{X} \tag{4.81}$$

$$\frac{\widehat{c}}{\widehat{\lambda}^2} = s^2 \tag{4.82}$$

Dividing the two quantities yields

$$\widehat{\lambda} = \frac{\overline{X}}{s^2} \tag{4.83}$$

which then gives

$$\widehat{c} = \frac{\overline{X}^2}{s^2} \tag{4.84}$$

4.6.4.2 MLEs

What about the MLEs of c and λ? Remember, the X_i are continuous random variables, so the likelihood function, i.e. the analog of (4.76), is the product of the density values:

$$L = \Pi_{i=1}^n \frac{1}{\Gamma(c)} \lambda^c {X_i}^{c-1} e^{-\lambda X_i} \tag{4.85}$$

In general, it is usually easier to maximize the log likelihood (and maximizing this is the same as maximizing the original likelihood):

$$l = (c-1) \sum_{i=1}^n \ln(X_i) - \frac{1}{\lambda} \sum_{k=1}^n X_i + nc\ln(\lambda) - n\ln(\Gamma(c)) \tag{4.86}$$

One then takes the partial derivatives of (4.86) with respect to c and λ, and sets the derivatives to zero. The solution values, $\check{c}$ and $\check{\lambda}$, are then the MLEs of c and λ. Unfortunately, these equations do not have closed-form solutions, so they must be solved numerically.

4.6.5 More Examples

Suppose $f_W(t) = ct^{c-1}$ for t in (0,1), with the density being 0 elsewhere, for some unknown $c > 0$. We have a random sample $W_1, ..., W_n$ from this density.

Let's find the Method of Moments estimator.

$$EW = \int_0^1 tct^{c-1}\, dt = \frac{c}{c+1} \tag{4.87}$$

So, set

$$\overline{W} = \frac{\widehat{c}}{\widehat{c}+1} \tag{4.88}$$

yielding

$$\widehat{c} = \frac{\overline{W}}{1-\overline{W}} \tag{4.89}$$

What about the MLE?

$$L = \Pi_{i=1}^{n} c W_i^{c-1} \tag{4.90}$$

so

$$l = n \ln c + (c-1) \sum_{i=1}^{n} \ln W_i \tag{4.91}$$

Then set

$$0 = \frac{n}{\widehat{c}} + \sum_{i=1}^{n} \ln W_i \tag{4.92}$$

and thus

$$\widehat{c} = -\frac{1}{\frac{1}{n}\sum_{i=1}^{n} \ln W_i} \tag{4.93}$$

As in Section 4.6.3, not every MLE can be determined by taking derivatives. Consider a continuous analog of the example in that section, with $f_W(t) = \frac{1}{c}$ on (0,c), 0 elsewhere, for some $c > 0$.

The likelihood is

$$\left(\frac{1}{c}\right)^n \tag{4.94}$$

as long as

$$c \geq \max_i W_i \tag{4.95}$$

and is 0 otherwise. So,

$$\widehat{c} = \max_i W_i \tag{4.96}$$

as before.

Let's find the bias of this estimator.

The bias is $E\widehat{C} - c$. To get $E\widehat{c}$ we need the density of that estimator, which we get as follows:

$$\begin{aligned} P(\widehat{c} \leq t) &= P(\text{all } W_i \leq t) \;\; \text{(definition)} && (4.97) \\ &= \left(\frac{t}{c}\right)^n \;\; \text{(density of } W_i\text{)} && (4.98) \end{aligned}$$

So,

$$f_{\widehat{c}}(t) = \frac{n}{c^n} t^{n-1} \tag{4.99}$$

Integrating against t, we find that

$$E\widehat{C} = \frac{n}{n+1}\, c \tag{4.100}$$

So the bias is c/(n+1), not bad at all.

4.6.6 What About Confidence Intervals?

Usually we are not satisfied with simply forming estimates (called **point estimates**). We also want some indication of how accurate these estimates are, in the form of confidence intervals (**interval estimates**).

In many special cases, finding confidence intervals can be done easily on an *ad hoc* basis. Look, for instance, at the Method of Moments Estimator in Section 4.6.2. Our estimator (4.73) is a linear function of $\overline{X}$, so we easily obtain a confidence interval for c from one for EX.

Another example is (4.93). Taking the limit as $n \to \infty$ the equation shows us (and we could verify) that

$$c = \frac{1}{E[\ln W]} \tag{4.101}$$

Defining $X_i = \ln W_i$ and $\overline{X} = (X_1 + ... + X_n)/$, we can obtain a confidence interval for EX in the usual way. We then see from (4.101) that we can form a confidence interval for c by simply taking the reciprocal of each endpoint of the interval, and swapping the left and right endpoints.

What about in general? For the Method of Moments case, our estimators are functions of the sample moments, and since the latter are formed from sums and thus are asymptotically normal, the delta method can be used to show that our estimators are asymptotically normal and to obtain asymptotic variances for them.

There is a well-developed asymptotic theory for MLEs, which under certain conditions not only shows asymptotic normality with a determined asymptotic variance, but also establishes that MLEs are in a certain sense optimal among all estimators. We will not pursue this here.

4.6.7 Bias Vs. Variance

Consider a general estimator Q of some population value b. Then a common measure of the quality (of course there are many others) of the estimator Q is the **mean squared error** (MSE),

$$E[(Q-b)^2] \tag{4.102}$$

Of course, the smaller the MSE, the better.

One can break (4.102) down into variance and (squared) bias components, as follows:[17]

$$\begin{align}
MSE(Q) &= E[(Q-b)^2] \text{ (definition)} \tag{4.103}\\
&= E[\{(Q-EQ)+(EQ-b)\}^2] \text{ (algebra)} \tag{4.104}\\
&= E[(Q-EQ)^2] + 2E\left[(Q-EQ)(EQ-b)\right] + E[(EQ-b)^2] \text{ (alg., props. of E)} \tag{4.105}\\
&= E[(Q-EQ)^2] + E[(EQ-b)^2] \text{ (factor out constant EQ-b)} \tag{4.106}\\
&= Var(Q) + (EQ-b)^2 \text{ (def. of Var(), fact that EQ-b is const.)} \tag{4.107}\\
&= \text{variance + squared bias} \tag{4.108}
\end{align}$$

In other words, in discussing the accuracy of an estimator—especially in comparing two or more candidates to use for our estimator—the average squared error has two main components, one for variance and one for bias. In building a model, these two components are often at odds with each other; we may be able to find an estimator with smaller bias but more variance, or vice versa.

This point will become central in Chapters 6 and 7.

4.7 Real Populations and Conceptual Populations

In our example in Section 4.2.3.1, we were sampling from a real population. However, in many, probably most applications of statistics, either the population or the sampling is more conceptual.

Consider the experiment comparing three scripting languages in Section 3.168. We think of our programmers as being a random sample from the population of all programmers, but that is probably an idealization.

[17] In reading the following derivation, keep in mind that EQ and b are constants.

It may be, for example, that they all work at the same company, in which case we must think of them as a "random sample" from the rather conceptual "population" of all programmers who *might* work at this company.[18]

And what about our raffle example in Section 4.6.1? Certainly we can imagine various kinds of randomness that contribute to the numbers people get on their raffle tickets. Maybe, for instance, you were in a traffic jam on the way to the the place where you bought the ticket, so you bought it a little later than you might have and thus got a higher number. But I've always emphasized the notion of a repeatable experiment in these notes. How can that happen here? You could imagine, for instance, the raffle chair suddenly losing all the tickets, and asking everyone to draw again, resulting in different ticket numbers. Or you can imagine the "population" of all raffles that you might submit to which have the same value of c.

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.

4.8 Nonparametric Distribution Estimation

Here we will be concerned with estimating distribution functions and densities in settings in which we do not assume our distribution belongs to some parametric model.

4.8.1 The Empirical cdf

Recall that F_X, the cdf of X, is defined as

$$F_X(t) = P(X \leq t), \; -\infty < t < \infty \tag{4.109}$$

Define its sample analog, called the **empirical distribution function**, by

$$\widehat{F}_X(t) = \frac{\# \; of \; X_i \; in \; (-\infty, t)}{n} \tag{4.110}$$

In other words, $F_X(t)$ is the proportion of X that are below t in the population, and $\widehat{F}_X(t)$ is the value of that proportion in our sample. $\widehat{F}_X(t)$ estimates $F_X(t)$ for each t.

Graphically, $\widehat{F}_X$ is a step function, with jumps at the values of the X_i. Specifically, let Y_j, j = 1,...,n denote

[18] 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 7.

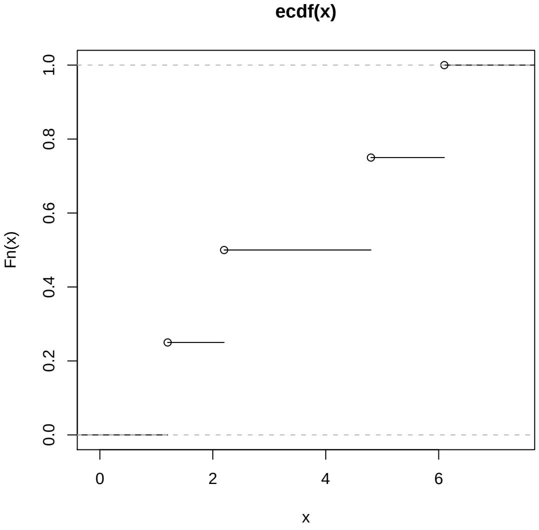
the sorted version of the X_I.[19] Then

$$\widehat{F}_X(t) = \begin{cases} 0, & \text{for } t < Y_1 \\ \frac{j}{n}, & \text{for } Y_j \leq t < Y_{j+1} \\ 1, & \text{for } t > Y_n \end{cases} \tag{4.111}$$

Here is a simple example. Say n = 4 and our data are 4.8, 1.2, 2.2 and 6.1. We can plot the empirical cdf by calling R's **ecdf()** function:

```
> plot(ecdf(x))
```

Here is the graph:



Consider the Bus Paradox example again. Recall that W denoted the time until the next bus arrives. This is called the **forward recurrence time**. The **backward recurrence time** is the time since the last bus was here, which we will denote by R.

Suppose we are interested in estimating the density of R, $f_R()$, based on the sample data $R_1, ..., R_n$ that we gather in our simulation in Section 4.2.1, where n = 1000. How can we do this?[20]

[19] A common notation for this is $Y_j = X_{(j)}$, meaning that Y_j is the j^{th} smallest of the X_i. These are called the **order statistics** of our sample.

[20] Actually, our unit on renewal theory, Chapter 9, proves that R has an exponential distribution. However, here we'll pretend we don't know that.

We could, of course, assume that f_R is a member of some parametric family of distributions, say the two-parameter gamma family. We would then estimate those two parameters as in Section 4.6, and possibly check our assumption using goodness-of-fit procedures, discussed in our unit on modeling, Chapter 6. On the other hand, we may wish to estimate f_R without making any parametric assumptions. In fact, one reason we may wish to do so is to visualize the data in order to search for a suitable parametric model.

If we do not assume any parametric model, we have in essence change our problem from estimating a finite number of parameters to an infinite-parameter problem; the "parameters" are the values of $f_X(t)$ for all the different values of t. Of course, we probably are willing to assume *some* structure on f_R, such as continuity, but then we still would have an infinite-parameter problem.

We call such estimation **nonparametric**, meaning that we don't use a parametric model. However, you can see that it is really infinite-parametric estimation.

Again discussed in our unit on modeling, Chapter 6, the more complex the model, the higher the variance of its estimator. **So, nonparametric estimators will have higher variance than parametric ones.** The nonparametric estimators will also generally have smaller bias, of course.

4.8.2 Basic Ideas

Recall that

$$f_R(t) = \frac{d}{dt} F_R(t) = \frac{d}{dt} P(R \leq t) \tag{4.112}$$

From calculus, that means that

$$\begin{aligned} f_R(t) &\approx \frac{P(R \leq t+h) - P(R \leq t-h)}{2h} && (4.113) \\ &= \frac{P(t-h < R \leq t+h)}{2h} && (4.114) \end{aligned}$$

if h is small. We can then form an estimate $\widehat{f}_R(t)$ by plugging in sample analogs in the right-hand side of (4.113):

$$\begin{aligned} \widehat{f}_R(t) &\approx \frac{\#(t-h, t+h))/n}{2h} && (4.115) \\ &= \frac{\#(t-h, t+h))}{2hn} && (4.116) \end{aligned}$$

where the notation $\#(a, b)$ means the number of R_i in the interval (a,b).

There is an important issue of how to choose the value of h here, but let's postpone that for now. For the moment, let's take

$$h = \frac{\max_i R_i - \min_i R_i}{100} \tag{4.117}$$

i.e. take h to be 0.01 of the range of our data.

At this point, we'd then compute (4.116) at lots of different points t. Although it would seem that theoretically we must compute (4.116) at infinitely many such points, the graph of the function is actually a step function. Imagine t moving to the right, starting at $\min_i R_i$. The interval $(t - h, t + h)$ moves along with it. Whenever the interval moves enough to the right to either pick up a new R_i or lose one that it had had, (4.116) will change value, but not at any other time. So, we only need to evaluate the function at about $2n$ values of t.

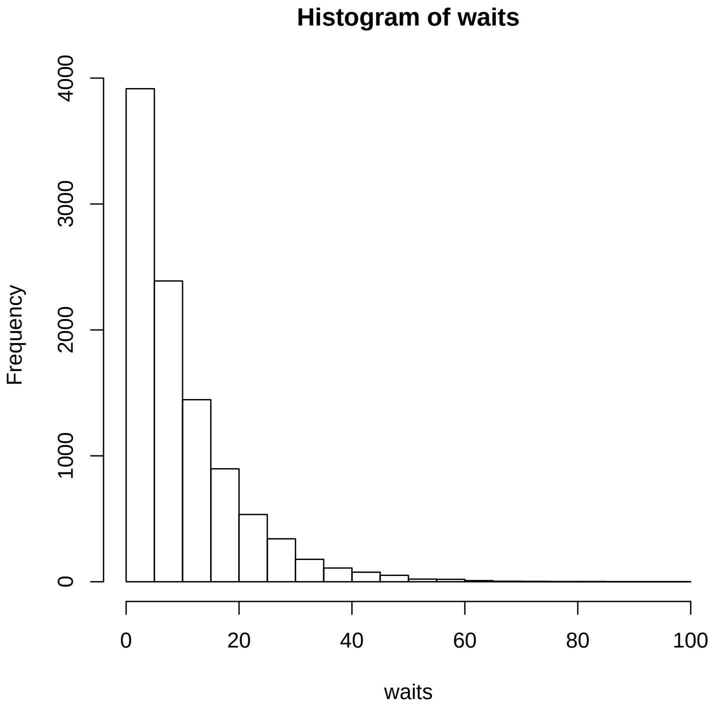
4.8.3 Histograms

If for some reason we really want to save on computation, let's say that we first break the interval $(\min_i R_i, \max_i R_i$ into 100 subintervals of size h given by (4.117). We then compute (4.116) only at the midpoints of those intervals, and pretend that the graph of $\widehat{f}_R(t)$ is constant within each subinterval. Do you know what we get from that? A histogram! Yes, a histogram is a form of density estimation. (Usually a histogram merely displays counts. We do so here too, but we have scaled things so that the total area under the curve is 1.)

Let's see how this works with our Bus Paradox simulation. We'll use R's **hist()** to draw a histogram. First, here's our simulation code:

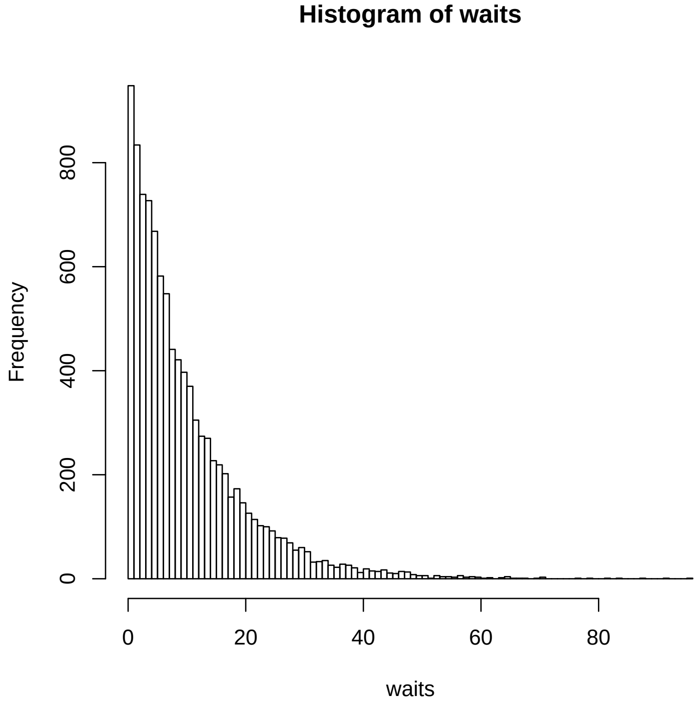
```
doexpt <- function(opt) {
   lastarrival <- 0.0
   while (TRUE) {
      newlastarrival = lastarrival + rexp(1,0.1)
      if (newlastarrival > opt)
         return(opt-lastarrival)
      else lastarrival <- newlastarrival
   }
}

observationpt <- 240
nreps <- 10000
waits <- vector(length=nreps)
for (rep in 1:nreps) waits[rep] <- doexpt(observationpt)
hist(waits)
```

Note that I used the default number of intervals, 20. Here is the result:



The density seems to have a shape like that of the exponential parametric family. (This is not surprising, because it *is* exponential, but remember we're pretending we don't know that.)

Here is the plot with 100 intervals:



Again, a similar shape, though more raggedy.

4.8.4 Kernel-Based Density Estimation

No matter what the interval width is, the histogram will consist of a bunch of rectanges, rather than a curve. That is basically because, for any particular value of t, $\widehat{f_X(t)}$, depends only on the X_i that fall into that interval. We could get a smoother result if we used all our data to estimate $f_X(t)$ but put more weight on the data that is closer to t. One way to do this is called **kernel-based** density estimation, which in R is handled by the function **density()**.

We need a set of weights, more precisely a weight function k, called the **kernel**. Any nonnegative function which integrates to 1—i.e. a density function in its own right—will work. Our estimator is then

$$\widehat{f_R}(t) = \frac{1}{nh} \sum_{i=1}^{n} k \left(\frac{t - R_i}{h} \right) \tag{4.118}$$

To make this idea concrete, take k to be the uniform density on (-1,1), which has the value 0.5 on (-1,1) and 0 elsewhere. Then (4.118) reduces to (4.116). Note how the parameter h, called the **bandwidth**, continues to control how far away from to t we wish to go for data points.

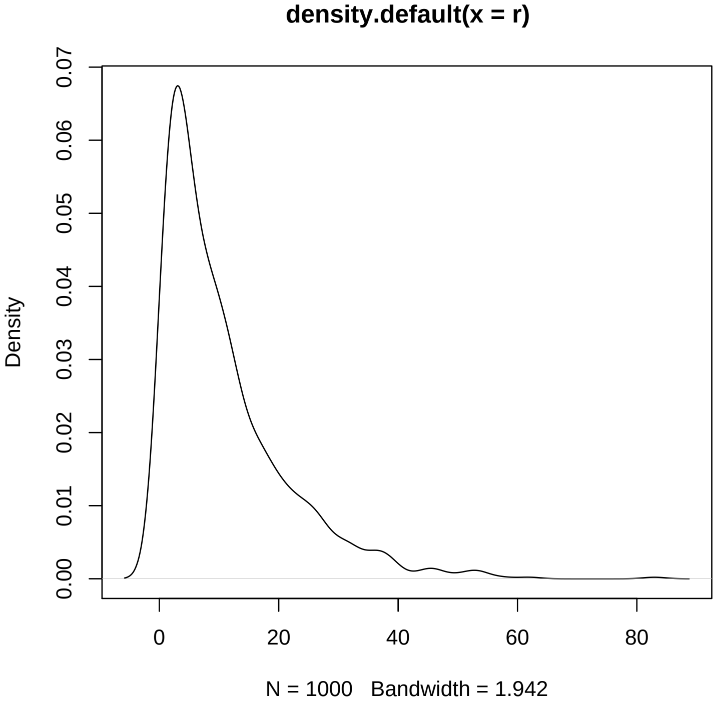


Figure 4.1: Kernel estimate, default bandwidth

But as mentioned, what we really want is to include all data points, so we typically use a kernel with support on all of $(-\infty, \infty)$. In R, the default kernel is that of the N(0,1) density. The bandwidth h controls how much smoothing we do; smaller values of h place heavier weights on data points near t and much lighter weights on the distant points. The default bandwidth in R is taken to the the standard deviation of k.

For our data here, I took the defaults:

```
plot(density(r))
```

The result is seen in Figure 4.1.

I then tried it with a bandwidth of 0.5. See Figure 4.2. This curve oscillates a lot, so an analyst might think 0.5 is too small. (We are prejudiced here, because we know the true population density is exponential.)

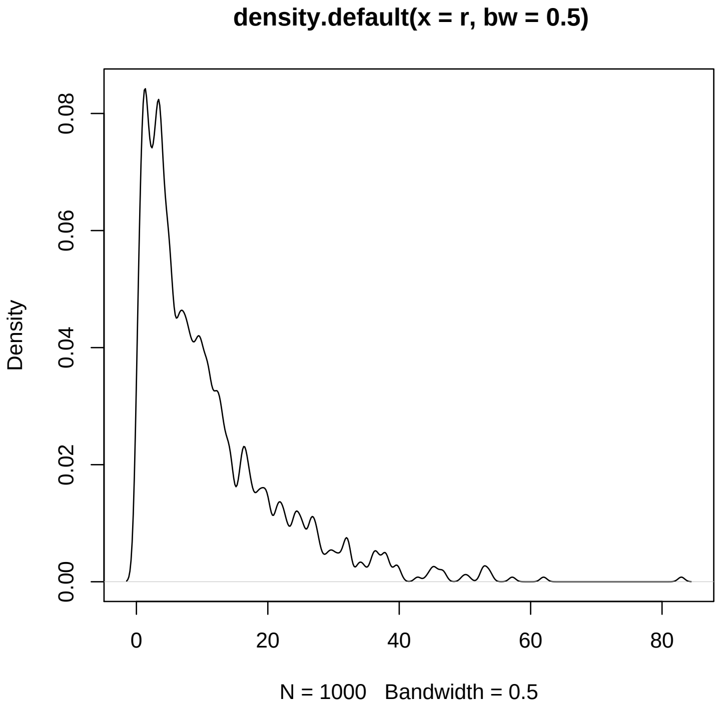


Figure 4.2: Kernel estimate, bandwidth 0.5

4.8.5 Proper Use of Density Estimates

There is no good, practical way to choose a good bin width or bandwdith. Moreover, there is also no good way to form a reasonable confidence band for a density estimate.

So, density estimates should be used as exploratory tools, not as firm bases for decision making. You will probably find it quite unsettling to learn that there is no exact answer to the problem. But that's real life!

Exercises

Note to instructor: See the Preface for a list of sources of real data on which exercises can be assigned to complement the theoretical exercises below.

1. Suppose we 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.

2. 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.

3. In Equation (4.17), 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.

4. Find the Method of Moments and Maximum Likelihood estimators of the following parameters in famous distribution families:

- p in the binomial family (n known)
- p in the geometric family
- μ in the normal family (σ known)
- λ in the Poisson family

5. For each of the following quantities, state whether the given estimator is unbiased in the given context:

(a) (4.15), p. 97, as an estimator of σ^2

(b) $\hat{p}$, as an estimator of p, p.105

(c) $\hat{p}(1-\hat{p})$, as an estimator of p(1-p), p.105

(d) $\bar{X} - \bar{Y}$, as an estimator of $\mu_1 - \mu_2$, p.107

(e) $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_1)^2$ (assuming μ_1 is known), as an estimator of σ_1^2, p.107

(f) $\bar{X}$, as an estimator of μ_1, p.107, *but sampling (from the population of Davis) without replacement*

6. Suppose lifetimes of lightbulbs are exponentially distributed with mean μ. In the past, μ was 1000, but there is a claim that the new light bulbs are improved and $\mu > 1000$. To test that claim, we will sample 100 lightbulbs, getting lifetimes $X_1, ..., X_20$, and compute $\bar{X} = (X_1 + ... + X_{20})/20$. We will then perform a significance test of $H_0 : \mu = 1000$ vs. $H_A : \mu > 1000$. It is natural to have our test take the form in which we reject H_0 if $\bar{X} > r$ for some constant r chosen so that $P(\bar{X} > r) = 0.05$ under H_0.

Suppose we want an exact test, not one based on a normal approximation. Find r.

7. Consider the Method of Moments Estimator $\hat{c}$ in the raffle example, Section 4.6.1. Find the exact value of $Var(\hat{c})$. Use the facts that $1 + 2 + ... + r = r(r+1)/2$ and $1^2 + 2^+ ..., r^2 = r(r+1)(2r+1)/6$.

8. Suppose W has a uniform distribution on (-c,c), and we draw a random sample of size n, $W_1, ..., W_n$. Find the Method of Moments and Maximum Likelihood estimators. (Note that in the Method of Moments case, the first moment won't work.)

9. An urn contains ω marbles, one of which is black and the rest being white. We draw marbles from the urn one at a time, without replacement, until we draw the black one; let N denote the number of draws needed. Find the Method of Moments estimator of ω based on X.

10. Suppose $X_1, ..., X_n$ are uniformly distributed on (0,c). Find the Method of Moments and Maximum Likelihood estimators of c, and compare their mean squared error.

Hint: You will need the density of $M = \max_i X_i$. Derive this by noting that $M \leq t$ if and only if $X_i \leq t$ for all i = 1,2,...,n.

11. Add a single line to the code on page 122 that will print out the estimated value of Var(W).

12. In the raffle example, Section 4.6.1, find a $(1-\alpha)\%$ confidence interval for c based on $\check{c}$, the Maximum Likelihood Estimate of c.

Hint: Use the example in Section 4.2.13 as a guide.

13. In many applications, observations come in correlated clusters. For instance, we may sample r trees at random, then s leaves within each tree. Clearly, leaves from the same tree will be more similar to each other than leaves on different trees.

In this context, suppose we have a random sample $X_1, ..., X_n$, n even, such that there is correlation within pairs. Specifically, suppose the pair (X_{2i+1}, X_{2i+2}) has a bivariate normal distribution with mean (μ, μ) and covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{4.119}$$

i = 0,...,n/2-1, with the n/2 pairs being independent. Find the Method of Moments estimators of μ and ρ.

14. Suppose we have a random sample $X_1, ..., X_n$ from some population in which $EX = \mu$ and $Var(X) = \sigma^2$. Let $\overline{X} = (X_1 + ... + X_n)/n$ be the sample mean. Suppose the data points X_i are collected by a machine, and that due to a defect, the machine always records the last number as 0, i.e. $X_n = 0$. Each of the other X_i is distributed as the population, i.e. each has mean μ and variance σ^2. Find the mean squared error of $\overline{X}$ as an estimator of μ, separating the MSE into variance and squared bias components as in Section 4.6.7.

15. Candidates A, B and C are vying for election. Let p_1, p_2 and p_3 denote the fractions of people planning to vote for them. We poll n people at random, yielding estimates $\widehat{p_1}$, $\widehat{p_2}$ and $\widehat{p_3}$. Y claims that she has more supporters than the other two candidates combined. Give a formula for an approximate 95% confidence interval for $p_2 - (p_1 + p_3)$.

Hint: There is a multinomial distribution involved.

16. Suppose we have a random sample $X_1, ..., X_n$ from a population in which X is uniformly distributed on the region $(0, 1) \cup (2, c)$ for some unknown c $>$ 2. Find closed-form expressions for the Method of Moments and Maximum Likelihood Estimators, to be denoted by T_1 and T_2, respectively.

Chapter 5

Advanced Statistical Inference

5.1 Slutsky's Theorem

(The reader should review Section 2.3.2.7 before continuing.)

Since one generally does not know the value of σ in (4.12), we replace it by s, yielding (4.16). Why was that legitimate?

The answer depends on the theorem below. First, we need a definition.

Definition 5 *We say that a sequence of random variables L_n* **converges in probability** *to the random variable L if for every $\epsilon > 0$,*

$$\lim_{n \to \infty} P(|L_n - L| > \epsilon) = 0 \tag{5.1}$$

This is a little weaker than convergence with probability 1, as in the Strong Law of Large Numbers (SLLN, Section 1.4.7). Convergence with probability 1 implies convergence in probability but not vice versa.

So for example, if $Q_1, Q_2, Q_3, ...$ are i.i.d. with mean ω, then the SLLN implies that

$$L_n = \frac{Q_1 + ... + Q_n}{n} \tag{5.2}$$

converges with probability 1 to ω, and thus L_n converges in probability to ω too.

5.1.1 The Theorem

Theorem 6 Slutsky's Theorem *(abridged version): Consider random variables $X_n, Y_n,$ and X, such that X_n converges in distribution to X and Y_n converges in probability to a constant c with probability 1,*

Then:

(a) $X_n + Y_n$ converges in distribution to $X + c$.

(b) X_n/Y_n converges in distribution to X/c.

5.1.2 Why It's Valid to Substitute s for σ

We now return to the question raised above. In our context here, that we take

$$X_n = \frac{\overline{W} - \mu}{\sigma/\sqrt{n}} \tag{5.3}$$

$$Y_n = \frac{s}{\sigma} \tag{5.4}$$

We know that (5.3) converges in distribution to N(0,1) while (5.4) converges in to 1. Thus for large n, we have that

$$\frac{\overline{W} - \mu}{s/\sqrt{n}} \tag{5.5}$$

has an approximate N(0,1) distribution, so that (4.16) is valid.

5.1.3 Example: Confidence Interval for a Ratio Estimator

Again consider the example in Section 4.2.3.1 of weights of men and women in Davis, but this time suppose we wish to form a confidence interval for the *ratio* of the means,

$$\gamma = \frac{\mu_1}{\mu_2} \tag{5.6}$$

Again, the natural estimator is

$$\widehat{\gamma} = \frac{\overline{X}}{\overline{Y}} \tag{5.7}$$

How can we construct a confidence interval from this estimator? If it were a linear combination of $\overline{X}$ and $\overline{Y}$, we'd have no problem, since a linear combination of multivariate normal random variables is again normal.

That is not exactly the case here, but it's close. Since $\overline{Y}$ converges in probability to μ_2, Slutsky's Theorem (Section 5.1) tells us that the problem here really is one of such a linear combination. We can form a confidence interval for μ_1, then divide both endpoints of the interval by $\overline{Y}$, yielding a confidence interval for γ.

5.2 The Delta Method: Confidence Intervals for General Functions of Means or Proportions

The **delta method** is a great way to derive asymptotic distributions of quantities that are functions of random variables whose asymptotic distributions are already known.

5.2.1 The Theorem

Theorem 7 *Suppose $R_1, ..., R_k$ are estimators of $\eta_1, ..., \eta_k$ based on a random sample of size n. Let R denote the vector whose components are the R_i, and let η denote the corresponding vector for the η_i. Suppose the random vector*

$$\sqrt{n}(R-\eta) = \sqrt{n} \left(\begin{array}{l} R_1 - \eta_1 \\ R_2 - \eta_2 \\ ... \\ R_k - \eta_k \end{array} \right) \tag{5.8}$$

is known to have an asymptotically multivariate normal distribution with mean 0 and nonsingular covariance matrix $\Sigma = (\sigma_{ij})$.

Let h be a smooth scalar function[1] *of k variables, with h_i denoting its i^{th} partial derivative. Consider the*

[1] The word "smooth" here refers to mathematical conditions such as existence of derivatives, which we will not worry about here.

Similarly, the reason that we multiply by $\sqrt{n}$ is also due to theoretical considerations we will not go into here, other than to note that it is related to the formal statement of the Central Limit Theorem in Section 2.3.2.7. If we replace $X_1 + ..., +X_n$ in (2.34), by *random variable*

$$Y = h(R_1, ..., R_k) \tag{5.10}$$

Then $\sqrt{n}[Y - h(\eta_1, ..., \eta_k)]$ *converges in distribution to a normal distribution with mean 0 and variance*

$$[\nu_1, ..., \nu_k]'\Sigma[\nu_1, ..., \nu_k] \tag{5.11}$$

provided not all of

$$\nu_i = h_i(\eta_1, ..., \eta_k), \ i = 1, ..., k \tag{5.12}$$

are 0.

Informally, the theorem says, with R, η, Σ, h() and Y defined above:

> Suppose R is asymptotically multivariate normally distributed with mean η and covariance matrix $\Sigma/n.$ Y will be approximately normal with mean $h(\eta_1, ..., \eta_k)$ and covariance matrix 1/n times (5.11).

Note carefully that the theorem is not saying, for example, that $E[h(R) = h(\eta)$ for fixed, finite n, which is not true. Nor is it saying that h(R) is normally distributed, which is definitely not true; recall for instance that if X has a N(0,1) distribution, then X^2 has a chi-square distribution with one degree of freedom, hardly the same as N(0,1). But the theorem says that for the purpose of asymptotic distributions, we can operate as if these things were true.

The theorem can be used to form confidence intervals for $h(\eta_1, ..., \eta_k)$, because it provides us with a standard error (Section 4.2.6):

$$\text{std. err. of } h(R) = \sqrt{\frac{1}{n} \ [\nu_1, ..., \nu_k]'\Sigma[\nu_1, ..., \nu_k]} \tag{5.13}$$

Of course, these quantities are typically estimated from the sample, e.g.

$$\widehat{\nu}_i = h_i(R_1, ..., R_k) \tag{5.14}$$

$n\overline{X}$, we get

$$Z = \sqrt{n} \cdot \frac{\overline{X} - m}{v} \tag{5.9}$$

So, our approximate 95% confidence interval for $h(\eta_1, ..., \eta_k)$ is

$$h(R_1, ..., R_k) \pm 1.96\sqrt{\frac{1}{n}[\widehat{\nu}_1, ..., \widehat{\nu}_k]'\widehat{\Sigma}[\widehat{\nu}_1, ..., \widehat{\nu}_k]} \tag{5.15}$$

Note that here we are considering scalar functions h(), but the theorem can easily be extended to vector-valued h().

Now, how is theorem derived?

Proof

We'll cover the case k = 1 (dropping the subscript 1 for convenience).

The intuitive version of the proof cites the fact from calculus[2] that a curve is close to its tangent line if we are close to the point of tangency. Here that means

$$h(R) \approx h(\eta) + h'(\eta)(R - \eta) \tag{5.16}$$

if R is near η, which will be the case for large n. Note that in the right-hand side of (5.16), the only random quantity is R; the rest are constants. In other words, the right-hand side has the form c+dQ, where Q is approximately normal. Since a linear function of a normally distributed random variable itself has a normal distribution, (5.16) implies that h(R) is approximately normal with mean $h(\eta)$ and variance $[h'(\eta)]^2 Var(R)$.

Reasoning more carefully, recall the Mean Value Theorem from calculus:

$$h(R) = h(\eta) + h'(W)(R - \eta) \tag{5.17}$$

for some W between η and R. Rewriting this, we have

$$\sqrt{n}[h(R) - h(\eta)] = \sqrt{n}\, h'(W)(R - \eta) \tag{5.18}$$

It can be shown—and should be intuitively plausible to you—that if a sequence of random variables converges in distribution to a constant, the convergence is in probability too. So, $R - \eta$ converges in probability to 0, forcing W to converge in probability to $h(\eta)$. Then from Slutsky's Theorem, the asymptotic distribution of (5.18) is the same as that of $\sqrt{n}\, h'(\eta)(R - \eta)$. The result follows.

■

[2] This is where the "delta" in the name of the method comes from, an allusion to the fact that derivatives are limits of difference quotients.

5.2.2 Example: Square Root Transformation

Here is an example of the delta method with k = 1. It will be a rather odd example, in that our goal is actually not to form a confidence interval for anything, but it will illustrate how the delta method is used.

It is used to be common, and to some degree is still common today, for statistical analysts to apply a square-root transformation to Poisson data. The delta method sheds light on the motivation for this, as follows.

First, note that we cannot even apply the delta method unless we have approximately normally distributed inputs, i.e. the R_i in the theorem. But actually, any Poisson-distributed random variable T is approximately normally distributed if its mean, λ, is large. To see this, recall from Section 3.8.1.2 that sums of independent Poisson random variables are themselves Poisson distributed. So, if for instance, ET is an integer k, then T has the same distribution as

$$U_1 + ... + U_m \tag{5.19}$$

where the U_i are i.i.d. Poisson random variables each having mean 1. By the Central Limit Theorem, T then has an approximate normal distribution, with mean and variance λ. (This is not quite a rigorous argument, so our treatment here is informal.)

Now that we know that T is approximately normal, we can apply the delta method. So, what h() should we use? The pioneers of statistics chose $h(t) = \sqrt{t}$. Let's see why.

Set $Y = h(T) = \sqrt{T}$ (so that T is playing the role of R in the theorem). Here η is $ET = \lambda$.

We have $h'(t) = 1/(2\sqrt{t})$. Then the delta method says that since T is approximately normally distributed with mean λ and variance λ, Y too has an approximate normal distribution, with mean

$$h(\eta) = \sqrt{\lambda} \tag{5.20}$$

What about the variance? Well, in one dimension, (5.11) reduces to

$$\nu^2 Var(R) \tag{5.21}$$

so we have

$$[h'(\eta)]^2 Var(R) = \left(\frac{1}{2\sqrt{t}} \Big|_{t=\lambda} \right)^2 \cdot \lambda = \frac{1}{4\lambda}\lambda = \frac{1}{4} \tag{5.22}$$

So, the (asymptotic) variance of $\sqrt{T}$ is a constant, independent of λ, and we say that the square root function is a **variance stabilizing transformation.** This becomes relevant in regression analysis, where, as we will discuss in Chapter 7, a classical assumption is that a certain collection of random variables all have the same variance. If those random variables are Poisson-distributed, then their square roots will all have approximately the same variance.

5.2.3 Example: Confidence Interval for σ^2

Recall that in Section 4.2.7 we noted that (4.17) is only an approximate confidence interval for the mean. An exact interval is available using the Student t-distribution, if the population is normally distributed. We pointed out that (4.17) is very close to the exact interval for even moderately large n anyway, and since no population is exactly normal, (4.17) is good enough. Note that one of the implications of this and the fact that (4.17) did not assume any particular population distribution is that a Student-t based confidence interval works well even for non-normal populations. We say that the Student-t interval is **robust** to the normality assumption.

But what about a confidence interval for a variance? It can be shown that one can form an exact interval based on the chi-square distribution, if the population is normal. In this case, though, the interval does NOT work well for non-normal populations; it is NOT robust to the normality assumption. So, let's derive an interval that doesn't assume normality; we'll use the delta method. (Warning: This will be a lengthy derivation, but it will cause you to review many concepts, which is good.)

As before, say we have $W_1, ..., W_n$, a random sample from our population, and with W representing a random variable having the population distribution.) Write

$$\sigma^2 = E(W^2) - (EW)^2 \tag{5.23}$$

and from (4.15) write our estimator of σ^2 as

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} W_i^2 - \overline{W}^2 \tag{5.24}$$

This suggests how we can use the delta method. We define

$$R_1 = \overline{W} \tag{5.25}$$

$$R_2 = \frac{1}{n}\sum_{i=1}^{n} W_i^2 \tag{5.26}$$

R_1 is an estimator of EW, and R_2 estimates $E(W^2)$. Furthermore, we'll see below that R_1 and R_2 are approximately bivariate normal, by the multivariate Central Limit Theorem, so we can use the delta method.

And most importantly, our estimator of interest, s^2, is a function of R_1 and R_2:

$$s^2 = R_2 - R_1^2 \tag{5.27}$$

So, we take our function h to be

$$h(u,v) = -u^2 + v \tag{5.28}$$

Now we must find Σ in the theorem. That means we'll need we'll need the covariance matrix of R_1 and R_2. But since

$$\left(\begin{array}{c} R_1 \\ R_2 \end{array} \right) = \frac{1}{n} \sum_{i=1}^{n} \left(\begin{array}{c} W_i \\ W_i^2 \end{array} \right) \tag{5.29}$$

we can derive the covariance matrix of R_1 and R_2, as follows.

Remember, the covariance matrix is the multidimensional analog of variance. So, after reviewing the reasoning in (4.9), we have in the vector-valued version of that derivation that

$$\begin{aligned} Cov \left[\left(\begin{array}{c} R_1 \\ R_2 \end{array} \right) \right] &= \frac{1}{n^2} Cov \left[\sum_{i=1}^{n} \left(\begin{array}{c} W_i \\ W_i^2 \end{array} \right) \right] && (5.30) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} Cov \left[\left(\begin{array}{c} W_i \\ W_i^2 \end{array} \right) \right] && (5.31) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} Cov \left[\left(\begin{array}{c} W \\ W^2 \end{array} \right) \right] && (5.32) \\ &= \frac{1}{n} Cov \left[\left(\begin{array}{c} W \\ W^2 \end{array} \right) \right] && (5.33) \end{aligned}$$

So

$$\Sigma = Cov \left[\left(\begin{array}{c} W \\ W^2 \end{array} \right) \right] \tag{5.34}$$

Now we must estimate Σ. Taking sample analogs of (3.82), we set

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \left(\begin{array}{c} W_i \\ W_i^2 \end{array} \right) (W_i, W_i^2) - R\,R' = \frac{1}{n} \sum_{i=1}^{n} \left(\begin{array}{cc} W_i^2 & W_i^3 \\ W_i^3 & W_i^4 \end{array} \right) - R\,R' \tag{5.35}$$

where $R = (R_1, R_2)'$.

Also, $h'(u, v) = (-2u, 1)'$, so

$$h'(R_1, R_2) = (-2R_1, 1)' \tag{5.36}$$

Whew! We're done. We can now plug everything into (5.15).

Note that all these quantities are expressions in $E(W^k)$ for various k. It should be noted that estimating means of higher powers of a random variable requires larger samples in order to achieve comparable accuracy. Our confidence interval here may need a rather large sample to be accurate, as opposed to the situation with (4.17), in which even n = 20 should work well.

5.2.4 Example: Confidence Interval for a Measurement of Prediction Ability

Suppose we have a random sample $X_1, ..., X_n$ from some population. In other words, the X_i are independent and each is distributed as in the population. Let X represent a generic random variable having that distribution. Here we are allowing the X_i and X to be random vectors, though they won't play much explicit role anyway.

Let A and B be events associated with X. If for example X is a random vector (U,V), we might have A and B being the events U $>$ 12 and U-V $<$ 5. The question of interest here will be to what extent we can predict A from B.

One measure of that might be the quantity $\nu = P(A|B) - P(A)$. The larger ν is (in absolute value), the stronger the ability of B to predict A. (We could look at variations of this, such as the quotient of those two probabilities, but will not do so here.)

Let's use the delta method to derive an approximate 95% confidence interval for ν. To that end, think of four categories—A and B; A and not B; not A and B; and not A and not B. Each X_i falls into one of those categories, so the four-component vector Y consisting of counts of the numbers of X_i falling into the four categories has a multinomial distribution with r = 4.

To use the theorem, set R = Y/n, so that R is the vector of the sample proportions. For instance, R_1 will be the number of X_i satisfying both events A and B, divided by n. The vector η will then be the corresponding population proportion, so that for instance

$$\eta_2 = P(A \text{ and not } B) \tag{5.37}$$

We are interested in

$$
\begin{aligned}
\nu &= P(A|B) - P(A) && (5.38)\\
&= \frac{P(A \text{ and } B)}{P(A \text{ and } B) + P(\text{not } A \text{ and } B)} - [P(A \text{ and } B) + P(A \text{ and not } B)] && (5.39)\\
&= \frac{\eta_1}{\eta_1 + \eta_3} - (\eta_1 + \eta_2) && (5.40)
\end{aligned}
$$

By the way, since η_4 is not involved, let's shorten R to $(R_1, R_2, R_3)'$.

What about Σ? Since Y is multinomial, Equation (3.119) provides us Σ:

$$
\Sigma = \frac{1}{n} \begin{pmatrix} \eta_1(1-\eta_1) & -\eta_1\eta_2 & -\eta_1\eta_3 \\ -\eta_2\eta_1 & \eta_2(1-\eta_2) & -\eta_2\eta_3 \\ -\eta_3\eta_1 & -\eta_3\eta_2 & \eta_3(1-\eta_3) \end{pmatrix} \qquad (5.41)
$$

We then get $\widehat{\Sigma}$ by substituting R_i for η_i. After deriving the $\widehat{\nu}_i$ from (5.38), we make the same substitution there, and then compute (5.15).

5.3 Simultaneous Confidence Intervals

Suppose in our study of heights, weights and so on of people in Davis, we are interested in estimating a number of different quantities, with our forming a confidence interval for each one. Though our confidence level for each one of them will be 95%, our *overall* confidence level will be less than that. In other words, we cannot say we are 95% confident that all the intervals contain their respective population values.

In some cases we may wish to construct confidence intervals in such a way that we can say we are 95% confident that all the intervals are correct. This branch of statistics is known as **simultaneous inference** or **multiple inference**.

Usually this kind of methodology is used in the comparison of several **treatments**. This term originated in the life sciences, e.g. comparing the effectiveness of several different medications for controlling hypertension, it can be applied in any context. For instance, we might be interested in comparing how well programmers do in several different programming languages, say Python, Ruby and Perl. We'd form three groups of programmers, one for each language, with say 20 programmers per group. Then we would have them write code for a given application. Our measurement could be the length of time T that it takes for them to develop the program to the point at which it runs correctly on a suite of test cases.

Let T_{ij} be the value of T for the jth programmer in the ith group, i = 1,2,3, j = 1,2,...,20. We would then wish to compare the three "treatments," i.e. programming languages, by estimating $\mu_i = ET_{i1}$, i = 1,2,3. Our estimators would be $U_i = \sum_{j=1}^{20} T_{ij}/20$, i = 1,2,3. Since we are comparing the three population means, we may not be satisfied with simply forming ordinary 95% confidence intervals for each mean. We may wish to form confidence intervals which *jointly* have confidence level 95%.[3]

Note very, very carefully what this means. As usual, think of our notebook idea. Each line of the notebook would contain the 60 observations; different lines would involve different sets of 60 people. So, there would be 60 columns for the raw data, three columns for the U_i. We would also have six more columns for the confidence intervals (lower and upper bounds) for the μ_i. Finally, imagine three more columns, one for each confidence interval, with the entry for each being either Right or Wrong. A confidence interval is labeled Right if it really does contain its target population value, and otherwise is labeled Wrong.

Now, if we construct individual 95% confidence intervals, that means that in a given Right/Wrong column, in the long run 95% of the entries will say Right. But for simultaneous intervals, we hope that within a line we see three Rights, and 95% of all lines will have that property.

In our context here, if we set up our three intervals to have individual confidence levels of 95%, their simultaneous level will be $0.95^3 = 0.86$, since the three confidence intervals are independent. Conversely, if we want a simultaneous level of 0.95, we could take each one at a 98.3% level, since $0.95^{\frac{1}{3}} \approx 0.983$.

However, in general the intervals we wish to form will not be independent, so the above "cube root method" would not work. Here we will give a short introduction to more general procedures.

Note that "nothing in life is free." If we want simultaneous confidence intervals, they will be wider.

Another reason to form simultaneous confidence intervals is that it gives you "license to browse," i.e. to rummage through the data looking for interesting nuggets.

5.3.1 The Bonferonni Method

One simple approach is **Bonferonni's Inequality**:

Lemma 8 *Suppose $A_1, ..., A_g$ are events. Then*

$$P(A_1 \text{ or ... or} A_g) \leq \sum_{i=1}^{g} P(A_i) \tag{5.42}$$

You can easily see this for g = 2:

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2) - P(A_1 \text{ and } A_2) \leq P(A_1) + P(A_2) \tag{5.43}$$

One can then prove the general case by mathematical induction.

[3] The word *may* is important here. It really is a matter of philosophy as to whether one uses simultaneous inference procedures.

Now to apply this to forming simultaneous confidence intervals, take A_i to be the event that the i^{th} confidence interval is incorrect, i.e. fails to include the population quantity being estimated. Then (5.42) says that if, say, we form two confidence intervals, each having individual confidence level (100-5/2)%, i.e. 97.5%, then the overall collective confidence level for those two intervals is at least 95%. Here's why: Let A_1 be the event that the first interval is wrong, and A_2 is the corresponding event for the second interval. Then

$$\begin{aligned} \text{overall conf. level} &= P(\text{not } A_1 \text{ and not } A_2) && (5.44) \\ &= 1 - P(A_1 \text{ or } A_2) && (5.45) \\ &\geq 1 - P(A_1) - P(A_2) && (5.46) \\ &= 1 - 0.025 - 0.025 && (5.47) \\ &= 0.95 && (5.48) \end{aligned}$$

5.3.2 Scheffe's Method

The Bonferonni method is unsuitable for more than a few intervals; each one would have to have such a high individual confidence level that the intervals would be very wide. Many alternatives exist, a famous one being **Scheffe's method**.[4]

Theorem 9 *Suppose $R_1, ..., R_k$ have an approximately multivariate normal distribution, with mean vector $\mu = (\mu_i)$ and covariance matrix $\Sigma = (\sigma_{ij})$. Let $\widehat{\Sigma}$ be a* **consistent** *estimator of Σ, meaning that it converges in probability to Σ as the sample size goes to infinity.*

For any constants $c_1, ..., c_k$, consider linear combinations of the R_i,

$$\sum_{i=1}^{k} c_i R_i \tag{5.49}$$

which estimate

$$\sum_{i=1}^{k} c_i \mu_i \tag{5.50}$$

[4] The name is pronounced "sheh-FAY."

Form the confidence intervals

$$\sum_{i=1}^{k} c_i R_i \pm \sqrt{k \chi^2_{\alpha;k}} s(c_1, ..., c_k) \tag{5.51}$$

where

$$[s(c_1, ..., c_k)]^2 = (c_1, ..., c_k)^T \widehat{\Sigma} (c_1, ..., c_k) \tag{5.52}$$

and where $\chi^2_{\alpha;k}$ is the upper-α percentile of a chi-square distribution with k degrees of freedom.[5]

Then all of these intervals (for infinitely many values of the c_i!) have simultaneous confidence level $1 - \alpha$.

By the way, if we are interested in only constructing confidence intervals for **contrasts**, i.e. c_i having the property that $\Sigma_i c_i = 0$, we the number of degrees of freedom reduces to k-1, thus producing narrower intervals.

Just as in Section 4.2.7 we avoided the t-distribution, here we have avoided the F distribution, which is used instead of ch-square in the "exact" form of Scheffe's method.

5.3.3 Example

For example, again consider the Davis heights example in Section 4.2.11. Suppose we want to find approximate 95% confidence intervals for two population quantities, μ_1 and μ_2. These correspond to values of c_1, c_2 of (1,0) and (0,1). Since the two samples are independent, $\sigma_{12} = 0$. The chi-square value is 5.99,[6] so the square root in (5.51) is 3.46. So, we would compute (4.17) for $\overline{X}$ and then for $\overline{Y}$, but would use 3.46 instead of 1.96.

This actually is not as good as Bonferonni in this case. For Bonferonni, we would find two 97.5% confidence intervals, which would use 2.24 instead of 1.96.

Scheffe's method is too conservative if we just are forming a small number of intervals, but it is great if we form a lot of them. Moreover, it is very general, usable whenever we have a set of approximately normal estimators.

[5] Recall that the distribution of the sum of squares of g independent N(0,1) random variables is called **chi-square with g degrees of freedom**. It is tabulated in the R statistical package's function **qchisq()**.

[6] Obtained from R via **qchisq(0.95,2)**.

5.3.4 Other Methods for Simultaneous Inference

There are many other methods for simultaneous inference. It should be noted, though, that many of them are limited in scope, in contrast to Scheffe's method, which is usable whenever one has multivariate normal estimators, and Bonferonni's method, which is universally usable.

5.4 The Bootstrap Method for Forming Confidence Intervals

Many statistical applications can be quite complex, which makes them very difficult to analyze mathematically. Fortunately, there is a fairly general method for finding confidence intervals called the **bootstrap**. Here is a brief overview of the type of bootstrap confidence interval construction called **Efron's percentile method**.

5.4.1 Basic Methodology

Say we are estimating some population value θ based on i.i.d. random variables Q_i, i = 1,...,n. Note that θ and the Q_i could be vector-valued.

Our estimator of θ is of course some function, which we'll call h(), of the Q_i. For example, if we are estimating a population mean by a sample mean, then

$$h(u_1, ..., u_n) = \frac{u_1 + ..., + u_n}{n} \tag{5.53}$$

Our procedure is as follows:

- Estimate θ based on the original sample, i.e. set

$$\widehat{\theta} = h(Q_1, ..., Q_n) \tag{5.54}$$

- For j = 1,2,...,k:
  - Create a new "sample," $\widetilde{Q}_1, .., \widetilde{Q}_n$, by drawing n times with replacement from $Q_1, .., Q_n$.
  - Calculate the value of $\widehat{\theta}$ based on the $\widetilde{Q}_i$ instead of the Q_i, i.e. set

$$\widetilde{\theta}_j = h(\widetilde{Q}_1, ..., \widetilde{Q}_n) \tag{5.55}$$

- Sort the values $\widetilde{\theta}_j$, j = 1,...,k, and let $\widetilde{\theta}_{(k)}$ be the kth-smallest value.[7]

  A and B denote the 0.025 and 0.975 quantiles, i.e.

$$A = \widehat{\theta}_{(0.025n)}, \; B = \widehat{\theta}_{(0.975n)} \tag{5.56}$$

  (The quantities 0.025n and 0.975n must be rounded, say to the nearest integer in the range 1,...,n.) Then your approximate 95% confidence interval for θ is (A,B).

In essence, we are performing a simulation, drawing samples from the empirical distribution function for our data.

5.4.2 Example: Confidence Intervals for a Population Variance

As noted in Section 5.2.3, the classical chi-square method for finding a confidence interval for a population variance σ^2 is not robust to the assumption of a normally distributed parent population. In that section, we showed how to find the desired confidence interval using the delta method.

That was a solution, but the derivation was complex. An alternative would be to use the bootstrap. We resample many times, calculate the sample variance on each of the new samples, and then form a confidence interval for σ^2 as in (5.56). We show the details using R in Section 5.4.3

5.4.3 Computation in R

R includes the **boot()** function to do the mechanics of this for us. To illustrate its usage, let's consider finding a confidence interval for the population variance σ^2, based on the sample variance, s^2. Here is the code:

```
library(boot)  # R base doesn't include the boot package

# finds the sample variance on x[c(inds)]
s2 <- function(x,inds) {
   return(var(x[inds]))
}

bt <- boot(x,s2,R=200)
cilow[rep] <- quantile(bt$t,alp)
cihi[rep] <- quantile(bt$t,1-alp)

print(mean(cilow <= 1.0 & 1.0 <= cihi))
```

How does this work? The line

[7] This is called the kth order statistic among the $\widetilde{\theta}_j$. The parenthetic notation we've used here is standard.

```
bt <- boot(x,s2,R=200)
```

instructs R to apply the bootstrap to the data set **x**, with the statistic of interest being specified by the user in the function **s2()**. The argument **R** here is what we called k in Section 5.4.1 above, i.e. the number of times we resample n items from **x**.

Our argument **inds** in **s2()** is less obvious. Here's what happens: As noted, the **boot()** function mere shortens our work. Without it, we could simply call **sample()** to do our resampling. Say for simplicity that n is 4. We might make the

```
j <- sample(1:4,replace=T)
```

and **j** might turn out to be, say, c(4,1,3,3). We would then apply the statistic to be bootstrapped, in our case here the sample variance, to the data $x[4], x[1], x[3], x[3]$—more compactly and efficiently expressed as $x[c(4, 1, 3, 3)]$. That's what **boot()** does for us. So, in our example above, the argument **inds** would be c(4,1,3,3) here.

In the example here, our statistic to be bootstrapped was a very common one, and thus there was already an R function for it, **var()**. In more complex settings, we'd write our own function. In more complex settings, we'd write our own function

5.4.4 General Applicability

Much theoretical work has been done on the bootstrap, and it is amazingly general. It has become the statistician's "Swiss army knife." However, there are certain types of estimators on which the bootstrap fails. How can one tell in general?

One approach would be to consult the excellent book *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley, Cambridge University Press, 1997.

But a simpler method would be to test the bootstrap in the proposed setting by simulation: Write R code to generate many samples; get a bootstrap confidence interval on each one; and then see whether the number of intervals containing the true population value is approximately 95%.

In the sample variance example above, the code could be:

```
sim <- function(n,nreps,alp) {
   cilow <- vector(length=nreps)
   cihi <- vector(length=nreps)
   for (rep in 1:nreps) {
      x <- rnorm(n)
      bt <- boot(x,s2,R=200)
      cilow[rep] <- quantile(bt$t,alp)
      cihi[rep] <- quantile(bt$t,1-alp)
   }
   print(mean(cilow <= 1.0 & 1.0 <= cihi))
}
```

5.4.5 Why It Works

The mathematical theory of the bootstrap can get extremely involved, but we can at least get a glimpse of why it works here.

Remember, to get any confidence interval from an estimator, we need the distribution of that estimator. In Section 4.2.2.2, for instance, we noted that the sample mean $\overline{W}$ has an approximately normal distribution with mean μ and variance σ^2. Here in our bootstrap context, our goal is to find the approximate distribution of $\widehat{\theta}$.

The bootstrap achieves that goal very simply. Since it provides us with all these "incarnations" of $\widehat{\theta}$, i.e. $\widetilde{\theta}_1, ..., \widetilde{\theta}_k$, we can in principle estimate any apsect of the distribution of $\widehat{\theta}$ that we wish.

For example, suppose we are in a situation covered by the by the delta method, so that $\widehat{\theta}$ is approximately normally distributed. Instead of calculating messy derivatives, we could get a standard error for $\widehat{\theta}$ as the sample standard deviation of $\widetilde{\theta}_1, ..., \widetilde{\theta}_k$.

Efron's percentile method is more general, though unfortunately not so intuitive.

5.5 Bayesian Methods

Whiskey's for drinkin' and water's for fightin'—Mark Twain

The most controversial topic in statistics by far is that of **Bayesian** methods. In fact, it is so controversial that a strident Bayesian colleague of mine even took issue with my calling it "controversial"!

The name stems from Bayes' Rule (Section 1.2.4). No one questions the validity of Bayes' Rule, but the debate stems from the cases in which Bayesians replace some of the probabilities in the theorem with NON-probabilities involving what they call **subjective prior distributions**.

The key word here is *subjective*. Here the analyst, before even collecting data, says, "Well, I think the population mean could be 1.2, with probability, oh, let's say 0.28,..." etc. The analyst then formally incorporates these "feelings" into his/her mathematical procedures for the analysis. Two different analysts working with the same data could come up with very different numerical results, due to having different "feelings."

The Bayesians justify this by saying one should use all available information, even if it is just a hunch. "The analyst is typically an expert in the field under study. You wouldn't want to throw away his/her expertise, would you?" The frequentists, on the other hand, dismiss this as unscientific and lacking in impartiality.

“In research on a controversial health issue, say, you wouldn’t want the researcher to incorporate his/her personal political biases into the number crunching, would you?”

5.5.1 How It Works

To introduce the idea, consider again the example of estimating p, the probability of heads for a certain penny. Suppose we were to say—before tossing the penny even once—“I think p could be any number, but more likely near 0.5, something like a normal distribution with mean 0.5 and standard deviation, oh, let’s say 0.1.” The prior distribution is then $N(0.5, 0.1^2)$. But again, note that the Bayesians do not consider it to be a distribution in the sense of probability. We are just using our “gut feeling” here, our “hunch.” This is an absolutely central point.

Though the Bayesians don’t view it this way at all, mathematically use of the above prior is equivalent to assuming that each penny has a different value of p—i.e. the penny-making machines vary subtantially from one penny to another, say due to vibrations that make some pennies weighted more heavily towards heads and others being more tails-biased—and that if we were to plot those values of p in a histogram, the plot would look like the normal density with mean 0.5 and standard deviation 0.1.

Note again the phrase *equivalent to* in the last paragraph. Bayesians would not regard p as random. They are simply using the normal “distribution” for p to describe a degree of belief, rather than a probability distribution. (I will continue to use quotation marks below for this reason.)

Under this “random p” assumption, the MLE would change. Our data here is X, the number of heads we get from n tosses of the penny. In contrast to the **frequentist** approach, in which the likelihood would be

$$L = \left(\begin{array}{c} n \\ X \end{array} \right) p^X (1-p)^{n-X} \tag{5.57}$$

it now becomes

$$L = \frac{1}{\sqrt{2\pi}\,0.1} \, \exp -0.5[(p-0.5)/0.1]^2 \left(\begin{array}{c} n \\ X \end{array} \right) p^X (1-p)^{n-X} \tag{5.58}$$

We would then find the value of p which maximizes L, and take that as our estimate.

Note how this procedure achieves a kind of balance between what our hunch says and what our data say. In (5.58), suppose the mean of p is 0.5 but n = 20 and X = 12 Then the frequentist estimator would be X/n = 0.6, while the Bayes estimator would be about 0.56. (Computation not shown here.) So our Bayesian approach “pulled” our estimate away from the frequentist estimate, toward our hunch that p is at or very near 0.5. This pulling effect would be stronger for smaller n or a smaller standard deviation of the prior distribution.

More generally, a Bayesian would use Bayes' Rule to compute the "distribution" of p given X, called the **posterior distribution**. The MLE would then be the **mode**, i.e. the point of maximal density of the posterior distribution, but we could use any measure of central tendency; it is common to use the mean.

5.5.2 Noninformative Priors

In choosing a subjective prior distribution in the penny problem, another approach is that of a **noninformative prior** distribution. In our penny example here, we might simply throw up our hands, and say "We have no idea what the value of p is for this coin, so let's just set the prior to be U(0,1)." Needless to say, this is even more controversial. For those who use it, though, it is viewed as achieving other goals, such as avoiding model overfit, a problem discussed (in a non-Bayesian context) in Section 7.3.9.1.

5.5.2.1 Empirical Bayes Methods

Note carefully that if the prior distribution in our model is not subjective, but is a real distribution verifiable from data, the above analysis on p would not be controversial at all. Say p does vary a substantial amount from one penny to another, so that there is a physical distribution involved. Suppose we have a sample of many pennies, tossing each one n times. If n is very large, we'll get a pretty accurate estimate of the value of p for each coin, and we can then plot these values in a histogram and compare it to the $N(0.5, 0.1^2)$ density, to check whether our prior is reasonable. This is called an **empirical Bayes** model, because we can empirically estimate our prior distribution, and check its validity. In spite of the name, frequentists would not consider this to be "Bayesian" analysis.

Note that we could also assume that p has a general $N(\mu, \sigma^2)$ distribution, and estimate μ and σ from the data. We could deal with the situation in which n is only moderately large, as long as it's large enough for the Central Limit Theorem to work well, but we will not pursue that point here.

5.5.3 Extent of Usage of Subjective Priors

Though some academics are staunch, often militantly proselytizing Bayesians, only a small minority of statisticians in practice use the Bayesian approach. For example, by my rough count in March 2010 of CRAN, the R repository, shows that only about 1% of R packages involve Bayesian techniques.[8] The popular commercial statistical package, SAS, also has only limited Bayesian capability.

Significantly, even among Bayesian academics, many use non-Bayesian (called **frequentist**) methods when they work on real, practical problems. Choose an academic statistician at random, and you'll likely find on the Web that he/she does not use Bayesian methods when working on real applications.

[8] There is, however, a book on the topic, *Bayesian Computation with R*, by Jim Albert, Springer, 2007, and among those who use Bayesian techniques, many use R for that purpose.

5.5.4 Arguments Against Use of Subjective Priors

Professor Andrew Gelman is a prominent Bayesian theoretician who has, among other things, published a book on Bayesian theory (*Bayesian Data Analysis* by Andrew Gelman, Donald B. Rubin, John B. Carlin, Hal S. Stern, pub. Taylor and Francis, 2003). His essay, *Why I Don't Like Bayesian Statistics* (`http://www.stat.columbia.edu/~cook/movabletype/archives/2008/04/problems_with_b.html`, should be required reading for anyone considering using subjective priors.

Gelman opens with the statement

> Bayesian inference is a coherent mathematical theory but I wouldn't trust it in scientific applications. Subjective prior distributions don't inspire confidence, and there's no good objective principle for choosing a noninformative prior (even if that concept were mathematically defined, which it's not). Where do prior distributions come from, anyway? I don't trust them and I see no reason to recommend that other people do, just so that I can have the warm feeling of philosophical coherence."

Given that, why is there so much fuss made? Though I personally am traditional, a frequentist, my goal in this section is not to convince the reader of the frequentist point of view. Instead, I want to make sure the reader clearly understands the difference between the two approaches, and the reasons cited by each camp for their views. Since in computer science academia there has recently been considerable interest in the topic, often resulting in what even strong Bayesian statisticians would consider flawed analysis, it's important to understand just what the Bayesian really does.

Since the prior "distribution" of p, $N(0.5, 0.1^2)$ arose from our "gut feeling" or "hunch," it is called a **subjective prior**. *Prior* to collecting any data, we have a certain belief about p. This is very controversial, and many people—including me—consider it to be highly inappropriate. They feel that there is nothing wrong using one's gut feelings to make a final decision, but it should not be part of the mathematical analysis of the data. One's hunches can play a role in deciding the "preponderance of evidence," as discussed in Section 4.5.3, but that should be kept separate from our data analysis. It bothers them that two different Bayesian statisticians can get wildly different answers from the same data—not because they use different tools, but because they have different "hunches." They view Bayesian methods as unscientific, anathema to the fundamental scientific notion of impartiality.

The Bayesians counter by pointing out that in certain situations, a Bayesian estimator may, for instance, produce smaller mean squared estimation error than its frequentist counterpart, even if the prior distribution was just in our imaginations.

On the other hand, say the frequentists, in practice we are typically interested in inference, i.e. confidence intervals and hypothesis tests, rather than point estimation. We are sampling from populations, and want to be able to legitimately make inferences about those populations. For instance, though one can derive a 95% confidence interval for p for our coin, that 95% confidence level is now predicated on the validity of our subjective prior distribution. Since that distribution is nothing more than a hunch, the confidence interval is a hunch too.

The issue thus really boils down to deciding whether one is willing to be subjective in one's approach to science. Use of a subjective prior gives us a subjective answer. For those who are comfortable with that, there is no problem, but most real-world statisticians see it as problematic.

5.5.4.1 What Would You Do?

In evaluating the frequentist/Bayesian debate, you might wish to ask yourself what you would do in the following situations:

- As a personal investor, you've developed a statistical model for the day-to-day price variation of Google stock prices, and will use it to decide whether to buy the stock today. You wish to predict the price of the stock tomorrow, based on its price the last few days. Here are your choices:

  - As a frequentist, you could use a classical mathematical model, say regression analysis (Chapter 7), say fitting a linear or polynomial model. You could use the data to estimate the parameters of the model. This would give you a predicted price for tomorrow. Note that you can still choose to ignore that predicted price in the end, based on a hunch, but you've kept that hunch separate from your data analysis.
  - As a Bayesian, you might use the say linear or polynomial regression model, but you would specify a subjective prior distribution for the parameters. Your predicted price would then be affected by that subjective prior.

  So, what would you deem wise here—a frequentist or Bayesian approach?

- We are in a presidential election, complete with opinion polls as to who is currently winning. As an involved citizen, would you rather that the pollsters simply report the data as is, with their reported margin of error being computed from the traditional frequentist methods we've seen in Chapter 4, or would you prefer that they factor in their own subjective priors?

  So, what would you deem wise here—a frequentist or Bayesian approach?

- You are a physician reading a medical journal article about the effectiveness of a certain drug for alleviating high blood pressure. Would you rather that the authors of the article simply report a straightforward analysis of the data, or would you prefer that the author incorporate a subjective prior distribution into his/her mathematical model?

  So, what would you deem wise here—a frequentist or Bayesian approach?

Chapter 6

Introduction to Model Building

All models are wrong, but some are useful.—George Box[1]

[Mathematical models] should be made as simple as possible, but not simpler.—Albert Einstein[2]

Beware of geeks bearing formulas.—Warrent Buffett, 2009, on the role of "quants" in the 2008 financial collapse.

The above quote by Box says it all. Consider for example the family of normal distributions. In real life, random variables are bounded—no person's height is negative or greater than 500 inches—and are inherently discrete, due to the finite precision of our measuring instruments. Thus, technically, no random variable in practice can have an exact normal distribution. Yet the assumption of normality pervades statistics, and has been enormously successful, provided one understands its approximate nature.

The situation is similar to that of physics. We know that in many analyses of bodies in motion, we can neglect the effect of air resistance, but that in some situations one must include that factor in our model.

So, the field of probability and statistics is fundamentally about *modeling*. The field is extremely useful, provided the user understands the modeling issues well. For this reason, this book contains this separate chapter on modeling issues.

[1] George Box (1919-) is a famous statistician, with several statistical procedures named after him.

[2] The reader is undoubtedly aware of Einstein's (1879-1955) famous theories of relativity, but may not know his connections to probability theory. His work on **Brownian motion**, which describes the path of a molecule as it is bombarded by others, is probabilistic in nature, and later developed into a major branch of probability theory. Einstein was also a pioneer in quantum mechanics, which is probabilistic as well. At one point, he doubted the validity of quantum theory, and made his famous remark, "God does not play dice with the universe."

6.1 "Desperate for Data"

Suppose we have the samples of men's and women's heights described in Section 4.2.11, say we wish to predict the height H of a new person who we know to be a man but for whom we know nothing else.

Recalling our notation from Section 4.2.11, assume that $n_1 = n_2$, and call the common value n. Also, for simplicity, let's assume that $\sigma_1 = \sigma_2 = \sigma$.

The question is, should we take gender into account in our prediction? If so, we might predict the man to be of height[3]

$$T_1 = \overline{X}, \tag{6.1}$$

our estimate for the mean height of all men. If on the other hand we don't take gender into account, then we predict this man's height to be

$$T_2 = \frac{\overline{X} + \overline{Y}}{2}, \tag{6.2}$$

our estimate of the mean height of all people (assuming that half the population is male).

Note that T_2 is based on a simpler model than is T_1, as T_2 ignores gender. We thus refer to T_1 as being based on the more complex model.

Which one is better? The answer will depend on our criterion for goodness of estimation, which we will take to be MSE. So, the question becomes, which has the smaller MSE, T_1 or T_2?

6.1.1 Bias and Variance of the Two Predictors

We could calculate MSE from scratch, but it would probably be better to make use of the work we already went through, producing (4.108). This is especially true in that we know a lot about variance of sample means, and we will take this route.

So, let's find the biases of the two estimators.

- T_1

  T_1 is unbiased, from (4.5). So,

  bias of $T_1 = 0$

[3] Assuming that predicting too high and too low are of equal concern to us, etc.

- T_2

$$\begin{aligned} E(T_2) &= E(0.5\overline{X} + 0.5\overline{Y}) \text{ (definition)} && (6.3)\\ &= 0.5E\overline{X} + 0.5E\overline{Y} \text{ (linearity of E())} && (6.4)\\ &= 0.5\mu_1 + 0.5\mu_2 \text{ [from (4.5)]} && (6.5) \end{aligned}$$

So,

bias of $T_2 = (0.5\mu_1 + 0.5\mu_2) - \mu_1$

On the other hand, T_2 has a smaller variance than T_1:

- T_1

  Recalling (4.9), we have

$$Var(T_1) = \frac{\sigma^2}{n} \tag{6.6}$$

- T_2

$$\begin{aligned} Var(T_2) &= Var(0.5\overline{X} + 0.5\overline{Y}) && (6.7)\\ &= 0.5^2 Var(\overline{X}) + 0.5^2 Var(\overline{Y}) \text{ (properties of Var())} && (6.8)\\ &= \frac{\sigma^2}{2n} \text{ [from 4.9]} && (6.9) \end{aligned}$$

6.1.2 Implications

These findings are highly instructive. You might at first think that "of course" T_1 would be the better predictor than T_2. But for a small sample size, the smaller (actually 0) bias of T_1 is not enough to counteract its larger variance. T_2 is biased, yes, but it is based on double the sample size and thus has half the variance.

In light of (4.108), we see that T_1, the "true" predictor, may not necessarily be the better of the two predictors. Granted, it has no bias whereas T_2 does have a bias, but the latter has a smaller variance.

So, under what circumstances will T_1 be better than T_2? Let's answer this by using (4.107):

$$MSE(T_1) = \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n} \tag{6.10}$$

$$MSE(T_2) = \frac{\sigma^2}{2n} + \left(\frac{\mu_1 + \mu_2}{2} - \mu_1 \right)^2 = \frac{\sigma^2}{2n} + \left(\frac{\mu_2 - \mu_1}{2} \right)^2 \tag{6.11}$$

T_1 is a better predictor than T_2 if (6.10) is smaller than (6.11), which is true if

$$\left(\frac{\mu_2 - \mu_1}{2} \right)^2 > \frac{\sigma^2}{2n} \tag{6.12}$$

So you can see that T_1 is better only if either

- n is large enough, or
- the difference in population mean heights between men and women is large enough, or
- there is not much variation within each population, e.g. most men have very similar heights

Since that third item, small within-population variance, is rarely seen, let's concentrate on the first two items. The big revelation here is that:

> A more complex model is more accurate than a simpler one only if either
>
> - we have enough data to support it, or
> - the complex model is sufficiently different from the simpler one

In height/gender example above, if n is too small, we are "desperate for data," and thus make use of the female data to augment our male data. Though women tend to be shorter than men, the bias that results from that augmentation is offset by the reduction in estimator variance that we get. But if n is large enough, the variance will be small in either model, so when we go to the more complex model, the advantage gained by reducing the bias will more than compensate for the increase in variance.

THIS IS AN ABSOLUTELY FUNDAMENTAL NOTION IN STATISTICS.

This was a very simple example, but you can see that in complex settings, fitting too rich a model can result in very high MSEs for the estimates. In essence, everything becomes noise. (Some people have cleverly coined the term **noise mining**, a play on the term **data mining**.) This is the famous **overfitting** problem.

In our unit on statistical relations, Chapter 7, we will show the results of a scary experiment done at the Wharton School, the University of Pennsylvania's business school. The researchers deliberately added fake data to a prediction equation, and standard statistical software identified it as "significant"! This is partly a problem with the word itself, as we saw in Section 4.5, but also a problem of using far too complex a model, as will be seen in that future unit.

Note that of course (6.12) contains several unknown population quantities. I derived it here merely to establish a principle, namely that a more complex model may perform more poorly under some circumstances.

It would be possible, though, to make (6.12) into a practical decision tool, by estimating the unknown quantities, e.g. replacing μ_1 by $\overline{X}$. This then creates possible problems with confidence intervals, whose derivation did not include this extra decision step. Such estimators, termed **adaptive**, are beyond the scope of this book.

6.2 Assessing "Goodness of Fit" of a Model

Our example in Section 4.6.4 concerned how to estimate the parameters of a gamma distribution, given a sample from the distribution. But that assumed that we had already decided that the gamma model was reasonable in our application. Here we will be concerned with how we might come to such decisions.

Assume we have a random sample $X_1, ..., X_n$ from a distribution having density f_X.

6.2.1 The Chi-Square Goodness of Fit Test

The classic way to do this would be the **Chi-Square Goodness of Fit Test**. We would set

$$H_0 : f_X \text{ is a member of the exponential parametric family} \tag{6.13}$$

This would involve partitioning $(0, \infty)$ into k intervals (s_{i-1}, s_i) of our choice, and setting

$$N_i = \text{number of } X_i \text{ in } (s_{i-1}, s_i) \tag{6.14}$$

We would then find the Maximum Likelihood Estimate (MLE) of λ, on the assumption that the distribution of X really is exponential. The MLE turns out to be the reciprocal of the sample mean, i.e.

$$\widehat{\lambda} = 1/\overline{X} \tag{6.15}$$

This would be considered the parameter of the "best-fitting" exponential density for our data. We would then estimate the probabilities

$$p_i = P[X \epsilon (s_{i-1}, s_i)] = e^{-\lambda s_{i-1}} - e^{-\lambda s_i},\ i = 1, ..., k. \tag{6.16}$$

by

$$\widehat{p_i} = e^{-\widehat{\lambda} s_{i-1}} - e^{-\widehat{\lambda} s_i},\ i = 1, ..., k. \tag{6.17}$$

Note that N_i has a binomial distribution, with n trials and success probability p_i. Using this, the expected value of EN_i is estimated to be

$$\nu_i = n(e^{-\widehat{\lambda} s_{i-1}} - e^{-\widehat{\lambda} s_i}),\ i = 1, ..., k. \tag{6.18}$$

Our test statistic would then be

$$Q = \sum_{i=1}^{k} \frac{(N_i - v_i)^2}{v_i} \tag{6.19}$$

where v_i is the expected value of N_i under the assumption of "exponentialness." It can be shown that Q is approximately chi-square distributed with k-2 degrees of freedom.[4] Note that only large values of Q should be suspicious, i.e. should lead us to reject H_0; if Q is small, it indicates a good fit. If Q were large enough to be a "rare event," say larger than $\chi_{0.95,k-2}$, we would decide NOT to use the exponential model; otherwise, we would use it.

Hopefully the reader has immediately recognized the problem here. If we have a large sample, this procedure will pounce on tiny deviations from the exponential distribution, and we would decide not to use the exponential model—even if those deviations were quite minor. Again, no model is 100% correct, and thus a goodness of fit test will eventually tell us not to use *any* model at all.

6.2.2 Kolmogorov-Smirnov Confidence Bands

Again consider the problem above, in which we were assessing the fit of a exponential model. In line with our major point that confidence intervals are far superior to hypothesis tests, we now present **Kolmogorov-Smirnov confidence bands**, which work as follows.

Recall the concept of empirical cdfs, presented in Section 4.8.1. It turns out that the distribution of

$$M = \max_{-\infty < t \infty} |\widehat{F}_X(t) - F_X(t)| \tag{6.20}$$

[4] We have k intervals, but the N_i must sum to n, so there are only k-1 free values. We then subtract one more degree of freedom, having estimated the parameter λ.

is the same for all distributions having a density. This fact (whose proof is related to the general method for simulating random variables having a given density, in Section 2.7) tells us that, without knowing anything about the distribution of X, we can be sure that M has the same distribution. And it turns out that

$$F_M(1.358n^{-1/2}) = 0.95 \tag{6.21}$$

Define "upper" and "lower" functions

$$U(t) = \widehat{F}_X(t) + 1.358n^{-1/2}, \;\; L(t) = \widehat{F}_X(t) - 1.358n^{-1/2} \tag{6.22}$$

So, what (6.20) and (6.21) tell us is

$$0.95 = P\left(\text{the curve } F_X \text{ is entirely between U and L}\right) \tag{6.23}$$

So, the pair of curves, (L(t), U(t)) is called a a **95% confidence band** for F_X.

The usefulness is similar to that of confidence intervals. If the band is very wide, we know we really don't have enough data to decide much about the distribution of X. But if the band is narrow but some member of the family comes reasonably close to the band, we would probably decide that the model is a good one, even if no member of the family falls within the band. Once again, we should NOT pounce on tiny deviations from the model.

Warning: The Kolmogorov-Smirnov procedure available in the R language performs only a hypothesis test, rather than forming a confidence band. In other words, it simply checks to see whether a member of the family falls within the band. This is not what we want, because we may be perfectly happy if a member is only *near* the band.

Of course, another way, this one less formal, of assessing data for suitability for some model is to plot the data in a histogram or something of that naure.

6.3 Bias Vs. Variance—Again

In our unit on estimation, Section 4.8, we saw a classic tradeoff in histogram- and kernel-based density estimators. With histograms, for instance, the wider bin width produces a graph which is smoother, but possibly *too* smooth, i.e. with less oscillation than the true population curve has. The same problem occurs with larger values of h in the kernel case.

This is actually yet another example of the bias/variance tradeoff, discussed in above and, as mentioned, **ONE OF THE MOST RECURRING NOTIONS IN STATISTICS**. A large bin width, or a large value of h, produces more bias. In general, the large the bin width or h, the further $E[\widehat{f}_R(t)$ is from the true value of $f_R(t)$. This occurs because we are making use of points which are not so near t, and thus at which the density height is different from that of $f_R(t)$. On the other hand, because we are making use of more points, $Var[\widehat{f}_r(t)]$ will be smaller.

THERE IS NO GOOD WAY TO CHOOSE THE BIN WIDTH OR h. Even though there is a lot of theory to suggest how to choose the bin width or h, no method is foolproof. This is made even worse by the fact that the theory generally has a goal of minimizing *integrated* mean squared error,

$$\int_{-\infty}^{\infty} E\left[\left(\widehat{f}_R(t) - f_R(t)\right)^2\right] \, dt \tag{6.24}$$

rather than, say, the mean squared error at a particular point of interest, v:

$$E\left[\left(\widehat{f}_R(t) - f_R(t)\right)^2\right] \tag{6.25}$$

6.4 Robustness

Traditionally, the term *robust* in statistics has meant resilience to violations in assumptions. For example, in Section 4.2.8, we presented Student-t, a method for finding exact confidence intervals for means, assuming normally-distributed populations. But as noted at the outset of this chapter, no population in the real world has an exact normal distribution. The question at hand (which we will address below) is, does the Student-t method still give approximately correct results if the sample population is not normal? If so, we say that Student-t is **robust** to the normality assumption.

Later, there was quite a lot of interest among statisticians in estimation procedures that do well even if there are **outliers** in the data, i.e. erroneous observations that are in the fringes of the sample. Such procedures are said to be robust to outliers.

Our interest here is on robustness to assumptions. Let us first consider the Student-t example. As discussed in Section 4.2.8, the main statistic here is

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \tag{6.26}$$

where μ is the population mean and s is the unbiased version of the sample variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \tag{6.27}$$

The distribution of T, under the assumption of a normal population, has been tabulated, and tables for it appear in virtually every textbook on statistics. But what if the population is not normal, as is inevitably the case?

The answer is that it doesn't matter. For large n, even for samples having, say, n = 20, the distribution of T is close to N(0,1) by the Central Limit Theorem regardless of whether the population is normal.

By contrast, consider the classic procedure for performing hypothesis tests and forming confidence intervals for a population variance σ^2, which relies on the statistic

$$K = \frac{(n-1)s^2}{\sigma^2} \tag{6.28}$$

where again s^2 is the unbiased version of the sample variance. If the sampled population is normal, then K can be shown to have a chi-square distribution with n-1 degrees of freedom. This then sets up the tests or intervals. However, it has been shown that these procedures are not robust to the assumption of a normal population. See *The Analysis of Variance: Fixed, Random, and Mixed Models*, by Hardeo Sahai and Mohammed I. Ageel, Springer, 2000, and the earlier references they cite, especially the pioneering work of Scheffe'.

Exercises

Note to instructor: See the Preface for a list of sources of real data on which exercises can be assigned to complement the theoretical exercises below.

1. In our example in Section 6.1, assume $\mu_1 = 70, \mu_2 = 66, \sigma = 4$ and the distribution of height is normal in the two populations. Suppose we are predicting the height of a man who, unknown to us, has height 68. We hope to guess within two inches. Find $P(|T_1 - 68|) < 2$ and $P(|T_2 - 68|) < 2$ for various values of n.

2. In Section 5.3 we discussed *simultaneous inference*, the forming of confidence intervals whose joint confidence level was 95% or some other target value. The Kolmogorov-Smirnov confidence band in Section 6.2.2 allows us to computer infinitely many confidence intervals for $F_X(t)$ at different values of t, at a "price" of only 1.358. Still, if we are just estimating $F_X(t)$ at a single value of t, an individual confidence interval using (4.34) would be narrower than that given to us by Kolmogorov-Smirnov. Compare the widths of these two intervals in a situation in which the true value of $F_X(t) = 0.4$.

3. Say we have a random sample $X_1, ..., X_n$ from a population with mean μ and variance σ^2. The usual estimator of μ is the sample mean $\bar{X}$, but here we will use what is called a *shrinkage estimator*: Our estimate of μ will be $0.9\bar{X}$. Find the mean squared error of this estimator, and give an inequality (you don't have to algebraically simplify it) that shows under what circumstances $0.9\bar{X}$ is better than $\bar{X}$. (Strong advice: Do NOT "reinvent the wheel." Make use of what we have already derived.)

Chapter 7

Statistical Relations Between Variables

7.1 The Goals: Prediction and Understanding

Prediction is difficult, especially when it's about the future.—Yogi Berra[1]

In this unit we are interested in relations between variables. Before beginning, it is important to understand the typical goals in analyzing such relations:

- **Prediction:** Here we are trying to predict one variable from one or more others.
- **Understanding:** Here we wish to determine which variables have a greater effect on a given variable.

Denote the predictor variables by, $X^{(1)}, ..., X^{(r)}$. The variable to be predicted, Y, is often called the **response 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**,[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 unit.

[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."

[2] Sometimes called a **dummy variable**.

7.2 Example Applications: Software Engineering, Networks, Text Mining

Example: As an aid in deciding which applicants to admit to a graduate program in computer science, we might try to predict Y, a faculty rating of a student after completion of his/her first year in the program, from $X^{(1)}$ = the student's CS GRE score, $X^{(2)}$ = the student's undergraduate GPA and various other variables. Here our goal would be Prediction, but educational researchers might do the same thing with the goal of Understanding. For an example of the latter, see Predicting Academic Performance in the School of Computing & Information Technology (SCIT), *35th ASEE/IEEE Frontiers in Education Conference*, by Paul Golding and Sophia McNamarah, 2005.

Example: In a paper, Estimation of Network Distances Using Off-line Measurements, *Computer Communications*, by Danny Raz, Nidhan Choudhuri and Prasun Sinha, 2006, the authors wanted to predict Y, the round-trip time (RTT) for packets in a network, using the predictor variables $X^{(1)}$ = geographical distance between the two nodes, $X^{(2)}$ = number of router-to-router hops, and other variables. The goal here is primarily Prediction.

Example: In a paper, Productivity Analysis of Object-Oriented Software Developed in a Commercial Environment, *Software—Practice and Experience*, by Thomas E. Potok, Mladen Vouk and Andy Rindos, 1999, the authors mainly had an Understanding goal: What impact, positive or negative, does the use of object-oriented programming have on programmer productivity? Here they predicted Y = number of person-months needed to complete the project, from $X^{(1)}$ = size of the project as measured in lines of code, $X^{(2)}$ = 1 or 0 depending on whether an object-oriented or procedural approach was used, and other variables.

Example: Most **text mining** applications are classification problems. For example, the paper Untangling Text Data Mining, *Proceedings of ACL'99*, by Marti Hearst, 1999 cites, *inter alia*, an application in which the analysts wished to know what proportion of patents come from publicly funded research. They were using a patent database, which of course is far too huge to feasibly search by hand. That meant that they 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.

7.3 Regression Analysis

7.3.1 What Does "Relationship" Really Mean?

Consider the Davis city population example again. In addition to the random variable W for weight, let H denote the person's height. Suppose we are interested in exploring the relationship between height and weight.

As usual, we must first ask, **what does that really mean**? What do we mean by "relationship"? Clearly, there is no exact relationship; for instance, we cannot exactly predict a person's weight from 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

$$m_{W;H}(t) = E(W|H = t) \tag{7.1}$$

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.

Now, let's again suppose we have a random sample of 1000 people from Davis, with

$$(H_1, W_1), ..., (H_{1000}, W_{1000}) \tag{7.2}$$

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. That means we are estimating infinitely many values, with one $m_{W;H}(t)$ value for each t.[3] How do we do this?

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.

Typically the parametric model chosen is linear, i.e. we assume that $m_{W;H}(t)$ is a linear function of t:

$$m_{W;H}(t) = ct + d \tag{7.3}$$

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 (7.3) 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.

[3] Of course, the population of Davis is finite, but there is the conceptual population of all people who *could* live in Davis.

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 will see how to estimate c and d in Section 7.3.7.

7.3.2 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

$$X = \left(X^{(1)}, X^{(2)}\right) = (H, A) = \text{(height, age)} \tag{7.4}$$

so as to study the relationship of weight with height and age. If we used a linear model, we would write for $t = (t_1, t_2)$,

$$m_{W;H,A}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 \tag{7.5}$$

In other words

$$\text{mean weight} = \beta_0 + \beta_1 \text{ height} + \beta_2 \text{ age} \tag{7.6}$$

(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.

7.3.3 Interaction Terms

Equation (7.5) 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 we are looking at tall people or short people. To see that, just plug 40 and 30 for age in (7.5), with the same number for height in both, and subtract; you get $10\beta_2$, an expression that has no height term.

If we feel that the assumption is not a good one (there are also data plotting techniques to help assess this), we can add an **interaction term** to (7.5), 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

$$m_{W;H}(t) = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_1 t_2 \tag{7.7}$$

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 7.1. The authors there set $X^{(3)} = X^{(1)}X^{(2)}$. The reader should make sure to understand that without this term, we are basically saying that the effect (whether positive or negative) of using object-oriented programming is the same for any code size.

Though the idea of adding interaction terms to a regression model is tempting, it can easily get out of hand. If we have k basic predictor variables, then there are $\binom{k}{2}$ potential two-way interaction terms, $\binom{k}{3}$ three-way terms and so on. Unless we have a very large amount of data, we run a big risk of overfitting (Section 7.3.9.1). And with so many interaction terms, the model would be difficult to interpret.

7.3.4 Nonrandom Predictor Variables

In our weight/height/age example above, all three variables are random. If we repeat the "experiment," i.e. we choose another sample of 1000 people, these new people will have different weights, different heights and different ages from the people in the first sample.

But we must point out that the function $m_{Y;X}$ makes sense even if X is nonrandom. To illustrate this, let's look at the ALOHA network example in our introductory unit on discrete probability, Section 1.1.

```
# 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
alohasim <- function(s,b,q,nslots) {
   # status[i,1] = 1 or 0, for node i active or not
   # status[i,2] = if node i active, then epoch in which msg was created
   # (could try a list structure instead a matrix)
   status <- matrix(nrow=s,ncol=2)
   # start with all active with msg created at time 0
   for (node in 1:s) status[node,] <- c(1,0)
   nsent <- 0  # number of successful transmits so far
   sumdelay <- 0  # total delay among successful transmits so far
   # now simulate the nslots slots
   for (slot in 1:nslots) {
      # check for new actives
      for (node in 1:s) {
         if (!status[node,1])  # inactive
            if (runif(1) < q) status[node,] <- c(1,slot)
      }
      # check for attempted transmissions
      ntrysend <- 0
      for (node in 1:s) {
         if (status[node,1])  # active
            if (runif(1) > b) {
               ntrysend <- ntrysend + 1
               whotried <- node
            }
      }
      if (ntrysend == 1) {  # something gets through iff exactly one tries
         # do our bookkeeping
         sumdelay <- sumdelay + slot - status[whotried,2]
         # this node now back to inactive
         status[whotried,1] <- 0
         nsent <- nsent + 1
      }
   }
   return(c(b,sumdelay/nsent))
}
```

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. We are particularly interested in the effect of b here on that mean. Our goal here, as described in Section 7.1, could be Prediction, so that we could have an idea of

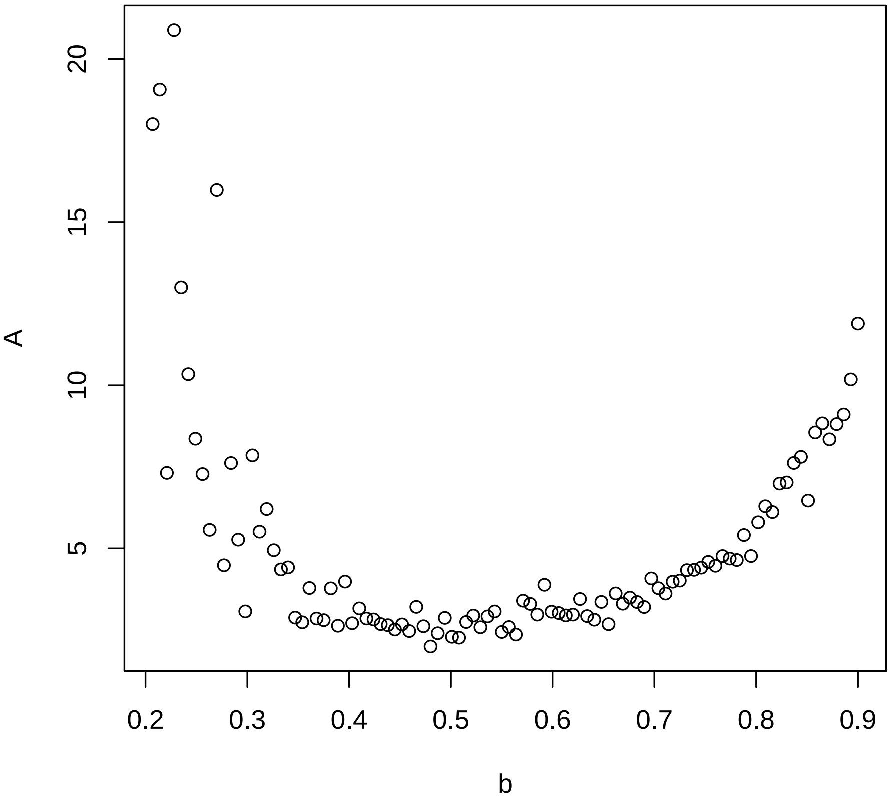


Figure 7.1: Scatter Plot

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")
```

The plot is shown in Figure 7.1.

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

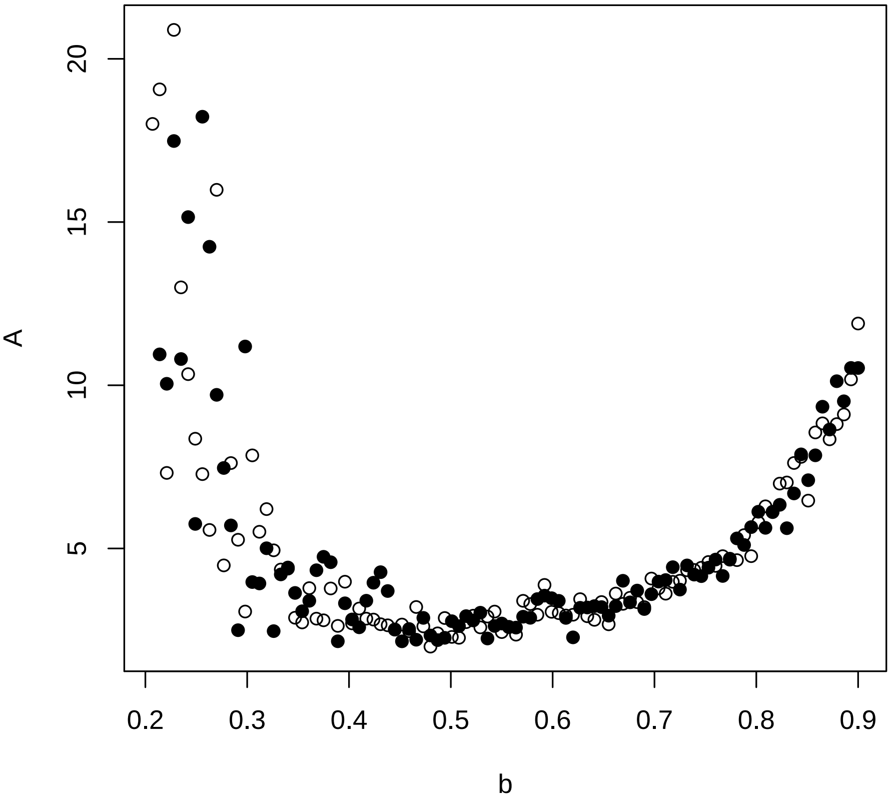


Figure 7.2: Scatter Plot, Two Data Sets

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 7.2.

We do expect some kind of U-shaped relation, as seen here. For b too small, the nodes are clashing with each other a lot, causing long delays to message transmission. For b too large, we are needlessly backing off in many cases in which we actually would get through.

This looks like a quadratic relationship, meaning the following. Take our response variable Y to be A, take our first predictor $X^{(1)}$ to be b, and take our second predictor $X^{(2)}$ to be b^2. Then when we say A and b have a quadratic relationship, we mean

$$m_{A;b}(b) = \beta_0 + \beta_1 b + \beta_2 b^2 \tag{7.8}$$

for some constants $\beta_0, \beta_1, \beta_2$. So, we are using a three-parameter family for our model of $m_{A;b}$. No model is exact, but our data seem to indicate that this one is reasonably good, and if further investigation confirms that, it provides for a nice compact summary of the situation.

Again, we'll see how to estimate the β_i in Section 7.3.7.

We could also try adding two more predictor variables, consisting of $X^{(3)} = q$ and $X^{(4)} = s$. We would collect more data, in which we varied the values of q and s, and then could entertain the model

$$m_{A;b}(b) = \beta_0 + \beta_1 b + \beta_2 b^2 + \beta_3 q + \beta_4 s \tag{7.9}$$

7.3.5 Prediction

So, we've taken our data on weight/height/age, and estimated the function m using that data, yielding $\widehat{m}$. Now, a new person comes in, of height 70.4 and age 24.8. What should we predict his weight to be?

The answer is that we predict his weight to be our estimated mean weight for his height/age group,

$$\widehat{m}_{W;H,A}(70.4, 24.8) \tag{7.10}$$

If our model is (7.5), then (7.10) is

$$m_{W;H}(t) = \widehat{\beta}_0 + \widehat{\beta}_1 70.4 + \widehat{\beta}_2 24.8 \tag{7.11}$$

where the $\widehat{\beta}_i$ are estimated from our data as in Section 7.3.7 below.

7.3.6 Optimality of the Regression Function

In predicting Y from X (with X random), we might assess our predictive ability by the **mean squared prediction error** (MSPE):

$$\text{MSPE} = E\left[(Y - w(X))^2\right] \tag{7.12}$$

where w is some function we will use to form our prediction for Y based on X. What w is best, i.e. which w minimizes MSPE?

To answer this question, condition on X in (7.12):

$$\text{MSPE} = E\left[E\{(Y - w(X))^2|X\}\right] \tag{7.13}$$

Theorem 10 *The best w is m, i.e. the best way to predict Y from X is to "plug in" X in the regression function.*

We need this lemma:

Lemma 11 *For any random variable Z, the constant c which minimizes*

$$E[(Z - c)^2] \tag{7.14}$$

is

$$c = EZ \tag{7.15}$$

Proof

Expand (7.14) to

$$E(Z^2) - 2cEZ + c^2 \tag{7.16}$$

and use calculus to find the best c.

■

Apply the lemma to the inner expectation in (7.13), with Z being Y and c being some function of X. The minimizing value is EZ, i.e. $E(Y|X)$ since our expectation here is conditional on X.

All of this tells us that the best function w in (7.12) is $m_{Y;X}$. This proves the theorem.

Note carefully that all of this was predicated on the use of a quadratic loss function, i.e. on minimizing mean squared error. If instead we wished to minimize mean absolute error, the solution would turn out to be to use the conditional median of Y given X, not the mean.

7.3.7 Parametric Estimation of Linear Regression Functions

7.3.7.1 Meaning of "Linear"

Here we model $m_{Y;X}$ as a linear function of $X^{(1)}, ..., X^{(r)}$:

$$m_{Y;X}(t) = \beta_0 + \beta_1 t^{(1)} + ... + \beta_r t^{(r)} \tag{7.17}$$

Note that the term **linear regression** does NOT necessarily mean that the graph of the regression function is a straight line or a plane. Instead, the word *linear* refers to the regression function being linear in the parameters. So, for instance, (7.8) is a linear model; if for example we multiple β_0, β_1 and β_2 by 8, then m is multiplied by 8.

A more literal look at the meaning of "linear" comes from the matrix formulation (7.21) below.

7.3.7.2 Point Estimates and Matrix Formulation

So, how do we estimate the β_i? Look for instance at (7.8). Keep in mind that in (7.8), the β_i are population values. We need to estimate them from our data. How do we do that?

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 $\hat{\beta}_i$. Using the result of Section 7.3.5 as a guide, 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:

$$\sum_{i=1}^{100} [A_i - (\hat{\beta}_0 + \hat{\beta}_1 b_i + \hat{\beta}_2 b_i^2)]^2 \tag{7.18}$$

Obviously, this is a calculus problem. We set the partial derivatives of (7.18) with respect to the $\hat{\beta}_i$ to 0, giving use three linear equations in three unknowns, and then solve.

For the general case (7.17), 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 (7.3.7.1), we have

$$E(Y_i | X_i^{(1)}, ..., X_i^{(r)}) = \beta_0 + \beta_1 X_i^{(1)} + ... + \beta_r X_i^{(r)}, \ i = 1, ..., n \tag{7.19}$$

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, then row 5 of Q would consist of a 1, then the height and age of the fifth person in our sample.

Also, let

$$V = (Y_1, ..., Y_n)', \tag{7.20}$$

Then the system (7.19) in matrix form is

$$E(V|Q) = Q\beta \tag{7.21}$$

where (with ′ denoting matrix transpose and a vector without a ′ being a row vector)

$$\beta = (\beta_0, \beta_1, ..., \beta_r)' \tag{7.22}$$

Now to estimate the β_i, let

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_r)' \tag{7.23}$$

with our goal now being to find $\hat{\beta}$. The matrix form of (7.18) (now for the general case, not just ALOHA) is

$$(V - Q\beta)'(V - Q\beta) \tag{7.24}$$

Then it can be shown that, after all the partial derivatives are taken and set to 0, the solution is

$$\hat{\beta} = (Q'Q)^{-1}Q'V \tag{7.25}$$

Note by the way that this implies that $\hat{\beta}$ is an unbiased estimate of β:

$$\begin{aligned}
E\hat{\beta} &= E[(Q'Q)^{-1}Q'V] \;\; (7.25) && (7.26)\\
&= (Q'Q)^{-1}Q'EV \;\; \text{(linearity of E())} && (7.27)\\
&= (Q'Q)^{-1}Q' \cdot Q\beta \;\; (7.21) && (7.28)\\
&= \beta && (7.29)
\end{aligned}$$

In some applications, we assume there is no constant term β_0 in (7.17). This means that our Q matrix no longer has the column of 1s on the left end, but everything else above is valid.

7.3.7.3 Back to Our ALOHA Example

R or any other statistical package does the work for us. In R, we can use the **lm()** ("linear model") function:

```
> md <- cbind(md,md[,1]ˆ2)
> 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.[4]

The return value from this call, which I've stored in **lmout**, is an object of class **lm**. One of the member variables of that class, **coefficients**, is the vector $\widehat{\beta}$:

```
> lmout$coefficients
(Intercept)     md[, 1]     md[, 3]
   27.56852   -90.72585    79.98616
```

So, $\widehat{\beta}_0 = 27.57$ and so on.

The result is

$$\widehat{m}_{A,b}(t) = 27.57 - 90.73t + 79.99t^2 \tag{7.30}$$

Another member variable in the **lm** class is **fitted.values**. This is the "fitted curve," meaning the values of (7.30) at $b_1, ..., b_{100}$. In other words, this is (7.30). I plotted this curve on the same graph,

```
> lines(cbind(md[,1],lmout$fitted.values))
```

See Figure 7.3. 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.

[4] 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.

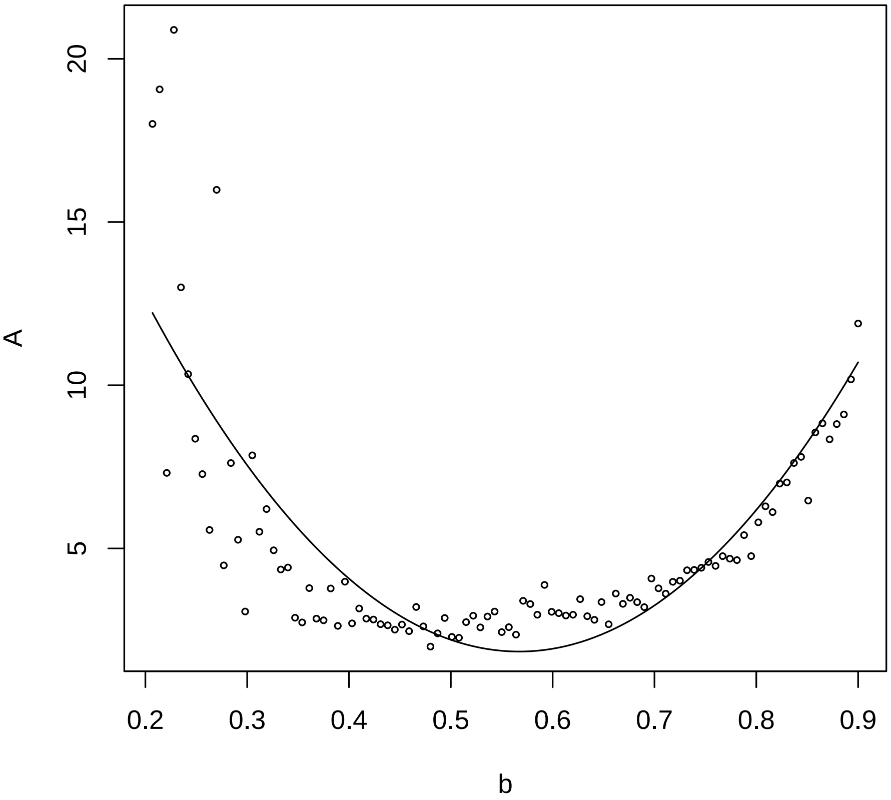


Figure 7.3: Quadratic Fit Superimposed

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

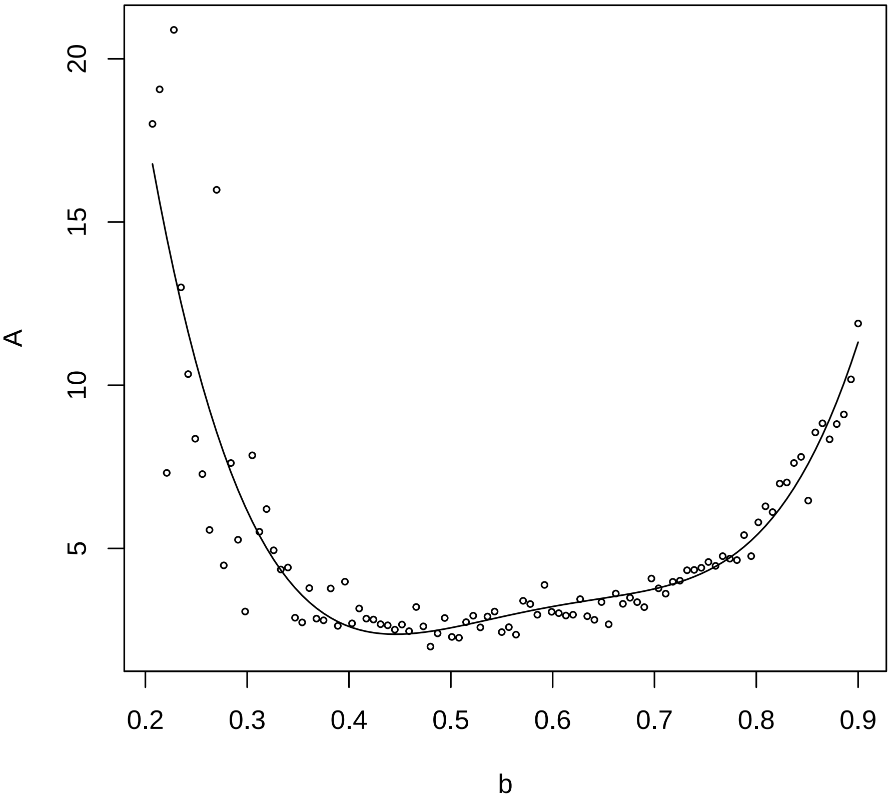


Figure 7.4: Fourth Degree Fit Superimposed

```
> lmout$coefficients
(Intercept)    md4[, 1]    md4[, 3]    md4[, 4]    md4[, 5]
   95.98882  -664.02780  1731.90848 -1973.00660   835.89714
```

In other words, we have an estimated regression function of

$$\widehat{m}_{A,b}(t) = 95.98882 - 664.02780\, t + 1731.90848\, t^2 - 1973.00660\, t^3 + 835.89714\, t^4 \tag{7.31}$$

The fit is shown in Figure 7.4. It looks much better. On the other hand, we have to worry about overfitting. We return to this issue in Section 7.3.9.1).

7.3.7.4 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 7.1. 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 (7.25) 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 (7.17). Now, though, we need to add an assumption:[5]

$$Var(Y|X = t) = \sigma^2 \tag{7.32}$$

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

$$Cov(V|Q) = \sigma^2 I \tag{7.33}$$

where I is the usual identiy matrix (1s on the diagonal, 0s off diagonal).

Be sure you understand what this means. In the Davis weights example, for instance, it means that the variance of weight among 72-inch tall people is the same as that for 65-inch-tall people. That is not quite true—the taller group has larger variance—but it's probably accurate enough for our purposes here.

Keep in mind that the derivation below is conditional on the $X_j^{(i)}$, which 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 Cov(V) instead of Cov(V|Q).

We can derive the covariance matrix of $\hat{\beta}$ as follows. Again to avoid clutter, let $B = (Q'Q)^{-1}$. Theorem from linear algebra say that Q'Q is symmetric and thus B is too. Another theorem says that for any conformable matrices U and V, then (UV)' = V'U'. Armed with that knowledge, here we go:

[5] Actually, we could derive some usable, though messy,standard errors without this assumption.

$$
\begin{aligned}
Cov(\hat{\beta}) &= Cov(BQ'V) \text{ ((7.25))} && (7.34)\\
&= BQ'Cov(V)(BQ')' \text{ (3.80)} && (7.35)\\
&= BQ'\sigma^2 I(BQ')' \text{ (7.33)} && (7.36)\\
&= \sigma^2 BQ'QB \text{ (lin. alg.)} && (7.37)\\
&= \sigma^2 (Q'Q)^{-1} \text{ (def. of B)} && (7.38)
\end{aligned}
$$

Whew! That's a lot of work for you, if your linear algebra is rusty. But it's worth it, because (7.38) now gives us what we need for confidence intervals. Here's how:

First, we need to estimate σ^2. Recall first that for any random variable U, $Var(U) = E[(U - EU)^2]$, we have

$$
\begin{aligned}
\sigma^2 &= Var(Y|X = t) && (7.39)\\
&= Var(Y|X^{(1)} = t_1, ..., X^{(r)} = t_r) && (7.40)\\
&= E\left[\{Y - m_{Y;X}(t)\}^2\right] && (7.41)\\
&= E\left[(Y - \beta_0 - \beta_1 t_1 - ... - \beta_r t_r)^2\right] && (7.42)
\end{aligned}
$$

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:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i^{(1)} - ... - \hat{\beta}_r X_i^{(r)})^2 \tag{7.43}$$

As in Chapter 4, 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

$$\widehat{Cov}(\hat{\beta}) = s^2(Q'Q)^{-1} \tag{7.44}$$

7.3.7.5 Once Again, Our ALOHA Example

In R we can obtain (7.44) 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 (7.44) 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

$$835.89714 \pm 1.96 \cdot 142.6 = (556.4, 1115.4) \tag{7.45}$$

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. If in an election poll the survey finds that 80% of the voters polled say they will vote for Candidate Z, but the margin of error is large, say 10%, it still means that Z is very comfortably ahead.

By the way, 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.

7.3.7.6 Estimation Vs. Prediction

In statistical parlance, there is a keen distinction made between the words *estimation* and *prediction*. To explain this, let's again consider the example of predicting Y = weight from X = (height,age). Say we have someone of height 67 inches and age 27, and want to guess—i.e. *predict*—her weight.

From Section 7.3.6, we know that the best prediction is m[(67,27)]. However, we do not know the value of that quantity, so we must *estimate* it from our data. So, our *predicted value* for this person's weight will be $\hat{m}[(67, 27)]$, i.e. our *estimate* for the value of the regression function at the point (67,27).

7.3.7.7 Exact Confidence Intervals

Note carefully that we have not assumed that Y, given X, is normally distributed. In the height/weight context, for example, such an assumption would mean that weights in a specific height subpopulation, say all people of height 70 inches, have a normal distribution.

If we do make such an assumption, then we can get exact confidence intervals (which of course, only hold if we really do have an exact normal distribution in the population). This again uses Student-t distributions.

In that analysis, s^2 has n-(r+1) in its denominator instead of our n, just as there was n-1 in the denominator for s^2 when we estimated a single population variance. The number of degrees of freedom in the Student-t distribution is likewise n-(r+1). But as before, for even moderately large n, it doesn't matter.

7.3.8 The Famous "Error Term" (advanced topic)

Books on 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

$$Y = \beta_0 + \beta_1 X + \epsilon, \ E\epsilon = 0 \tag{7.46}$$

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

$$m_{Y;X}(t) = E(Y|X = t) = \beta_0 + \beta_1 t \tag{7.47}$$

Note that in our context, we would define ϵ as

$$\epsilon = Y - m_{Y;X}(X) \tag{7.48}$$

Equation (7.46) is consistent with (7.47): The former has $E\epsilon = 0$, and so does the latter, since

$$E\epsilon = EY - E[m_{Y;X}(X)] = EY - E[E(Y|X)] = EY - EY = 0 \tag{7.49}$$

In order to produce confidence intervals, we later added the assumption (7.32), which you can see is consistent with (7.46) since the latter assumes that $Var(\epsilon) = \sigma^2$ no matter what value X has.

Now, what about the normality assumption in (7.46)? 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.

7.3.9 Model Selection

The issues raised in Chapter 6 become crucial in regression and classification problems. In this unit, we will typically deal with models having large numbers of parameters. A central principle will be that simpler models are preferable, provided of course they are accurate. Hence the Einstein quote above. Simpler models are often called **parsimonious**.

Here I use the term *model selection* to mean which predictor variables we will use. If we have data on many predictors, we almost certainly will not be able to use them all, for the following reason:

7.3.9.1 The Overfitting Problem in Regression

Recall (7.8). There we assumed a second-degree polynomial for $m_{A;b}$. Why not a third-degree, or fourth, and so on?

You can see that if we carry this notion to its extreme, we get absurd results. If we fit a polynomial of degree 99 to our 100 points, we can make our fitted curve exactly pass through every point! This clearly would give us a meaningless, useless curve. We are simply fitting the noise.

Recall that we analyzed this problem in Section 6.1.2 in our unit on modeling. testing. There we noted an absolutely fundamental principle in statistics:

> In choosing between a simpler model and a more complex one, the latter is more accurate only if either
>
> - we have enough data to support it, or
> - the complex model is sufficiently different from the simpler one

This is extremely important in regression analysis. For example, look at our regression model for A against b in the ALOHA simulation in earlier sections. We did analyses for a simpler model, a quadratic polynomial, and a more complex model, a quartic (polynomial of degree 4). Rephrasing the above points in this context, we would say,

> In choosing between the quadratic and quartic models, the latter is more accurate only if either
>
> - we have enough data to support it, or
> - at least one of the coefficients β_3 and β_4 is quite different from 0

In the weight/height/age example in Section 7.3.1, this would be phrased as

> In deciding whether to predict from height only, versus from both height and age, the latter is more accurate only if either
>
> - we have enough data to support it, or
> - the coefficient β_2 is quite different from 0

If we use too many predictor variables,[6], our data is "diluted," by being "shared" by so many β_i. As a result, $Var(\beta_i)$ will be large, with big implications: Whether our goal is Prediction or Understanding, our estimates will be so poor that neither goal is achieved.

The questions raised in turn by the above considerations, i.e. **How much data** is enough data?, and **How different** from 0 is "quite different"?, are addressed below in Section 7.3.9.2.

A detailed mathematical example of overfitting in regression is presented in my paper A Careful Look at the Use of Statistical Methodology in Data Mining (book chapter), by N. Matloff, in *Foundations of Data Mining and Granular Computing*, edited by T.Y. Lin, Wesley Chu and L. Matzlack, Springer-Verlag Lecture Notes in Computer Science, 2005.

7.3.9.2 Methods for Predictor Variable Selection

So, we typically must discard some, maybe many, of our predictor variables. In the weight/height/age example, we may need to discard the age variable. In the ALOHA example, we might need to discard b^4 and even b^3. How do we make these decisions?

Note carefully that **this is an unsolved problem.** If anyone ever claims they have a foolproof way to do this, they do not understand the problem in the first place. Entire books have been written on this subject (e.g. *Subset Selection in Regression*, by Alan Miller, pub. by Chapman and Hall, 2002), discussing myriad different methods, but again, none of them is foolproof.

Most of the methods for variable selection use hypothesis testing in one form or another. Typically this takes the form

$$H_0 : \beta_i = 0 \tag{7.50}$$

In the context of (7.6), this would mean testing

$$H_0 : \beta_2 = 0 \tag{7.51}$$

If we reject H_0, then we use the age variable; otherwise we discard it.

[6] In the ALOHA example above, b, b^2, b^3 and b^4 are separate predictors, even though they are of course correlated.

I hope I've convinced you 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, whereas *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 (7.50). 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.[7] This problem has always been known by thinking statisticians, but the Wharton study certainly dramatized it.

Well, then, what can be done instead? First, there is the same alternative to hypothesis testing that we discussed before—confidence intervals. We saw an example of that in (7.45). 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 pretty safe to use b^4 as a predictor.

Moreover, a confidence interval for β_i tells us whether the variable $X^{(i)}$ would have much value as a predictor. Once again, consider the weight/height/age example. Suppose our confidence interval for β_2 is (0.04,0.06). In other words, we estimate β_2 to be 0.05, with a margin of error of 0.01. The 0.01 is telling us that our sample size is good enough for an accurate assessment of the situation, but the interval's location—centered at 0.05–says, for instance, a 10-year difference in age only makes about half a pound difference in mean weight. In that situation age would be of almost no value in predicting weight.

A method that enjoys some popularity in certain circles is the **Akaike Information Criterion** (AIC). It uses a formula, backed by some theoretical analysis, which creates a tradeoff between richness of the model and size of the standard errors of the $\hat{\beta}_i$. The R statistical package includes a function **AIC()** for this, which is used by **step()** in the regression case.

The most popular alternative to hypothesis testing for variable selection today is probably **cross validation**. Here we split our data into a **training set**, which we use to estimate the β_i, and a **validation set**, in which we see how well our fitted model predicts new data, say in terms of average squared prediction error. We do this for several models, i.e. several sets of predictors, and choose the one which does best in the validation set. I like this method very much, though I often simply stick with confidence intervals.

A rough rule of thumb is that one should have $r < \sqrt{n}$.[8]

[7] They added so many predictors that r became greater than n. However, the problems they found would have been there to a large degree even if r were less than n but r/n was substantial.

[8] Asymptotic Behavior of Likelihood Methods for Exponential Families When the Number of Parameters Tends to Infinity, Stephen Portnoy, *Annals of Statistics*, 1968.

7.3.10 Nonlinear Parametric Regression Models

We pointed out in Section 7.3.7.1 that the word *linear* in *linear regression model* means linear in β, not in t. This is the most popular approach, as it is computationally easy, but nonlinear models are often used.

The most famous of these is the **logistic** model, for the case in which Y takes on only the values 0 and 1. As we have seen before, in this case the expected value becomes a probability. The logistic model for a nonvector X is then

$$m_{Y;X}(t) = P(Y = 1|X = t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 t})} \tag{7.52}$$

It extends to the case of vector-valued X in the obvious way.

The logistic model is quite widely used in computer science, in medicine, economics, psychology and so on.

Here is an example of a nonlinear model used in kinetics of chemical reactions, with r = 3: [9]

$$m_{Y;X}(t) = \frac{\beta_1 t^{(2)} - t^{(3)}/\beta_5}{1 + \beta_2 t^{(1)} + \beta_3 t^{(2)} + \beta_4 t^{(3)}} \tag{7.53}$$

Here the X vector is (hydrogen, n-pentane, isopentane)'.

Unfortunately, in most cases, the least-squares estimates of the parameters in nonlinear regression do not have closed-form solutions, and numerical methods must be used. But R does that for you, via the **nls()** function in general, and via **glm()** for the logistic and related models in particular.

7.3.11 Nonparametric Estimation of Regression Functions

In some applications, there may be no obvious parametric model for $m_{Yl;X}$. Or, we may have a parametric model that we are considering, but we would like to have some kind of nonparametric estimation method available as a means of checking the validity of our parametric model. So, how do we estimate a regression function nonparametrically?

To guide our intuition on this, let's turn again of the Davis example of the relationship between height and weight. Consider estimation of the quantity $m_{W;H}(68.2)$, the *population* mean weight of all people of height 68.2. We could take our estimate $\hat{m}_{W;H}(68.2)$ to be the average weight of all the people in our sample who have that height. But we may have very few people of that height, so that our estimate may have a high variance, i.e. may not be very accurate.

[9] See `http://www.mathworks.com/index.html?s_cid=docframe_homepage`.

What we could do instead is to take the mean weight of all the people in our sample whose heights are *near* 68.2, say between 67.7 and 68.7. That would bias things a bit, but we'd get a lower variance. All nonparametric regression methods work like this, though with many variations.

As our definition of "near," we could take all people in our sample whose heights are within h amount of 68.2. This should remind you of our density estimators in Section 4.8 of our unit on estimation and testing. As we saw there, a generalization would be to use a kernel method. For instance, for univariate X and t:

$$\hat{m}_{Y;X}(t) = \frac{\sum_{i=1}^n Y_i k \left(\frac{t-X_i}{h} \right)}{\sum_{i=1}^n k \left(\frac{t-X_i}{h} \right)} \tag{7.54}$$

There is an R package that includes a function **nkreg()** to do this. The R base has a similar method, called **LOESS**. Note: That is the method name, but the R function is called **lowess()**.

Other types of nonparametric methods include **Classification and Regression Trees** (CART), nearest-neighbor methods, support vector machines, splines etc.

7.3.12 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. (7.32)], 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.

7.3.13 Nominal Variables

Recall our example in Section 7.2 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 called 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**.

7.3.14 The Case in Which All Predictors Are Nominal Variables: "Analysis of Variance"

Continuing the ideas in Section 7.3.13, suppose in the software engineering study they had kept the project size constant, and instead of $X^{(1)}$ being project size, this variable recorded whether the programmer uses an integrated development environment (IDE). Say $X^{(1)}$ is 1 or 0, depending on whether the programmer uses the Eclipse IDE or no IDE, respectively. Continue to assume the study included the nominal Language variable, i.e. assume the study included the indicator variables $X^{(2)}$ (C++) and $X^{(3)}$ (Java). Now all of our predictors would be nominal/indicator variables. Regression analysis in such settings is called **analysis of variance** (ANOVA).

Each nominal variable is called a **factor**. So, in our software engineering example, the factors are IDE and Language. Note again that in terms of the actual predictor variables, each factor is represented by one or more indicator variables; here IDE has one indicator variables and Language has two.

Analysis of variance is a classic statistical procedure, used heavily in agriculture, for example. We will not go into details here, but mention it briefly both for the sake of completeness and for its relevance to Sections 7.3.3 and 7.6. (The reader is strongly advised to review Sections 7.3.3 before continuing.)

7.3.14.1 It's a Regression!

The term *analyisis of variance* is a misnomer. A more appropriate name would be **analysis of means**, as it is in fact a regression analysis, as follows.

First, note in our software engineering example we basically are talking about six groups, because there are six different combinations of values for the triple $(X^{(1)}, X^{(2)}, X^{(3)})$. For instance, the triple (1,0,1) means that the programmer is using an IDE and programming in Java. Note that triples of the form (w,1,1) are impossible.

So, all that is happening here is that we have six groups with six means. But that is a regression! Remember, for variables U and V, $m_{V;U}(t)$ is the mean of all values of V in the subpopulation group of people (or cars or whatever) defined by U = s. If U is a continuous variable, then we have infinitely many such groups, thus infinitely many means. In our software engineering example, we only have six groups, but the principle is the same. We can thus cast the problem in regression terms:

$$m_{Y;X}(i,j,k) = E(Y|X^{(1)} = i, X^{(2)} = j, X^{(3)} = k),\ i,j,k = 0,1,j+k \leq 1 \tag{7.55}$$

Note the restriction $j + k \leq 1$, which reflects the fact that j and k can't both be 1.

Again, keep in mind that we are working with means. For instance, $m_{Y;X}(0,1,0)$ is the population mean project completion time for the programmers who do not use Eclipse and who program in C++.

Since the triple (i,j,k) can take on only six values, m can be modeled fully generally in the following six-parameter linear form:

$$m_{Y;X}(i,j,k) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 ij + \beta_5 ik \tag{7.56}$$

where β_4 and β_5 are the coefficients of two interaction terms, as in Section 7.3.3.

7.3.14.2 Interaction Terms

It is crucial to understand the interaction terms. Without the ij and ik terms, for instance, our model would be

$$m_{Y;X}(i,j,k) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k \tag{7.57}$$

which would mean (as in Section 7.3.3) that the difference between using Eclipse and and no IDE is the same for all three programming languages, C++, Java and C. That common difference would be β_1. If this condition—the impact of using an IDE is the same across languages—doesn't hold, at least approximately, then would use the full model, (7.56). More on this below.

Note carefully that there is no interaction term corresponding to jk, since that quantity is 0, and thus there is no three-way interaction term corresponding to ijk either.

But suppose we add a third factor, Education, represented by the indicator $X^{(4)}$, having the value 1 if the programmer has a least a Master's degree, 0 otherwise. Then m would take on 12 values, and the full model would have 12 parameters:

$$m_{Y;X}(i,j,k,l) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 l + \beta_5 ij + \beta_6 ik + \beta_7 il + \beta_8 jl + \beta_9 kl + \beta_{10} ijl + \beta_{11} ikl \tag{7.58}$$

Again, there would be no ijkl term, as jk = 0.

Here β_1, β_2, β_3 and β_4 are called the **main effects**, as opposed to the coefficients of the interaction terms, called of course the **interaction effects**.

The no-interaction version would be

$$m_{Y;X}(i,j,k,l) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 l \tag{7.59}$$

7.3.14.3 Now Consider Parsimony

In the three-factor example above, we have 12 groups and 12 means. Why not just treat it that way, instead of applying the powerful tool of regression analysis? The answer lies in our desire for parsimony, as noted in Section 7.3.9.1.

If for example (7.59) were to hold, at least approximately, we would have a far more satisfying model. We could for instance then talk of "the" effect of using an IDE, rather than qualifying such a statement by stating what the effect would be for each different language and education level. Moreover, if our sample size is not very large, we would get more accurate estimates of the various subpopulation means.

Or it could be that, while (7.59) doesn't hold, a model with only two-way interactions,

$$m_{Y;X}(i,j,k,l) = \beta_0 + \beta_1 i + \beta_2 j + \beta_3 k + \beta_4 l + \beta_5 ij + \beta_6 ik + \beta_7 il + \beta_8 jl + \beta_9 kl \tag{7.60}$$

does work well. This would not be as nice as (7.59), but it still would be more parsimonious than (7.58).

Accordingly, the major thrust of ANOVA is to decide how rich a model is needed to do a good job of describing the situation under study. There is an implied hierarchy of models of interest here:

- the full model, including two- and three-way interactions, (7.58)
- the model with two-factor interactions only, (7.60)
- the no-interaction model, (7.59)

Traditionally these are determined via hypothesis testing, which involves certain partitionings of sums of squares similar to (7.18). (This is where the name *analysis of variance* stems from.) The null distribution of the test statistic often turns out to be an F-distribution. Of course, in this book, we consider hypothesis testing inappropriate, preferring to give some careful thought to the estimated parameters, but it is standard. Further testing can be done on individual β_1 and so on. Often people use simultaneous inference procedures, discussed briefly in Section 5.3 of our unit on estimation and testing, since many tests are performed.

7.3.14.4 Reparameterization

Classical ANOVA uses a somewhat different parameterization than that we've considered here. For instance, consider a single-factor setting (called **one-way ANOVA**) with three levels. Our predictors are then $X^{(1)}$ and $X^{(2)}$. Taking our approach here, we would write

$$m_{Y;X}(i,j) = \beta_0 + \beta_1 i + \beta_2 j \tag{7.61}$$

The traditional formulation would be

$$\mu_i = \mu + \alpha_i, \ i = 1,2,3 \tag{7.62}$$

where

$$\mu = \frac{\mu_1 + \mu_2 + \mu_3}{3} \tag{7.63}$$

and

$$\alpha_i = \mu_i - \mu \tag{7.64}$$

Of course, the two formulations are equivalent. It is left to the reader to check that, for instance,

$$\mu = \beta_0 + \frac{\beta_1 + \beta_2}{2} \tag{7.65}$$

There are similar formulations for ANOVA designs with more than one factor.

Note that the classical formulation overparameterizes the problem. In the one-way example above, for instance, there are four parameters (μ, α_1, α_2, α_3) but only three groups. This would make the system indeterminate, but we add the constraint

$$\sum_{i=1}^{3} \alpha_i = 0 \tag{7.66}$$

Equation (7.25) then must make use of **generalized matrix inverses**.

7.4 The Classification Problem

As mentioned earlier, in the special case in which Y is an indicator variable, with the value 1 if the object is in a class and 0 if not, the regression problem is called the **classification problem**. It is also sometimes called **pattern recognition**, in which case the predictors are called **features**. Also, the term **machine learning** usually refers to classification problems.

If there are c classes, we need c (or c-1) Y variables, which I will denote by $Y^{(i)}$, i = 1,...,c.

7.4.1 Meaning of the Regression Function

7.4.1.1 The Mean Here Is a Probability

Now, here is a key point: Since the mean of any indicator random variable is the probability that the variable is equal to 1, the regression function in classification problems reduces to

$$m_{Y;X}(t) = P(Y = 1|X = t) \tag{7.67}$$

(Remember that X and t are vector-valued.)

For concreteness, let's look at the patent example in Section 7.1. Again, Y will be 1 or 0, depending on whether the patent had public funding. We'll take $X^{(1)}$ to be an indicator variable for the presence or absence of "NSF" in the patent, $X^{(2)}$ to be an indicator variable for "NIH," and take $X^{(3)}$ to be the number of claims in the patent. This last predictor might be relevant, e.g. if industrial patents are lengthier.

So, $m_{Y;X}[(1, 0, 5)]$ would be the population proportion of all patents that are publicly funded, among those that contain the word "NSF," do not contain "NIH," and make five claims.

7.4.1.2 Optimality of the Regression Function

Again, our context is that we want to guess Y, knowing X. Since Y is 0-1 valued, our guess for Y based on X, g(X), should be 0-1 valued too. What is the best g?

Again, since Y and g are 0-1 valued, our criterion should be what will I call Probability of Correct Classification (PCC):

$$\text{PCC} = P[Y = g(X)] \tag{7.68}$$

Now proceed as in (7.13):

$$\text{PCC} = E\left[P\{Y = g(X)|X\}\right] \tag{7.69}$$

The analog of Lemma 11 is

Lemma 12 *Suppose W takes on values in the set A = $\{$0,1$\}$, and consider the problem of maximizing*

$$P(W = c),\ c \epsilon A \tag{7.70}$$

The solution is

$$\begin{cases} 1, & \textit{if P(W = 1)} > \textit{0.5} \\ 0, & \textit{otherwise} \end{cases} \tag{7.71}$$

Proof

Again recalling that c is either 1 or 0, we have

$$\begin{aligned} P(W = c) &= P(W = 1)c + \left[1 - P(W = 1)\right](1 - c) && (7.72) \\ &= \left[2P(W = 1) - 1\right]c + 1 - P(W = 1) && (7.73) \end{aligned}$$

The result follows.

■

Applying this to (7.69), we see that the best g is given by

$$g(t) = \begin{cases} 1, & \text{if } m_{Y;X}(t) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{7.74}$$

So we find that the regression function is again optimal, in this new context.

7.4.2 Parametric Models for the Regression Function in Classification Problems

Remember, we often try a parametric model for our regression function first, as it means we are estimating a finite number of quantities, instead of an infinite number.

7.4.2.1 The Logistic Model: Form

The most common parametric model in the classification problem is the logistic model (often called the *logit* model), seen in Section 7.3.10. In its r-predictor form, it is

$$m_{Y;X}(t) = P(Y = 1|X = t) = \frac{1}{1 + e^{-(\beta_0+\beta_1 t_1+...+\beta_r t_r)}} \tag{7.75}$$

For instance, consider the patent example. Under the logistic model, the population proportion of all patents that are publicly funded, among those that contain the word “NSF,” do not contain “NIH,” and make five claims would have the value

$$\frac{1}{1 + e^{-(\beta_0+\beta_1+5\beta_3)}} \tag{7.76}$$

7.4.2.2 The Logistic Model: Intuitive Motivation

The logistic function itself,

$$\frac{1}{1 + e^{-u}} \tag{7.77}$$

has values between 0 and 1, and is thus a candidate for modeling a probability. Also, it is monotonic in u, making it further attractive, as in many classification problems we believe that $m_{Y;X}(t)$ should be monotonic in the predictor variables.

7.4.2.3 The Logistic Model: Theoretical Foundation

But there are much stronger reasons to use the logit model, as it includes many common parametric models for X. To see this, note that we can write, for vector-valued discrete X and t,

$$P(Y=1|X=t) = \frac{P(Y=1 \text{ and } X=t)}{P(X=t)} \tag{7.78}$$

$$= \frac{P(Y=1)P(X=t|Y=1)}{P(X=t)} \tag{7.79}$$

$$= \frac{P(Y=1)P(X=t|Y=1)}{P(Y=1)P(X=t|Y=1)+P(Y=0)P(X=t|Y=0)} \tag{7.80}$$

$$= \frac{1}{1+\frac{(1-q)P(X=t|Y=0)}{qP(X=t|Y=1)}} \tag{7.81}$$

where $q = P(Y = 1)$ is the proportion of members of the population which have $Y = 1$. (Keep in mind that this probability is unconditional!!!! In the patent example, for instance, if say $q = 0.12$, then 12% of all patents in the patent population—without regard to words used, numbers of claims, etc.—are publicly funded.)

If X is a continuous random vector, then the analog of (7.81) is

$$P(Y=1|X=t) = \frac{1}{1+\frac{(1-q)f_{X|Y=0}(t)}{qf_{X|Y=1}(t)}} \tag{7.82}$$

Now suppose X, given Y, has a normal distribution. In other words, within each class, Y is normally distributed. Consider the case of just one predictor variable, i.e. r = 1. Suppose that given $Y = i$, X has the distribution $N(\mu_i, \sigma^2)$, i = 0,1. Then

$$f_{X|Y=i}(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-0.5\left(\frac{t-\mu_i}{\sigma}\right)^2\right] \tag{7.83}$$

After doing some elementary but rather tedious algebra, (7.82) reduces to the logistic form

$$\frac{1}{1+e^{-(\beta_0+\beta_1 t)}} \tag{7.84}$$

where β_0 and β_1 are functions of μ_0, μ_0 and σ.

In other words, if X is normally distributed in both classes, with the same variance but different means, then $m_{Y;X}$ has the logistic form! And the same is true if X is multivariate normal in each class, with different mean vectors but equal covariance matrices. (The algebra is even more tedious here, but it does work out.)

So, not only does the logistic model have an intuitively appealing form, it is also implied by one of the most famous distributions X can have within each class—the multivariate normal.

If you reread the derivation above, you will see that the logit model will hold for any within-class distributions for which

$$\ln \left(\frac{f_{X|Y=0}(t)}{f_{X|Y=1}(t)} \right) \tag{7.85}$$

(or its discrete analog) is linear in t. Well guess what—this condition is true for exponential distributions too! Work it out for yourself.

In fact, a number of famous distributions imply the logit model.

7.4.3 Nonparametric Estimation of Regression Functions for Classification (advanced topic)

7.4.3.1 Use the Kernel Method, CART, Etc.

Since the classification problem is a special case of the general regression problem, nonparametric regression methods can be used here too.

7.4.3.2 SVMs

There are also some methods which have been developed exclusively, or mainly, for classification. One of them which has been getting a lot of publicity in computer science circles is **support vector machines** (SVMs). To explain the SVM concept, consider the case r = 2, i.e. two predictor variables $X^{(1)}$ and $X^{(2)}$. What an SVM would do is use our sample data to draw a curve in the $X^{(1)}$-$X^{(2)}$ plane, with our classification rule then being, "Guess Y to be 1 if X is on one side of the curve, and guess it to be 0 if X is on the other side."

Beware! There are no "magic" solutions to statistical problems, especially prediction problems, and the statements one sees in some computer science research to the effect that SVMs are generally superior to other prediction methods are unfounded. SVMs do very well in some situations, BUT not so well in others. I highly recommend the site `www.dtreg.com/benchmarks.htm`, which compares six different types of classification function estimators—including logistic regression and SVM—on several dozen real data sets. The overall percent misclassification rates, averaged over all the data sets, was fairly close, ranging from a high of 25.3% to a low of 19.2%. The much-vaunted SVM came in at 20.3%. That's nice, but it was only a tad better than logit's 20.9%. Considering that the latter has a big advantage in that one gets an actual equation for the classification function, complete with parameters which we can estimate and make confidence intervals for, it is not clear just what role SVM and the other nonparametric estimators should play, in general, though in specific applications they may be appropriate.

7.4.4 Variable Selection in Classification Problems

7.4.4.1 Problems Inherited from the Regression Context

In Section 7.3.9.2, it was pointed out that the problem of predictor variable selection in regression is unsolved. Since the classification problem is a special case of regression, there is no surefire way to select predictor variables there either.

7.4.4.2 Example: Forest Cover Data

And again, using hypothesis testing to choose predictors is not the answer. To illustrate this, let's look again at the forest cover data we saw in Section 4.2.12.

There were seven classes of forest cover there. Let's restrict attention to classes 1 and 2. In my R analysis I had the class 1 and 2 data in objects **cov1** and **cov2**, respectively. I combined them,

```
> cov1and2 <- rbind(cov1,cov2)
```

and created a new variable to serve as Y:

```
cov1and2[,56] <- ifelse(cov1and2[,55] == 1,1,0)
```

Let's see how well we can predict a site's class from the variable HS12 (hillside shade at noon) that we investigated in that past unit, using a logistic model.

In R we fit logistic models via the **glm()** function, for generalized linear models. The word *generalized* here refers to models in which some function of $m_{Y;X}(t)$ is linear in parameters β_i. For the classification model,

$$\ln\left(m_{Y;X}(t)/[1-m_{Y;X}(t)]\right) = \beta_0 + \beta_1 t^{(1)} + ... + \beta_r t^{(r)} \tag{7.86}$$

This kind of generalized linear model is specified in R by setting the named argument **family** to **binomial**. Here is the call:

```
> g <- glm(cov1and2[,56] ~ cov1and2[,8],family=binomial)
```

The result was:

```
> summary(g)

Call:
glm(formula = cov1and2[, 56] ~ cov1and2[, 8], family = binomial)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.165  -0.820  -0.775   1.504   1.741

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.515820   1.148665   1.320   0.1870
cov1and2[, 8] -0.010960   0.005103  -2.148   0.0317 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 959.72  on 810  degrees of freedom
Residual deviance: 955.14  on 809  degrees of freedom
AIC: 959.14

Number of Fisher Scoring iterations: 4
```

So, $\widehat{\beta}_1 = -0.01$. This is tiny, as can be seen from our data in the last unit. There we found that the estimated mean values of HS12 for cover types 1 and 2 were 223.8 and 226.3, a difference of only 2.5. That difference in essence gets multiplied by 0.01. More concretely, in (7.52), plug in our estimates 1.52 and -0.01 from our R output above, first taking t to be 223.8 and then 226.3. The results are 0.328 and 0.322, respectively. In other words, HS12 isn't having much effect on the probability of cover type 1, and so it cannot be a good predictor of cover type.

Yet the R output says that β_1 is "significantly" different from 0, with a p-value of 0.03. Thus, we see once again that hypothesis testing does not achieve our goal. Again, cross validation is a better method for choosing predictors.

7.4.5 Y Must Have a Marginal Distribution!

In our material here, we have tacitly assumed that the vector (Y,X) has a distribution. That may seem like an odd and puzzling remark to make here, but **it is absolutely crucial**. Let's see what it means.

Consider the study on object-oriented programming in Section 7.1, but turned around. (This example will be somewhat contrived, but it will illustrate the principle.) Suppose we know how many lines of code are in a project, which we will still call $X^{(1)}$, and we know how long it took to complete, which we will now take as $X^{(2)}$, and from this we want to guess whether object-oriented or procedural programming was used (without being able to look at the code, of course), which is now our new Y.

Here is our huge problem: Given our sample data, there is no way to estimate q in (7.81). That's because the authors of the study simply took two groups of programmers and had one group use object-oriented programming and had the other group use procedural programming. If we had sampled programmers at random from actual projects done at this company, that would enable us to estimate q, the population proportion of projects done with OOP. But we can't do that with the data that we do have. Indeed, in this setting, it may not even make sense to speak of q in the first place.

Mathematically speaking, if you think about the process under which the data was collected in this study, there does exist some conditional distribution of X given Y, but Y itself has no distribution. So, we can NOT estimate P(Y=1|X). About the best we can do is try to guess Y on the basis of whichever value of i makes $f_{X|Y=i}(X)$ larger.

7.5 Principal Components Analysis

7.5.1 Dimension Reduction and the Principle of Parsimony

Consider a random vector $X = (X_1, X_2)^T$. Suppose the two components of X are highly correlated with each other. Then for some constants c and d,

$$X_2 \approx c + dX_1 \tag{7.87}$$

Then in a sense there is really just one random variable here, as the second is nearly equal to some linear combination of the first. The second provides us with almost no new information, once we have the first.

In other words, even though the vector X roams in two-dimensional space, it usually sticks close to a one-dimensional object, namely the line (7.87). We saw a graph illustrating this in our unit on multivariate distributions, page 97.

In general, consider a k-component random vector

$$X = (X_1, ..., X_k)^T \tag{7.88}$$

We again wish to investigate whether just a few, say w, of the X_i tell almost the whole story, i.e. whether most X_j can be expressed approximately as linear combinations of these few X_i. In other words, even though X is k-dimensional, it tends to stick close to some w-dimensional subspace.

Note that although (7.87) is phrased in prediction terms, we are not (or more accurately, not necessarily) interested in prediction here. We have not designated one of the $X^{(i)}$ to be a response variable and the rest to be predictors.

Once again, the Principle of Parsimony is key. If we have, say, 20 or 30 variables, it would be nice if we could reduce that to, for example, three or four. This may be easier to understand and work with, albeit with the complication that our new variables would be linear combinations of the old ones.

7.5.2 How to Calculate Them

Here's how it works. The theory of linear algebra says that since Σ is a symmetric matrix, it is diagonalizable, i.e. there is a real matrix Q for which

$$Q^T \Sigma Q = D \tag{7.89}$$

where D is a diagonal matrix. (This is a special case of **singular value decomposition**.) The columns C_i of Q are the eigenvectors of Σ, and it turns out that they are orthogonal to each other, i.e. their dot product is 0.

Let

$$W_i = C_i^T X, \ i = 1, ..., k \tag{7.90}$$

so that the W_i are scalar random variables, and set

$$W = (W_1, ..., W_k)^T \tag{7.91}$$

Then

$$W = Q^T X \tag{7.92}$$

Now, use the material on covariance matrices from our unit on multivariate analysis, page 86,

$$Cov(W) = Cov(Q^T X) = Q^T Cov(X) Q = D \ \text{(from (7.89))} \tag{7.93}$$

Note too that if X has a multivariate normal distribution (which we are not assuming), then W does too.

Let's recap:

- We have created new random variables W_i as linear combinations of our original X_j.
- The W_i are uncorrelated. Thus if in addition X has a multivariate normal distribution, so that W does too, then the W_i will be independent.
- The variance of W_i is given by the ith diagonal element of D.

The W_i are called the **principal components** of the distribution of X.

It is customary to relabel the W_i so that W_1 has the largest variance, W_2 has the second-largest, and so on. We then choose those W_i that have the larger variances, and discard the others, because the latter, having small variances, are close to constant and thus carry no information.

All this will become clearer in the example below.

7.5.3 Example: Forest Cover Data

Let's try using principal component analysis on the forest cover data set we've looked at before. There are 10 continuous variables (also many discrete ones, but there is another tool for that case, the log-linear model, discussed in Section 7.6).

In my R run, the data set (. not restricted to just two forest cover types, but consisting only of the first 1000 observations) was in the object **f**. Here are the call and the results:

```
> prc <- prcomp(f[,1:10])
> summary(prc)
Importance of components:
                            PC1      PC2      PC3      PC4      PC5 PC6
Standard deviation     1812.394 1613.287 1.89e+02 1.10e+02 96.93455 30.16789
Proportion of Variance    0.552    0.438 6.01e-03 2.04e-03  0.00158 0.00015
Cumulative Proportion     0.552    0.990 9.96e-01 9.98e-01  0.99968 0.99984
                            PC7      PC8 PC9  PC10
Standard deviation     25.95478 16.78595 4.2 0.783
Proportion of Variance  0.00011  0.00005 0.0 0.000
Cumulative Proportion   0.99995  1.00000 1.0 1.000
```

You can see from the variance values here that R has scaled the W_i so that their variances sum to 1.0. (It has not done so for the standard deviations, which are for the nonscaled variables.) This is fine, as we are only interested in the variances relative to each other, i.e. saving the principal components with the larger variances.

What we see here is that eight of the 10 principal components have very small variances, i.e. are close to constant. In other words, though we have 10 variables $X_1, ..., X_{10}$, there is really only two variables' worth of information carried in them.

So for example if we wish to predict forest cover type from these 10 variables, we should only use two of them. We could use W_1 and W_2, but for the sake of interpretability we stick to the original X vector; we can use any two of the X_i.

The coefficients of the linear combinations which produce W from X, i.e. the Q matrix, are available via **prc$rotation**.

7.6 Log-Linear Models

Here we discuss a procedure which is something of an analog of principal components for discrete variables. Our material on ANOVA will also come into play. It is recommended that the reader review Sections 7.3.14 and 7.5 before continuing.

7.6.1 The Setting

Let's consider a variation on the software engineering example in Sections 7.2 and 7.3.14. Assume we have the factors, IDE, Language and Education. Our change—**of extreme importance**—is that we will now assume that these factors are **RANDOM**. What does this mean?

In the original example described in Section 7.2, programmers were *assigned* to languages, and in our extensions of that example, we continued to assume this. Thus for example the number of programmers who use an IDE and program in Java was fixed; if we repeated the experiment, that number would stay the same. If we were sampling from some programmer population, our new sample would have new programmers, but the number using and IDE and Java would be the same as before, as our study procedure specifies this.

By contrast, let's now assume that we simply sample programmers at random, and ask them whether they prefer to use an IDE or not, and which language they prefer.[10] Then for example the number of programmers who prefer to use an IDE and program in Java will be random, not fixed; if we repeat the experiment, we will get a different count.

Suppose now we now wish to investigate relations between the factors. Are choice of platform and language related to education, for instance?

7.6.2 The Data

Denote our three factors by $X^{(s)}$, s = 1,2,3. Here $X^{(1)}$, IDE, will take on the values 1 and 2 instead of 1 and 0 as before, 1 meaning that the programmer prefers to use an IDE, and 2 meaning not so. $X^{(3)}$ changes this way too, and $X^{(2)}$ will take on the values 1 for C++, 2 for Java and 3 for C. Note that we no longer use indicator variables.

Let $X_r^{(s)}$ denote the value of $X^{(s)}$ for the rth programmer in our sample, r = 1,2,...,n. Our data are the counts

$$N_{ijk} = \text{number of r such that } X_r^{(1)} = i, X_r^{(2)} = j \text{ and } X_r^{(3)} = k \tag{7.94}$$

For instance, if we sample 100 programmers, our data might look like this:

[10] Other sampling schemes are possible too.

```
prefers to use IDE:

                   Bachelor's or less     Master's or more
          C++                      18                   15
         Java                      22                   10
            C                       6                    4


prefers not to use IDE:

                   Bachelor's or less     Master's or more
          C++                       7                    4
         Java                       6                    2
            C                       3                    3
```

So for example $N_{122} = 10$ and $N_{212} = 4$.

Here we have a three-dimensional **contingency table**. Each N_{ijk} value is a **cell** in the table.

7.6.3 The Models

Let p_{ijk} be the population probability of a randomly-chosen programmer falling into cell ijk, i.e.

$$p_{ijk} = P\left(X^{(1)} = i \text{ and } X^{(2)} = j \text{ and } X^{(3)} = k\right) = E(N_{ijk})/n \tag{7.95}$$

As mentioned, we are interested in relations between the factors, in the form of independence, full and partial. Consider first the case of full independence:

$$\begin{aligned} p_{ijk} &= P\left(X^{(1)} = i \text{ and } X^{(2)} = j \text{ and } X^{(3)} = k\right) & (7.96) \\ &= P\left(X^{(1)} = i\right) \cdot P\left(X^{(2)} = j\right) \cdot P\left(X^{(3)} = k\right) & (7.97) \end{aligned}$$

Taking logs of both sides in (7.96), we see that independence of the three factors is equivalent to saying

$$\log(p_{ijk}) = a_i + b_j + c_k \tag{7.98}$$

for some numbers a_i, b_j and c_j. The numbers must be nonpositive, and since

$$\sum_m P(X^{(s)} = m) = 1 \tag{7.99}$$

we must have, for instance,

$$\sum_{g=1}^{2} \exp(c_g) = 1 \tag{7.100}$$

The point is that (7.98) looks like our no-interaction ANOVA models, e.g. (7.57). On the other hand, if we assume instead that Education is independent of IDE and Language but that IDE and Language are not independent of each other, our model would be

$$\begin{aligned} \log(p_{ijk}) &= P\left(X^{(1)} = i \text{ and } X^{(2)} = j\right) \cdot P\left(X^{(3)} = k\right) & (7.101) \\ &= a_i + b_j + d_{ij} + c_k & (7.102) \end{aligned}$$

Here we have written $P\left(X^{(1)} = i \text{ and } X^{(2)} = j\right)$ as a sum of "main effects" a_i and b_j, and "interaction effects," d_{ij}, analogous to ANOVA.

Another possible model would have IDE and Language conditionally independent, given Education, meaning that at any level of education, a programmer's preference to use IDE or not, and his choice of programming language, are not related. We'd write the model this way:

$$\begin{aligned} \log(p_{ijk}) &= P\left(X^{(1)} = i \text{ and } X^{(2)} = j\right) \cdot P\left(X^{(3)} = k\right) & (7.103) \\ &= a_i + b_j + f_{ik} + h_{jk} + c_k & (7.104) \end{aligned}$$

Note carefully that the type of independence in (7.104) has a quite different interpretation than that in (7.102).

The full model, with no independence assumptions at all, would have three two-way interaction terms, as well as a three-way interaction term.

7.6.4 Parameter Estimation

Remember, whenever we have parametric models, the statistician's "Swiss army knife" is maximum likelihood estimation. That is what is most often used in the case of log-linear models.

How, then, do we compute the likelihood of our data, the N_{ijk}? It's actually quite straightforward, because the N_{ijk} have the multinomial distribution we studied in Section 3.6.1.1 of our unit on multivariate distributions.

$$L = \frac{n!}{\Pi_{i,j,k} N_{ijk}!} p_{ijk}^{N_{ijk}} \tag{7.105}$$

We then write the p_{ijk} in terms of our model parameters. Take for example (7.102), where we write

$$p_{ijk} = e^{a_i + b_j + d_{ij} + c_k} \tag{7.106}$$

We then substitute (7.106) in (7.105), and maximize the latter with respect to the a_i, b_j, d_{ij} and c_k, subject to constraints such as (7.100).

The maximization may be messy. But certain cases have been worked out in closed form, and in any case today one would typically do the computation by computer. In R, for example, there is the **loglin()** function for this purpose.

7.6.5 The Goal: Parsimony Again

Again, we'd like "the simplest model possible, but not simpler." This means a model with as much independence between factors as possible, subject to the model being accurate.

Classical log-linear model procedures do model selection by hypothesis testing, testing whether various interaction terms are 0. The tests often parallel ANOVA testing, with chi-square distributions arising instead of F-distributions.

7.7 Simpson's (Non-)Paradox

Suppose each individual in a population either possesses or does not possess traits *A*, *B* and *C*, and that we wish to predict trait *A*. Let $\bar{A}$, $\bar{B}$ and $\bar{C}$ denote the situations in which the individual does not possess the given trait. Simpson's Paradox then describes a situation in which

$$P(A|B) > P(A|\bar{B}) \tag{7.107}$$

and yet

$$P(A|B,C) < P(A|\bar{B},C) \tag{7.108}$$

In other words, the possession of trait B seems to have a positive predictive power for A by itself, but when in addition trait C is held constant, the relation between B and A turns negative.

An example is given by Fabris and Freitas,[11] concerning a classic study of tuberculosis mortality in 1910. Here the attribute A is mortality, B is city (Richmond, with $\bar{B}$ being New York), and C is race (African-American, with $\bar{C}$ being Caucasian). In probability terms, the data show that (these of course are sample estimates)

- P(mortality | Richmond) = 0.0022
- P(mortality | New York) = 0.0019
- P(mortality | Richmond, black) = 0.0033
- P(mortality | New York, black) = 0.0056
- P(mortality | Richmond, white) = 0.0016
- P(mortality | New York, white) = 0.0018

The data also show that

- P(black | Richmond) = 0.37
- P(black | New York) = 0.002

a point which will become relevant below.

At first, New York looks like it did a better job than Richmond. However, once one accounts for race, we find that New York is actually worse than Richmond. Why the reversal? The answer stems from the fact that racial inequities being what they were at the time, blacks with the disease fared much worse than whites. Richmond's population was 37% black, proportionally far more than New York's 0.2%. So, Richmond's heavy concentration of blacks made its overall mortality rate look worse than New York's, even though things were actually much worse in New York.

But is this really a "paradox"? Closer consideration of this example reveals that the only reason this example (and others like it) is surprising is that the predictors were used in the wrong order. One normally looks for predictors one at a time, first finding the best single predictor, then the best pair of predictors, and so on. If this were done on the above data set, the first predictor variable chosen would be race, not city. In other words, the sequence of analysis would look something like this:

[11] C.C. Fabris and A.A. Freitas. Discovering Surprising Patterns by Detecting Occurrences of Simpson's Paradox. In *Research and Development in Intelligent Systems XVI (Proc. ES99, The 19th SGES Int. Conf. on Knowledge-Based Systems and Applied Artificial Intelligence)*, 148-160. Springer-Verlag, 1999

- P(mortality | Richmond) = 0.0022
- P(mortality | New York) = 0.0019
- P(mortality | black) = 0.0048
- P(mortality | white) = 0.0018
- P(mortality | black, Richmond) = 0.0033
- P(mortality | black, New York) = 0.0056
- P(mortality | white, Richmond) = 0.0016
- P(mortality | white, New York) = 0.0018

The analyst would have seen that race is a better predictor than city, and thus would have chosen race as the best single predictor. The analyst would then investigate the race/city predictor pair, and would never reach a point in which city alone were in the selected predictor set. Thus no anomalies would arise.

Exercises

Note to instructor: See the Preface for a list of sources of real data on which exercises can be assigned to complement the theoretical exercises below.

1. Suppose we are interested in documents of a certain type, which we'll call Type 1. Everything that is not Type 1 we'll call Type 2, with a proportion q of all documents being Type 1. Our goal will be to try to guess document type by the presence of absence of a certain word; we will guess Type 1 if the word is present, and otherwise will guess Type 2.

Let T denote document type, and let W denote the event that the word is in the document. Also, let p_i be the proportion of documents that contain the word, among all documents of Type i, i = 1,2. The event C will denote our guessing correctly.

Find the overall probability of correct classification, $P(C)$, and also $P(C|W)$.

Hint: Be careful of your conditional and unconditional probabilities here.

2. In the quartic model in ALOHA simulation example, find an approximate 95% confidence interval for the true population mean wait if our backoff parameter b is set to 0.6.

Hint: You will need to use the fact that a linear combination of the components of a multivariate normal random vector has a univariate normal distributions as discussed in Section 3.6.2.1.

3. Consider the linear regression model with one predictor, i.e. r = 1. Let Y_i and X_i represent the values of the response and predictor variables for the ith observation in our sample.

(a) Assume as in Section 7.3.7.4 that $Var(Y|X = t)$ is a constant in t, σ^2. Find the exact value of $Cov(\hat{\beta}_0, \hat{\beta}_1)$, as a function of the X_i and σ^2. Your final answer should be in scalar, i.e. non-matrix form.

(b) Suppose we wish to fit the model $m_{Y;X}(t) = \beta_1 t$, i.e. the usual linear model but without the constant term, β_0. Derive a formula for the least-squares estimate of β_1.

4. Suppose the random pair (X, Y) has density $8st$ on $0 < t < s < 1$. Find $m_{Y;X}(s)$ and $Var(Y|X = t)$, $0 < s < 1$.

5. We showed that (7.82) reduces to the logistic model in the case in which the distribution of X given Y is normal. Show that this is also true in the case in which that distribution is exponential, i.e.

$$f_{X|Y}(t, i) = \lambda_i e^{-\lambda_i t}, \ t > 0 \tag{7.109}$$

6. The code below reads in a file, **data.txt**, with the header record

```
"age", "weight", "systolic blood pressure", "height"
```

and then does the regression analysis.

Suppose we wish to estimate β in the model

$$\text{mean weight} = \beta_0 + \beta_1 \text{height} + \beta_2 \text{age}$$

Fill in the blanks in the code:

```
dt <- ____________(____________________________________)
regr <- lm(_____________________________________________)
cvmat <- _______________(regr)
print("the estimated value of beta2-beta0 is",
   ____________________________________________________)
print("the estimated variance of beta2 - beta0 is",
   _______________________ %*% cvmat %*%  _________________________)
# calculate the matrix Q
q <- cbind(_________________________________________)
```

7. In this problem, you will conduct an R simulation experiment similar to that of Foster and Stine on overfitting, discussed in Section 7.3.9.2.

Generate data $X_i^{(j)}, \ i = 1, ..., n, \ j = 1, ..., r$ from a N(0,1) distribution, and $\epsilon_i, \ i = 1, ..., n$ from N(0,4). Set $Y_i = X_i^{(1)} + \epsilon_i, \ i = 1, ..., n$. This simulates drawing a random sample of n observations from an (r+1)-variate population.

Now suppose the analyst, unaware that Y is related to only $X^{(1)}$, fits the model

$$m_{Y;X^{(1)},...,X^{(r)}}(t_1, ..., t_r) = \beta_0 + \beta_1 t^{(1)} + ... + \beta_r t^{(r)} \tag{7.110}$$

In actuality, $\beta_j = 0$ for $j > 1$ (and for $i = 0$). But the analyst wouldn't know this. Suppose the analyst selects predictors by testing the hypotheses $H_0 : \beta_i = 0$, as in Section 7.3.9.2, with $\alpha = 0.05$.

Do this for various values of r and n. You should find that, for fixed n and increasing r. You begin to find that some of the predictors are declared to be "significantly" related to Y (complete with asterisks) when in fact they are not (while $X^{(1)}$, which really is related to Y, may be declared NOT "significant." This illustrates the folly of using hypothesis testing to do variable selection.

8. Suppose given X = t, the distribution of Y has mean γt and variance σ^2, for all t in (0,1). This is a fixed-X regression setting, i.e. X is nonrandom: For each i = 1,...,n we observe Yi drawn at random from the distribution of Y given X = i/n. The quantities γ and σ^2 are unknown.

Our goal is to estimate $m_{Y;X}(0.75)$. We have two choices for our estimator:

- We can estimate in the usual least-squares manner, denoting our estimate by G, and then use as our estimator $T_1 = 0.75G$.
- We can take our estimator T_2 to be $(Y_1 + ... + Y_n)/n$,

Perform a tradeoff analysis similar to that of Section 4.6.7, determining under what conditions T_1 is superior to T_2 and vice versa. Our criterion is mean squared error (MSE), $E[(T_i - m_{Y;X}(0.75)^]$. Make your expressions as closed-form as possible.

Advice: This is a linear model, albeit one without an intercept term. The quantity G here is simply $\hat{\sigma}$. G will turn out to be a linear combination of the Xs (which are constants), so its variance is easy to find.

9. Suppose X has an $N(\mu, \mu^2)$ distribution, i.e. with the standard deviation equal to the mean. (A common assumption in regression contexts.) Show that $h(X) = \ln(X)$ will be a variance-stabilizing transformation, a concept discussed in Section 5.2.2.

10. Consider a random pair (X, Y) for which the linear model $E(Y|X) = \beta_0 + \beta_1 X$ holds, and think about predicting Y, first without X and then with X, minimizing mean squared prediction error (MSPE) in each case. From Section 7.3.6, we know that without X, the best predictor is EY, while with X it is $E(Y|X)$, which under our assumption here is $\beta_0 + \beta_1 X$. Show that the reduction in MSPE accrued by using X, i.e.

$$\frac{E\left[(Y - EY)^2\right] - E\left[\{Y - E(Y|X)\}^2\right]}{E\left[(Y - EY)^2\right]} \tag{7.111}$$

is equal to $\rho^2(X, Y)$.

Chapter 8

Markov Chains

One of the most famous stochastic models is that of a Markov chain. This type of model is widely used in computer science, biology, physics and so on.

8.1 Discrete-Time Markov Chains

8.1.1 Example: Finite Random Walk

One of the most commonly used stochastic models is that of a **Markov chain**. To motivate this discussion, let us start with a simple example: Consider a **random walk** on the set of integers between 1 and 5, moving randomly through that set, say one move per second, according to the following scheme. If we are currently at position i, then one time period later we will be at either i-1, i or i+1, according to the outcome of rolling a fair die—we move to i-1 if the die comes up 1 or 2, stay at i if the die comes up 3 or 4, and move to i+1 in the case of a 5 or 6. For the special cases i = 1 and i = 5, we simply move back to 2 or 4, respectively. (In random walk terminology, these are called **reflecting barriers**.)

The integers 1 through 5 form the **state space** for this process; if we are currently at 4, for instance, we say we are in state 4. Let X_t represent the position of the particle at time t, t = 0, 1,2,. . . .

The random walk is a **Markov process**. The process is "memoryless," meaning that we can "forget the past"; given the present and the past, the future depends only on the present:

$$P(X_{t+1} = s_{t+1}|X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) = P(X_{t+1} = s_{t+1}|X_t = s_t) \tag{8.1}$$

The term *Markov process* is the general one. If the state space is discrete, i.e. countably infinite, then we usually use the more specialized term, *Markov chain*.

Although this equation has a very complex look, it has a very simple meaning: The distribution of our next position, given our current position and all our past positions, is dependent only on the current position.[1] It is clear that the random walk process above does have this property; for instance, if we are now at position 4, the probability that our next state will be 3 is 1/3—no matter where we were in the past.

Continuing this example, let p_{ij} denote the probability of going from position i to position j in one step. For example, $p_{21} = p_{23} = \frac{1}{3}$ while $p_{24} = 0$ (we can reach position 4 from position 2 in two steps, but not in one step). The numbers p_{ij} are called the **one-step transition probabilities** of the process. Denote by P the matrix whose entries are the p_{ij}:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{8.2}$$

By the way, it turns out that the matrix P^k gives the k-step transition probabilities. In other words, the element (i,j) of this matrix gives the probability of going from i to j in k steps.

8.1.2 Long-Run Distribution

In typical applications we are interested in the long-run distribution of the process, for example the long-run proportion of the time that we are at position 4. For each state i, define

$$\pi_i = \lim_{t \to \infty} \frac{N_{it}}{t} \tag{8.3}$$

where N_{it} is the number of visits the process makes to state i among times 1, 2,..., t. In most practical cases, this proportion will exist and be independent of our initial position X_0. The π_i are called the **steady-state probabilities**, or the **stationary distribution** of the Markov chain.

Intuitively, the existence of π_i implies that as t approaches infinity, the system approaches steady-state, in the sense that

$$\lim_{t \to \infty} P(X_t = i) = \pi_i \tag{8.4}$$

Actually, the limit (8.4) may not exist in some cases. We'll return to that point later, but for typical cases it does exist, and we will usually assume this.

[1] This can be generalized, so that the future depends on the present and also on the state one unit of time ago, etc. However, such models become quite unwieldy.

8.1.2.1 Derivation of the Balance Equations

Equation (8.4) suggests a way to calculate the values π_i, as follows.

First note that

$$P(X_{t+1} = i) = \sum_k P(X_t = k \text{ and } X_{t+1} = i) = \sum_k P(X_t = k) P(X_{t+1} = i | X_t = k) = \sum_k P(X_t = k) p_{ki} \tag{8.5}$$

where the sum goes over all states k. For example, in our random walk example above, we would have

$$P(X_{t+1} = 3) = \sum_{k=1}^{5} P(X_t = k \text{ and } X_{t+1} = 3) = \sum_{k=1}^{5} P(X_t = k) P(X_{t+1} = 3 | X_t = k) = \sum_{k=1}^{5} P(X_t = k) p_{k3} \tag{8.6}$$

Then as $t \rightarrow \infty$ in Equation (8.5), intuitively we would have

$$\pi_i = \sum_k \pi_k p_{ki} \tag{8.7}$$

Remember, here we know the p_{ki} and want to find the π_i. Solving these equations (one for each i), called the **balance equations**, give us the π_i.

For the random walk problem above, for instance, the solution is $\pi = (\frac{1}{11}, \frac{3}{11}, \frac{3}{11}, \frac{3}{11}, \frac{1}{11})$. Thus in the long run we will spend 1/11 of our time at position 1, 3/11 of our time at position 2, and so on.

8.1.2.2 Solving the Balance Equations

A matrix formulation is also useful. Letting π denote the row vector of the elements π_i, i.e. $\pi = (\pi_1, \pi_2, ...)$, these equations (one for each i) then have the matrix form

$$\pi = \pi P \tag{8.8}$$

or

$$(I - P')\pi = 0 \tag{8.9}$$

where as usual ’ denotes matrix transpose.

Note that there is also the constraint

$$\sum_i \pi_i = 1 \tag{8.10}$$

One of the equations in the system is redundant. We thus eliminate one of them, say by removing the last row of I-P in (8.9). To reflect This can be used to calculate the π_i. It turns out that one of the equations in the system is redundant. We thus eliminate one of them, say by removing the last row of I-P in (8.9).

To reflect (8.10), which in matrix form is

$$1_n' \pi = 1 \tag{8.11}$$

where 1_n is a column vector of all 1s, we replace the removed row in I-P by a row of all 1s, and in the right-hand side of (8.9) we replace the last 0 by a 1. We can then solve the system.

All this can be done with R’s **solve()** function:

```
findpi1 <- function(p) {
   n <- nrow(p)
   imp <- diag(n) - t(p)  # I-P
   imp[n,] <- rep(1,n)
   rhs <- c(rep(0,n-1),1)
   pivec <- solve(imp,rhs)
   return(pivec)
}
```

Or one can note from (8.8) that π is a left eigenvector of P with eigenvalue 1, so one can use R’s **eigen()** function. It can be proven that if P is irreducible and aperiodic (defined later in this chapter), every eigenvalue other than 1 is smaller than 1 (so we can speak of *the* eigenvalue 1), and the eigenvector corresponding to 1 has all components real.

Since π is a left eigenvector, the argument in the call must be P’ rather than P. In addition, since an eigenvector is only unique up to scalar multiplication, we must deal with the fact that the return value of **eigen()** may have negative components, and will likely not satisfy (8.10). Here is the code:

```
findpi2 <- function(p) {
   n <- nrow(p)
   # find first eigenvector of P'
   pivec <- eigen(t(p))$vectors[,1]
   # guaranteed to be real, but could be negative
   if (pivec[1] < 0) pivec <- -pivec
   # normalize
   pivec <- pivec / sum(pivec)
   return(pivec)
}
```

But Equation (8.9) may not be easy to solve. For instance, if the state space is infinite, then this matrix equation represents infinitely many scalar equations. In such cases, you may need to try to find some clever trick which will allow you to solve the system, or in many cases a clever trick to analyze the process in some way other than explicit solution of the system of equations.

And even for finite state spaces, the matrix may be extremely large. In some cases, you may need to resort to numerical methods.

8.1.2.3 Periodic Chains

Note again that even if Equation (8.9) has a solution, this does not imply that (8.4) holds. For instance, suppose we alter the random walk example above so that

$$p_{i,i-1} = p_{i,i+1} = \frac{1}{2} \tag{8.12}$$

for i = 2, 3, 4, with transitions out of states 1 and 5 remaining as before. In this case, the solution to Equation (8.9) is $(\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8})$. This solution is still valid, in the sense that Equation (8.3) will hold. For example, we will spend 1/4 of our time at Position 4 in the long run. But the limit of $P(X_i = 4)$ will not be 1/4, and in fact the limit will not even exist. If say X_0 is even, then X_i can be even only for even values of i. We say that this Markov chain is **periodic** with period 2, meaning that returns to a given state can only occur after amounts of time which are multiples of 2.

8.1.2.4 The Meaning of the Term "Stationary Distribution"

Though we have informally defined the term *stationary distribution* in terms of long-run proportions, the technical definition is this:

Definition 13 *Consider a Markov chain. Suppose we have a vector π of nonnegative numbers that sum to 1. Let X_0 have the distribution π. If that results in X_1 having that distribution too (and thus also all X_n), we say that π is the* **stationary distribution** *of this Markov chain.*

Note that this definition stems from (8.5).

In our (first) random walk example above, this would mean that if we have X_0 distributed on the integers 1 through 5 with probabilities $(\frac{1}{11}, \frac{3}{11}, \frac{3}{11}, \frac{3}{11}, \frac{1}{11})$, then for example $P(X_1 = 1) = \frac{1}{11}$, $P(X_1 = 4) = \frac{3}{11}$ etc. This is indeed the case, as you can verify using (8.5) with t = 0.

In our "notebook" view, here is what we would do. Imagine that we generate a random integer between 1 and 5 according to the probabilities $(\frac{1}{11}, \frac{3}{11}, \frac{3}{11}, \frac{3}{11}, \frac{1}{11})$,[2] and set X_0 to that number. We would then generate another random number, by rolling an ordinary die, and going left, right or staying put, with probability 1/3 each. We would then write down X_1 and X_2 on the first line of our notebook. We would then do this experiment again, recording the results on the second line, then again and again. In the long run, 3/11 of the lines would have, for instance, $X_0 = 4$, and 3/11 of the lines would have $X_1 = 4$. In other words, X_1 would have the same distribution as X_0.

8.1.3 Example: Stuck-At 0 Fault

8.1.3.1 Description

In the above example, the labels for the states consisted of single integers i. In some other examples, convenient labels may be r-tuples, for example 2-tuples (i,j).

Consider a serial communication line. Let $B_1, B_2, B_3, ...$ denote the sequence of bits transmitted on this line. It is reasonable to assume the B_i to be independent, and that $P(B_i = 0)$ and $P(B_i = 1)$ are both equal to 0.5.

Suppose that the receiver will eventually fail, with the type of failure being **stuck at 0**, meaning that after failure it will report all future received bits to be 0, regardless of their true value. Once failed, the receiver stays failed, and should be replaced. Eventually the new receiver will also fail, and we will replace it; we continue this process indefinitely.

Let ρ denote the probability that the receiver fails on any given bit, with independence between bits in terms of receiver failure. Then the lifetime of the receiver, that is, the time to failure, is geometrically distributed with "success" probability ρ i.e. the probability of failing on receipt of the i-th bit after the receiver is installed is $(1-\rho)^{i-1}\rho$ for i = 1,2,3,...

However, the problem is that we will not know whether a receiver has failed (unless we test it once in a while, which we are not including in this example). If the receiver reports a long string of 0s, we should suspect that the receiver has failed, but of course we cannot be sure that it has; it is still possible that the message being transmitted just happened to contain a long string of 0s.

Suppose we adopt the policy that, if we receive k consecutive 0s, we will replace the receiver with a new unit. Here k is a design parameter; what value should we choose for it? If we use a very small value, then we will incur great expense, due to the fact that we will be replacing receiver units at an unnecessarily high rate. On the other hand, if we make k too large, then we will often wait too long to replace the receiver, and the resulting error rate in received bits will be sizable. Resolution of this tradeoff between expense and accuracy depends on the relative importance of the two. (There are also other possibilities, involving the

[2] Say by rolling an 11-sided die.

addition of redundant bits for error detection, such as parity bits. For simplicity, we will not consider such refinements here. However, the analysis of more complex systems would be similar to the one below.)

8.1.3.2 Initial Analysis

A natural state space in this example would be

$$\{(i,j) : i = 0, 1, ..., k-1; j = 0, 1; i + j \neq 0\} \tag{8.13}$$

where i represents the number of consecutive 0s that we have received so far, and j represents the state of the receiver (0 for failed, 1 for nonfailed). Note that when we are in a state of the form (k-1,j), if we receive a 0 on the next bit (whether it is a true 0 or the receiver has failed), our new state will be (0,1), as we will install a new receiver. Note too that there is no state (0,0), since if the receiver is down it must have received at least one bit.

The calculation of the transition matrix P is straightforward, though it requires careful thought. For example, suppose the current state is (2,1), and that we are investigating the expense and bit accuracy corresponding to a policy having k = 5. What can happen upon receipt of the next bit? The next bit will have a true value of either 0 or 1, with probability 0.5 each. The receiver will change from working to failed status with probability ρ. Thus our next state could be:

- (3,1), if a 0 arrives, and the receiver does not fail;
- (0,1), if a 1 arrives, and the receiver does not fail; or
- (3,0), if the receiver fails

The probabilities of these three transitions out of state (2,1) are:

$$\begin{aligned} p_{(2,1),(3,1)} &= 0.5(1-\rho) && (8.14)\\ p_{(2,1),(0,1)} &= 0.5(1-\rho) && (8.15)\\ p_{(2,1),(3,0)} &= \rho && (8.16)\end{aligned}$$

Other entries of the matrix P can be computed similarly. Note by the way that from state (4,1) we will go to (0,1), no matter what happens.

Formally specifying the matrix P using the 2-tuple notation as above would be very cumbersome. In this case, it would be much easier to map to a one-dimensional labeling. For example, if k = 5, the nine states

(1,0),...,(4,0),(0,1),(1,1),...,(4,1) could be renamed states 1,2,...,9. Then we could form P under this labeling, and the transition probabilities above would appear as

$$
\begin{aligned}
p_{78} &= 0.5(1-\rho) && (8.17)\\
p_{75} &= 0.5(1-\rho) && (8.18)\\
p_{73} &= \rho && (8.19)
\end{aligned}
$$

8.1.3.3 Going Beyond Finding π

Finding the π_i should be just the first step. We then want to use them to calculate various quantities of interest.[3] For instance, in this example, it would also be useful to find the error rate ϵ, and the mean time (i.e., the mean number of bit receptions) between receiver replacements, μ. We can find both ϵ and μ in terms of the π_i, in the following manner.

The quantity ϵ is the proportion of the time during which the true value of the received bit is 1 but the receiver is down, which is 0.5 times the proportion of the time spent in states of the form (i,0):

$$\epsilon = 0.5(\pi_1 + \pi_2 + \pi_3 + \pi_4) \qquad (8.20)$$

This should be clear intuitively, but it would also be instructive to present a more formal derivation of the same thing. Let E_n be the event that the n-th bit is received in error, with D_n denoting the event that the receiver is down. Then

$$
\begin{aligned}
\epsilon &= \lim_{n\to\infty} P(E_n) && (8.21)\\
&= \lim_{n\to\infty} P(X_n = 1 \text{ and } D_n) && (8.22)\\
&= \lim_{n\to\infty} P(X_n = 1)P(D_n) && (8.23)\\
&= 0.5(\pi_1 + \pi_2 + \pi_3 + \pi_4) && (8.24)
\end{aligned}
$$

Here we used the fact that X_n and the receiver state are independent.

Equations (8.21) follow a pattern we'll use repeatedly in this chapter. In subsequent examples we will not show the steps with the limits, but the limits are indeed there. Make sure to mentally go through these steps yourself.[4]

[3] Note that unlike a classroom setting, where those quantities would be listed for the students to calculate, in research we must decide on our own which quantities are of interest.

[4] The other way to work this out rigorously is to assume that X_0 has the distribution π, as in Section 8.1.2.4. Then no limits are needed in (8.21. But this may be more difficult to understand.

Now to get μ in terms of the π_i note that since μ is the long-run average number of bits between receiver replacements, it is then the reciprocal of η, the long-run fraction of bits that result in replacements. For example, say we replace the receiver on average every 20 bits. Over a period of 1000 bits, then (speaking on an intuitive level) that would mean about 50 replacements. Thus approximately 0.05 (50 out of 1000) of all bits results in replacements.

$$\mu = \frac{1}{\eta} \tag{8.25}$$

Again suppose k = 5. A replacement will occur only from states of the form (4,j), and even then only under the condition that the next reported bit is a 0. In other words, there are three possible ways in which replacement can occur:

(a) We are in state (4,0). Here, since the receiver has failed, the next reported bit will definitely be a 0, regardless of that bit's true value. We will then have a total of k = 5 consecutive received 0s, and therefore will replace the receiver.

(b) We are in the state (4,1), and the next bit to arrive is a true 0. It then will be reported as a 0, our fifth consecutive 0, and we will replace the receiver, as in (a).

(c) We are in the state (4,1), and the next bit to arrive is a true 1, but the receiver fails at that time, resulting in the reported value being a 0. Again we have five consecutive reported 0s, so we replace the receiver.

Therefore,

$$\eta = \pi_4 + \pi_9(0.5 + 0.5\rho) \tag{8.26}$$

Again, make sure you work through the full version of (8.26), using the pattern in (8.21).

Thus

$$\mu = \frac{1}{\eta} = \frac{1}{\pi_4 + 0.5\pi_9(1+\rho)} \tag{8.27}$$

This kind of analysis could be used as the core of a cost-benefit tradeoff investigation to determine a good value of k. (Note that the π_i are functions of k, and that the above equations for the case k = 5 must be modified for other values of k.)

8.1.4 Example: Shared-Memory Multiprocessor

(Adapted from *Probabiility and Statistics, with Reliability, Queuing and Computer Science Applicatiions*, by K.S. Trivedi, Prentice-Hall, 1982 and 2002, but similar to many models in the research literature.)

8.1.4.1 The Model

Consider a shared-memory multiprocessor system with m memory modules and m CPUs. The address space is partitioned into m chunks, based on either the most-significant or least-significant $\log_2 m$ bits in the address.[5]

The CPUs will need to access the memory modules in some random way, depending on the programs they are running. To make this idea concrete, consider the Intel assembly language instruction

```
add %eax, (%ebx)
```

which adds the contents of the EAX register to the word in memory pointed to by the EBX register. Execution of that instruction will (absent cache and other similar effects, as we will assume here and below) involve two accesses to memory—one to fetch the old value of the word pointed to by EBX, and another to store the new value. Moreover, the instruction itself must be fetched from memory. So, altogether the processing of this instruction involves three memory accesses.

Since different programs are made up of different instructions, use different register values and so on, the sequence of addresses in memory that are generated by CPUs are modeled as random variables. In our model here, the CPUs are assumed to act independently of each other, and successive requests from a given CPU are independent of each other too. A CPU will choose the ith module with probability q_i. A memory request takes one unit of time to process, though the wait may be longer due to queuing. In this very simplistic model, as soon as a CPU's memory request is fulfilled, it generates another one. On the other hand, while a CPU has one memory request pending, it does not generate another.

Let's assume a crossbar interconnect, which means there are m^2 separate paths from CPUs to memory modules, so that if the m CPUs have memory requests to m different memory modules, then all the requests can be fulfilled simultaneously. Also, assume as an approximation that we can ignore communication delays.

How good are these assumptions? One weakness, for instance, is that many instructions, for example, do not use memory at all, except for the instruction fetch, and as mentioned, even the latter may be suppressed due to cache effects.

Another example of potential problems with the assumptions involves the fact that many programs will have code like

[5] You may recognize this as high-order and low-order interleaving, respectively.

```
for (i = 0; i < 10000; i++) sum += x[i];
```

Since the elements of the array x will be stored in consecutive addresses, successive memory requests from the CPU while executing this code will not be independent. The assumption would be more justified if we were including cache effects, or (noticed by Earl Barr) if we are studying a timesharing system with a small quantum size.

Thus, many models of systems like this have been quite complex, in order to capture the effects of various things like caching, nonindependence and so on in the model. Nevertheless, one can often get some insight from even very simple models too. In any case, for our purposes here it is best to stick to simple models, so as to understand more easily.

Our state will be an m-tuple $(N_1, ..., N_m)$, where N_i is the number of requests currently pending at memory module i. Recalling our assumption that a CPU generates another memory request immediately after the previous one is fulfilled, we always have that $N_1 + ... + N_m = m$.

It is straightforward to find the transition probabilities p_{ij}. Here are a couple of examples, with m = 2:

- $p_{(2,0),(1,1)}$: Recall that state (2,0) means that currently there are two requests pending at Module 1, one being served and one in the queue, and no requests at Module 2. For the transition $(2, 0) \rightarrow (1, 1)$ to occur, when the request being served at Module 1 is done, it will make a new request, this time for Module 2. This will occur with probability q_2. Meanwhile, the request which had been queued at Module 1 will now start service. So, $p_{(2,0),(1,1)} = q_2$.
- $p_{(1,1),(1,1)}$: In state (1,1), both pending requests will finish in this cycle. To go to (1,1) again, that would mean that the two CPUs request different modules from each other—CPUs 1 and 2 choose Modules 1 and 2 or 2 and 1. Each of those two possibilities has probability q_1q_2, so $p_{(1,1),(1,1)} = 2q_1q_2$.

We then solve for the π, using (8.7). It turns out, for example, that

$$\pi_{(1,1)} = \frac{q_1q_2}{1 - 2q_1q_2} \tag{8.28}$$

8.1.4.2 Going Beyond Finding π

Let B denote the number of memory requests in a given memory cycle. Then we may be interested in E(B), the number of requests completed per unit time, i.e. per cycle. We can find E(B) as follows. Let S denote the current state. Then, continuing the case m = 2, we have from the Law of Total Expectation,[6]

[6] Actually, we could take a more direct route in this case, noting that B can only take on the values 1 and 2. Then $EB = P(B = 1) + 2P(B = 2) = \pi_{(2,0} + \pi_{s(0,2)} + 2\pi_{(1,1)}$. But the analysis below extends better to the case of general m.

$$
\begin{aligned}
E(B) &= E[E(B|S)] && (8.29)\\
&= P(S=(2,0))E(B|S=(2,0)) + P(S=(1,1))E(B|S=(1,1)) + P(S=(0,2))E(B|S=(0,2)) && (8.30)\\
&= \pi_{(2,0)}E(B|S=(2,0)) + \pi_{(1,1)}E(B|S=(1,1)) + \pi_{(0,2)}E(B|S=(0,2)) && (8.31)
\end{aligned}
$$

All this equation is doing is finding the overall mean of B by breaking down into the cases for the different states.

Now if we are in state (2,0), only one request will be completed this cycle, and B will be 1. Thus $E(B|S = (2,0)) = 1$. Similarly, $E(B|S = (1,1)) = 2$ and so on. After doing all the algebra, we find that

$$EB = \frac{1 - q_1 q_2}{1 - 2q_1 q_2} \tag{8.32}$$

The maximum value of E(B) occurs when $q_1 = q_2 = \frac{1}{2}$, in which case E(B)=1.5. This is a lot less than the maximum capacity of the memory system, which is m = 2 requests per cycle.

So, we can learn a lot even from this simple model, in this case learning that there may be a substantial underutilization of the system. This is a common theme in probabilistic modeling: Simple models may be worthwhile in terms of insight provided, even if their numerical predictions may not be too accurate.

8.1.5 Example: Slotted ALOHA

Recall the slotted ALOHA model from Chapter 1:

- Time is divided into slots or epochs.
- There are n nodes, each of which is either idle or has a **single** message transmission pending. So, a node doesn't generate a new message until the old one is successfully transmitted (a very unrealistic assumption, but we're keeping things simple here).
- In the middle of each time slot, each of the idle nodes generates a message with probability q.
- Just before the end of each time slot, each active node attempts to send its message with probability p.
- If more than one node attempts to send within a given time slot, there is a **collision**, and each of the transmissions involved will fail.
- So, we include a **backoff** mechanism: At the middle of each time slot, each node with a message will with probability q attempt to send the message, with the transmission time occupying the remainder of the slot.

So, q is a design parameter, which must be chosen carefully. If q is too large, we will have too mnay collisions, thus increasing the average time to send a message. If q is too small, a node will often refrain from sending even if no other node is there to collide with.

Define our state for any given time slot to be the number of nodes currently having a message to send at the very beginning of the time slot (before new messages are generated). Then for $0 < i < n$ and $0 < j < n - i$ (there will be a few special boundary cases to consider too), we have

$$p_{i,i-1} = \underbrace{(1-q)^{n-i}}_{\text{no new msgs}} \cdot \underbrace{i(1-p)^{i-1}p}_{\text{one xmit}} \tag{8.33}$$

$$p_{ii} = \underbrace{(1-q)^{n-i} \cdot [1 - i(1-p)^{i-1}p]}_{\text{no new msgs and no succ xmits}} + \underbrace{(n-i)(1-q)^{n-i-1}q \cdot (i+1)(1-p)^{i}p}_{\text{one new msg and one xmit}} \tag{8.34}$$

$$\begin{aligned} p_{i,i+j} \quad &= \quad \underbrace{\binom{n-i}{j} q^j (1-q)^{n-i-j} \cdot [1 - (i+j)(1-p)^{i+j-1}p]}_{\text{j new msgs and no succ xmit}} \\ &+ \quad \underbrace{\binom{n-i}{j+1} q^{j+1} (1-q)^{n-i-j-1} \cdot (i+j+1)(1-p)^{i+j}p}_{\text{j+1 new msgs and succ xmit}} \end{aligned} \tag{8.35}$$

Note that in (8.34) and (8.35), we must take into account the fact that a node with a newly-created messages might try to send it. In (8.35), for instance, in the first term we have j new messages, on top of the i we already had, so i+j messages might try to send. The probability that there is no successful transmission is then $1 - (i + j)(1 - p)^{i+j-1}p$.

The matrix P is then quite complex. We always hope to find a closed-form solution, but that is unlikely in this case. Solving it on a computer is easy, though, say by using the **solve()** function in the R statistical language.

8.1.5.1 Going Beyond Finding π

Once again various interesting quantities can be derived as functions of the π, such as the system throughput τ, i.e. the number of successful transmissions in the network per unit time. Here's how to get τ:

First, suppose for concreteness that in steady-state the probability of there being a successful transmission in a given slot is 20%. Then after, say, 100,000 slots, about 20,000 will have successful transmissions—a throughput of 0.2. So, the long-run probability of successful transmission is the same as the long-run fraction of slots in which there are successful transmissions! That in turn can be broken down in terms of the various states:

$$\begin{aligned}\tau &= P(\text{success xmit}) \qquad (8.36)\\ &= \sum_s P(\text{success xmit} \mid \text{in state s}) P(\text{in state s})\end{aligned}$$

Now, to calculate $P(\text{success xmit} \mid \text{in state s})$, recall that in state s we start the slot with s nonidle nodes, but that we may acquire some new ones; each of the n-s idle nodes will create a new message, with probability q. So,

$$P(\text{success xmit} \mid \text{in state s}) = \sum_{j=0}^{n-s} \binom{n-s}{j} q^j (1-q)^{n-s-j} \cdot (s+j)(1-p)^{s+j-1} p \qquad (8.37)$$

Substituting into (8.36), we have

$$\tau = \sum_{s=0}^{n} \sum_{j=0}^{n-s} \binom{n-s}{j} q^j (1-q)^{n-s-j} \cdot (s+j)(1-p)^{s+j-1} p \cdot \pi_s \qquad (8.38)$$

With some more subtle reasoning, one can derive the mean time a message waits before being successfully transmitted, as follows:

Focus attention on one particular node, say Node 0. It will repeatedly cycle through idle and busy periods, I and B. We wish to find E(B). I has a geometric distribution with parameter q,[7] so

$$E(I) = \frac{1}{q} \qquad (8.39)$$

Then if we can find E(I+B), we will get E(B) by subtraction.

To find E(I+B), note that there is a one-to-one correspondence between I+B cycles and successful transmissions; each I+B period ends with a successful transmission at Node 0. Imagine again observing this node for, say, 100000 time slots, and say E(I+B) is 2000. That would mean we'd have about 50 cycles, thus 50 successful transmissions from this node. In other words, the throughput would be approximately 50/100000

[7] If a message is sent in the same slot in which it is created, we will count B as 1. If it is sent in the following slot, B = 2, etc. B will have a modified geometric distribution starting at 0 instead of 1, but we will ignore this here for the sake of simplicity.

= 0.02 = 1/E(I+B). So, a fraction

$$\frac{1}{E(I+B)} \tag{8.40}$$

of the time slots have successful transmissions from this node.

But that quantity is the throughput for this node (number of successful transmissions per unit time), and due to the symmetry of the system, that throughput is 1/n of the total throughput of the n nodes in the network, which we denoted above by τ.

So,

$$E(I+B) = \frac{n}{\tau} \tag{8.41}$$

Thus from (8.39) we have

$$E(B) = \frac{n}{\tau} - \frac{1}{q} \tag{8.42}$$

where of course τ is the function of the π_i in (8.36).

Now let's find the proportion of attempted transmissions which are successful. This will be

$$\frac{E(\text{number of successful transmissions in a slot})}{E(\text{number of attempted transmissions in a slot})} \tag{8.43}$$

(To see why this is the case, again think of watching the network for 100,000 slots.) Then the proportion of successful transmissions during that period of time is the number of successful transmissions divided by the number of attempted transmissions. Those two numbers are approximately the numerator and denominator of 8.43.

Now, how do we evaluate (8.43)? Well, the numerator is easy, since it is τ, which we found before. The denominator will be

$$\sum_s \pi_s[sp + (n-s)pq] \tag{8.44}$$

The factor sp+spq comes from the following reasoning. If we are in state s, the s nodes which already have something to send will each transmit with probability p, so there will be an expected number sp of them that try to send. Also, of the n-s which are idle at the beginning of the slot, an expected sq of them will generate new messages, and of those sq, and estimated sqp will try to send.

8.2 Hidden Markov Models

The word *hidden* in the term *Hidden Markov Model* (HMM) refers to the fact that the state of the process is hidden, i.e. unobservable.

Actually, we've already seen an example of this, back in Section 8.1.3. There the state, actually just part of it, was unobservable, namely the status of the receiver being up or down. But here we are not trying to guess X_n from Y_n (see below), so it probably would not be considered an HMM. HMMs.

An HMM consists of a Markov chain X_n which is unobservable, together with observable values Y_n. The X_n are governed by the transition probabilities p_{ij}, and the Y_n are generated from the X_n according to

$$r_{km} = P(Y_n = m|X_n = k) \tag{8.45}$$

Typically the idea is to guess the X_n from the Y_n and our knowledge of the p_{ij} and r_{km}. The details are too complex to give here, but you can at least understand that Bayes' Rule comes into play.

A good example of HMMs would be in text mining applications. Here the Y_n might be words in the text, and X_n would be their parts of speech (POS)—nouns, verbs, adjectives and so on. Consider the word *round*, for instance. Your first thought might be that it is an adjective, but it could be a noun (e.g. an elimination round in a tournament) or a verb (e.g. to round off a number or round a corner). The HMM would help us to guess which, and therefore guess the true meaning of the word.

HMMs are also used in speech process, DNA modeling and many other applications.

8.3 Continuous-Time Markov Chains

In the Markov chains we analyzed above, events occur only at integer times. However, many Markov chain models are of the **continuous-time** type, in which events can occur at any times. Here the **holding time**, i.e. the time the system spends in one state before changing to another state, is a continuous random variable.

The state of a Markov chain at any time now has a continuous subscript. Instead of the chain consisting of the random variables $X_n, \ n = 1, 2, 3, ...$ (you can also start n at 0 in the sense of Section 8.1.2.4), it now consists of $\{X_t : t \in [0, \infty)\}$. The Markov property is now

$$P(X_{t+u} = k|X_s \text{ for all } 0s \leq t) = P(X_{t+u} = k|X_t) \text{ for all } t, u \geq 0 \tag{8.46}$$

8.3.1 Holding-Time Distribution

In order for the Markov property to hold, the distribution of holding time at a given state needs to be "memoryless." You may recall that exponentially distributed random variables have this property. In other words, if a random variable W has density

$$f(t) = \lambda e^{-\lambda t} \tag{8.47}$$

for some λ then

$$P(W > r + s | W > r) = P(W > s) \tag{8.48}$$

for all positive r and s. Actually, one can show that exponential distributions are the only continuous distributions which have this property. Therefore, *holding times in Markov chains must be exponentially distributed.*

It is difficult for the beginning modeler to fully appreciate the memoryless property. You are urged to read the material on exponential distributions in Section 2.3.4.1 before continuing.

Because it is central to the Markov property, the exponential distribution is assumed for all basic activities in Markov models. In queuing models, for instance, both the interarrival time and service time are assumed to be exponentially distributed (though of course with different values of λ). In reliability modeling, the lifetime of a component is assumed to have an exponential distribution.

Such assumptions have in many cases been verified empirically. If you go to a bank, for example, and record data on when customers arrive at the door, you will find the exponential model to work well (though you may have to restrict yourself to a given time of day, to account for nonrandom effects such as heavy traffic at the noon hour). In a study of time to failure for airplane air conditioners, the distribution was also found to be well fitted by an exponential density. On the other hand, in many cases the distribution is not close to exponential, and purely Markovian models cannot be used for anything more than a rough approximation.

8.3.2 The Notion of "Rates"

A key point is that the parameter λ in (8.47) has the interpretation of a rate, in the sense we will now discuss. First, recall that $1/\lambda$ is the mean. Say light bulb lifetimes have an exponential distribution with mean 100 hours, so $\lambda = 0.01$. In our lamp, whenever its bulb burns out, we immediately replace it with a new on. Imagine watching this lamp for, say, 100,000 hours. During that time, we will have done approximately 100000/100 = 1000 replacements. That would be using 1000 light bulbs in 100000 hours, so we are using bulbs at the rate of 0.01 bulb per hour. For a general λ, we would use light bulbs at the rate of λ bulbs per hour. This concept is crucial to what follows.

8.3.3 Stationary Distribution

We again define π_i to be the long-run proportion of time the system is in state i, and we again will derive a system of linear equations to solve for these proportions.

8.3.3.1 Derivation

To this end, let λ_i denote the parameter in the holding-time distribution at state i, and define the following:

- $U(i,t)$ is the total time spent at state i up through time t
- $N(i,t)$ is the number of visits to state i up through time t
- H_{ij} is the holding time during the jth visit to state i

The reason $U(i,t)$ is of interest to us is that

$$\lim_{t \rightarrow \infty} \frac{U(i,t)}{t} = \pi_i \tag{8.49}$$

Next, write

$$U(i,t) = H_{i1} + H_{i2} + ... + H_{i,N(i,t)} + \text{small error} \tag{8.50}$$

by definition.

The reason for the "small error" is that at time t, we may be currently at state i, in a visit that has not yet finished. At $t \rightarrow \infty$, this term vanishes, so we'll ignore it.

Now in taking the expected value in (8.50), we need to deal with the fact that there is a random number of terms in the sum on the right-hand side. This we do using the Theorem of Total Expectation, as seen in the example in Section 3.8.1.3, yielding

$$E[U(i,t)] = \frac{1}{\lambda_i} \cdot E[N(i,t)] \tag{8.51}$$

since holding times at state i have mean $1/\lambda_i$. Then

$$\lim_{t \rightarrow \infty} \frac{U(i,t)}{t} = \frac{1}{\lambda_i} \lim_{t \rightarrow \infty} \frac{N(i,t)}{t} \tag{8.52}$$

(Both U and N are essentially cumulative sums, so dividing by t sets up something like the Strong Law of Large Numbers, discussed in Section 1.4.7.)

And then for large t

$$U(i,t) \approx \frac{1}{\lambda_i} \cdot N(i,t) \tag{8.53}$$

The next point is to look at the rates of transitions into and out of state i. These should be equal in the long run, and that will be the basis for our balance equations.

The number of transitions out of i up through time t (except for the "small error") is equal to $N(i,t)$. What about inbound transitions? Let p_{ji} be the probability that, when a holding time at state j ends, our transition is to i. Then for large t, the number of transitions from state j to state i is approximately $N(j,t)p_{ji}$. Equating the two, we have, again for large t

$$\sum_{j \neq i} N(j,t)p_{ji} \approx N(i,t) \tag{8.54}$$

Combining (8.53) and (8.54), we have

$$U(i,t)\lambda_i \approx N(i,t) \approx \sum_{j \neq i} N(j,t)p_{ji} \approx \sum_{j \neq i} U(j,t)\lambda_j p_{ji} \tag{8.55}$$

Dividing by t and taking limits, we have

$$\pi_i \lambda_i = \sum_{j \neq i} \pi_j \lambda_j p_{ji} \tag{8.56}$$

So, *voila!*, there are our balance equations (one for each i).

8.3.3.2 Quicker (But Less Clear?) Derivation

We will sometimes refer to quantities

$$\rho_{rs} = \lambda_r p_{rs} \tag{8.57}$$

with the following interpretation. In the context of the ideas in our example of the rate of light bulb replacements in Section 8.3.2, one can view (8.57) as the rate of transitions from r to s, *during the time we are in state r.* Equation (8.56) can then be interpreted as equating the rate of transitions into i and the rate out of i.

8.3.3.3 Computation

Motivated by (8.56), define the matrix Q by

$$q_{ij} = \begin{cases} \lambda_j p_{ji}, & \text{if } i \neq j \\ -\lambda_i, & \text{if } i = j \end{cases} \tag{8.58}$$

Q is called the **infinitesimal generator** of the system, so named because it is the basis of the system of differential equations that can be used to find the finite-time probabilistic behavior of X_t.

The name also reflects the rates notion we've been discussing, due to the fact that, say in our light bulb example in Section 8.3.2,

$$P(\text{bulb fails in next h time}) = \lambda h + o(h) \tag{8.59}$$

Then (8.56) is stated in matrix form as

$$Q'\pi = 0 \tag{8.60}$$

Here is R code to solve the system:

```
findpicontin <- function(q) {
   n <- nrow(q)
   newq <- t(q)
   newq[n,] <- rep(1,n)
   rhs <- c(rep(0,n-1),1)
   pivec <- solve(newq,rhs)
   return(pivec)
}
```

8.3.4 Minima of Independent Exponentially Distributed Random Variables

In setting up Equations (8.56) are fine, there will be an issue with finding the p_{ji}. The material in this section will be used for that purpose in later sections.

Theorem 14 *Suppose $W_1, ..., W_k$ are independent random variables, with W_i being exponentially distributed with parameter λ_i. Let $Z = \min(W_1, ..., W_k)$. Then*

(a) Z is exponentially distributed with parameter $\lambda_1 + ... + \lambda_k$

(b) $P(Z = W_i) = \frac{\lambda_i}{\lambda_1 + ... + \lambda_k}$

The sum $\lambda_1 + ... + \lambda_n$ in (a) should make good intuitive sense to you, for the following reasons. 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

$$F_Z(t) = 1 - P(Z > t) = 1 - P(W_1 > t \text{ and ... and } W_k > t) = 1 - \Pi_i\, e^{-\lambda_i t} = 1 - e^{-(\lambda_1 + ... + \lambda_n)t} \quad (8.61)$$

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

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

$$P(Z = W_1) = P(W_1 < W_2) = \int_0^\infty \int_t^\infty \lambda_1 e^{-\lambda_1 t} \lambda_2 e^{-\lambda_2 s}\, ds\, dt = \frac{\lambda_1}{\lambda_1 + \lambda_2} \quad (8.62)$$

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, $P(W_1 = W_2) = 0$.

8.3.5 Example: Machine Repair

Suppose the operations in a factory require the use of a certain kind of machine. The manager has installed two of these machines. This is known as a **gracefully degrading system**: When both machines are working, the fact that there are two of them, instead of one, leads to a shorter wait time for access to a machine. When one machine has failed, the wait is longer, but at least the factory operations may continue. Of course, if both machines fail, the factory must shut down until at least one machine is repaired.

Suppose the time until failure of a single machine, carrying the full load of the factory, has an exponential distribution with mean 20.0, but the mean is 25.0 when the other machine is working, since it is not so loaded. Repair time is exponentially distributed with mean 8.0.

We can take as our state space {0,1,2}, where the state is the number of working machines. Now, let us find the parameters λ_i and p_{ji} for this system. For example, what about λ_2? The holding time in state 2 is the minimum of the two lifetimes of the machines, and thus from the results of Section 8.3.4, has parameter $\frac{1}{25.0} + \frac{1}{25.0} = 0.08$.

For λ_1, a transition out of state 1 will be either to state 2 (the down machine is repaired) or to state 0 (the up machine fails). The time until transition will be the minimum of the lifetime of the up machine and the repair time of the down machine, and thus will have parameter $\frac{1}{20.0} + \frac{1}{8.0} = 0.175$. Similarly, $\lambda_0 = \frac{1}{8.0} + \frac{1}{8.0} = 0.25$.

It is important to understand how the Markov property is being used here. Suppose we are in state 1, and the down machine is repaired, sending us into state 2. Remember, the machine which had already been up has "lived" for some time now. But the memoryless property of the exponential distribution implies that this machine is now "born again."

What about the parameters p_{ji}? Well, p_{21} is certainly easy to find; since the transition 2 → 1 is the *only* transition possible out of state 2, $p_{21} = 1$.

For p_{12}, recall that transitions out of state 1 are to states 0 and 2, with rates 1/20.0 and 1/8.0, respectively. So,

$$p_{12} = \frac{1/8.0}{1/20.0 + 1/8.0} = 0.72 \tag{8.63}$$

Working in this manner, we finally arrive at the complete system of equations (8.56):

$$\begin{aligned} \pi_2(0.08) &= \pi_1(0.125) & (8.64)\\ \pi_1(0.175) &= \pi_2(0.08) + \pi_0(0.25) & (8.65)\\ \pi_0(0.25) &= \pi_1(0.05) & (8.66) \end{aligned}$$

Of course, we also have the constraint $\pi_2 + \pi_1 + \pi_0 = 1$. The solution turns out to be

$$\pi = (0.072, 0.362, 0.566) \tag{8.67}$$

Thus for example, during 7.2% of the time, there will be no machine available at all.

Several variations of this problem could be analyzed. We could compare the two-machine system with a one-machine version. It turns out that the proportion of down time (i.e. time when no machine is available) increases to 28.6%. Or we could analyze the case in which only one repair person is employed by this factory, so that only one machine can be repaired at a time, compared to the situation above, in which we (tacitly) assumed that if both machines are down, they can be repaired in parallel. We leave these variations as exercises for the reader.

8.3.6 Example: Migration in a Social Network

The following is a simplified version of research in online social networks.

There is a town with two social groups. Everyone is in exactly one group People arrive from outside town, with exponentially distributed interarrival times at rate α, and join one of the groups with probability 0.5 each. Each person will occasionally switch groups, with one possible "switch" being to leave town entirely. A person's time before switching is exponentially distributed with rate σ; the switch will either be to the other group or to the outside world, with probabilities q and 1-q, respectively. Let the state of the system be (i,j), where i and j are the number of current members in groups 1 and 2, respectively.

Let's find a typical balance equation, say for the state (8,8):

$$\pi_{(8,8)}(\alpha + 2 \cdot 8 \cdot \sigma) = (\pi_{(9,8)} + \pi_{(8,9)}) \cdot \sigma(1-q) + (\pi_{(9,7)} + \pi_{(7,9)}) \cdot \sigma q + (\pi_{(8,7)} + \pi_{(7,8)}) \cdot 0.5\alpha \quad (8.68)$$

The reasoning is straightforward. How can we move out of state (8,8)? Well, there could be an arrival (rate α), or any one of the 16 people could switch groups (rate 16σ), etc.

Now, in a "going beyond finding the ı" vein, let's find the long-run fraction of transfers into group 1 that come from group 2, as opposed to from the outside.

The rate of transitions into that group from outside is 0.5α. When the system is in state (i,j), the rate of transitions into group 1 from group 2 is $j\sigma q$, so the overall rate is $\sum_{i,j} \pi_{(i,j)} j\sigma q$. Thus the fraction of new members coming in to group 1 from transfers is

$$\frac{\sum_{i,j} \pi_{(i,j)} j\sigma q}{\alpha + \sum_{i,j} \pi_{(i,j)} j\sigma q} \quad (8.69)$$

The above reasoning is very common, quite applicable in many situations. By the way, note that $\sum_{i,j} \pi_{(i,j)} j\sigma q = \sigma q EN$, where N is the number of members of group 1.

8.3.7 Continuous-Time Birth/Death Processes

We noted earlier that the system of equations for the π_i may not be easy to solve. In many cases, for instance, the state space is infinite and thus the system of equations is infinite too. However, there is a rich class of Markov chains for which closed-form solutions have been found, called **birth/death processes**.[8]

Here the state space consists of (or has been mapped to) the set of nonnegative integers, and p_{ji} is nonzero only in cases in which $|i - j| = 1$. (The name "birth/death" has its origin in Markov models of biological populations, in which the state is the current population size.) Note for instance that the example of the gracefully degrading system above has this form. An M/M/1 queue—one server, "Markov" (i.e. exponential) interarrival times and Markov service times—is also a birth/death process, with the state being the number of jobs in the system.

Because the p_{ji} have such a simple structure, there is hope that we can find a closed-form solution to (8.56), and it turns out we can. Let $u_i = \rho_{i,i+1}$ and $d_i = \rho_{i,i-1}$ ('u' for "up," 'd' for "down"). Then (8.56) is

$$\pi_{i+1} d_{i+1} + \pi_{i-1} u_{i-1} = \pi_i \lambda_i = \pi_i (u_i + d_i), \ i \geq 1 \tag{8.70}$$

$$\pi_1 d_1 = \pi_0 \lambda_0 = \pi_0 u_0 \tag{8.71}$$

In other words,

$$\pi_{i+1} d_{i+1} - \pi_i u_i = \pi_i d_i - \pi_{i-1} u_{i-1}, \ i \geq 1 \tag{8.72}$$

$$\pi_1 d_1 - \pi_0 u_0 = 0 \tag{8.73}$$

Applying (8.72) recursively to the base (8.73), we see that

$$\pi_i d_i - \pi_{i-1} u_{i-1} = 0, \ i \geq 1 \tag{8.74}$$

so that

$$\pi_i = \pi_{i-1} \frac{u_{i-1}}{d_i} \ i \geq 1 \tag{8.75}$$

[8] Though we treat the continuous-time case here, there is also a discrete-time analog.

and thus

$$\pi_i = \pi_0 r_i \tag{8.76}$$

where

$$r_i = \Pi_{k=1}^{i} \frac{u_{k-1}}{d_k} \tag{8.77}$$

where $r_i = 0$ for $i > m$ if the chain has no states past m.

Then since the π_i must sum to 1, we have that

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} r_i} \tag{8.78}$$

and the other π_i are then found via (8.76).

Note that the chain might be finite, i.e. have $u_i = 0$ for some i. In that case it is still a birth/death chain, and the formulas above for π still apply.

8.3.8 Example: Computer Worm

Not all interesting Markov chains have stationary distributions. Here is an example in which other considerations come into play. This chain happens to be a birth/death chain, but it is pure birth, and thus does not have a stationary distribution, or more accurately, has its stationary distribution concentrated on the absorbing state.

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 this as a pure birth process. 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.

Suppose that for each infected/noninfected pair of hosts, the time to infection of the noninfected member by the infected member has an exponential distribution with mean 1.0. Assume independence among the various infection attempts. Since in state i there are i(g-i) such pairs, and since we go to state i+1 when the first infection among these occurs, we have $\lambda_i = i(g-i)$. 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

$$\sum_{i=1}^{g-1} \frac{1}{i(g-i)} \tag{8.79}$$

Using a calculus approximation, we have

$$\int_1^{g-1} \frac{1}{x(g-x)} \, dx = \frac{1}{g} \int_1^{g-1} (\frac{1}{x} + \frac{1}{g-x}) \, dx = \frac{2}{g} \ln(g-1) \tag{8.80}$$

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, (8.79) 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.

8.4 Hitting Times Etc.

In this section we're interested in the amount of time it takes to get from one state to another, including cases in which this might be infinite.

8.4.1 Some Mathematical Conditions

There is a rich mathematical theory regarding the asymptotic behavior of Markov chains. We will not present such material here in this brief introduction, but we will give an example of the implications the theory can have.

A state in a Markov chain is called **recurrent** if it is guaranteed that, if we start at that state, we will return to the state infinitely many times. A nonrecurrent state is called **transient.**

Let T_{ii} denote the time needed to return to state i if we start there.[9] Note that an equivalent definition of recurrence is that $P(T_{ii} < \infty) = 1$, i.e. we are sure to return to i at least once. By the Markov property, if we are sure to return once, then we are sure to return again once after that, and so on, so this implies infinitely many visits.

[9] Keep in mind that T_{ii} is the time from one entry to state i to the next entry to state i. So, it includes time spent in i, which is 1 unit of time for a discrete-time chain and a random exponential amount of time in the continuous-time case, and then time spent away from i, up to the time of next entry to i.

A recurrent state i is called **positive recurrent** if $E(T_{ii}) < \infty$, while a state which is recurrent but not positive recurrent is called **null recurrent**.

Let T_{ij} be the time it takes to get to state j if we are now in i. Note that this is measured from the time that we enter state i to the time we enter state j.

One can show that in the discrete time case, a state i is recurrent if and only if

$$\sum_{n=0}^{\infty} P(T_{ii} = n) = \infty \tag{8.81}$$

This makes intuitive sense: Let A_n denote the indicator random variable for the event $T_{ii} = n$, i.e. A_n = 1 or 0, depending on whether $T_{ii} = n$. Then $P(T_{ii} = n) = EA_n$, so the left-hand side of (8.81) is the expected value of the total number of visits to state i. If state i is recurrent, then we will visit i infinitely often, and thus that sum should be equal to infinity.

Consider an **irreducible** Markov chain, meaning one which has the property that one can get from any state to any other state (though not necessarily in one step). One can show that in an irreducible chain, if one state is recurrent then they all are. The same statement holds if "recurrent" is replaced by "positive recurrent."

8.4.2 Example: Random Walks

Consider the famous **random walk** on the full set of integers: At each time step, one goes left one integer or right one integer (e.g. to +3 or +5 from +4), with probability 1/2 each. In other words, we flip a coin and go left for heads, right for tails.

If we start at 0, then we return to 0 when we have accumulated an equal number of heads and tails. So for even-numbered n, i.e. n = 2m, we have

$$P(T_{ii} = n) = P(\text{m heads and m tails}) = \binom{2m}{m} \frac{1}{2^{2m}} \tag{8.82}$$

One can use Stirling's approximation,

$$m! \approx \sqrt{2\pi} e^{-m} m^{m+1/2} \tag{8.83}$$

to show that the series (8.81) diverges in this case. So, this chain (meaning all states in the chain) is recurrent. However, it turns out not to be not positive recurrent, as we'll see below.

The same is true for the corresponding random walk on the two-dimensional integer lattice (moving up, down, left or right with probability 1/4 each). However, in the three-dimensional case, the chain is not even null recurrent; it is transient.

8.4.3 Finding Hitting and Recurrence Times

For a positive recurrent state i in a discrete-time Markov chain,

$$\pi_i = \frac{1}{E(T_{ii})} \tag{8.84}$$

The approach to deriving this is similar to that of Section 8.1.5.1. Define alternating On and Off subcycles, where On means we are at state i and Off means we are elsewhere. An On subcycle has duration 1, and an Off subcycle has duration $T_{ii} - 1$. Define a full cycle to consist of an On subcycle followed by an Off subcycle.

Then intuitively the proportion of time we are in state i is

$$\pi_i = \frac{E(\text{On})}{E(\text{On}) + E(\text{Off})} = \frac{1}{1 + E(T_{ii} - 1)} = \frac{1}{ET_{ii}} \tag{8.85}$$

The equation is similar for the continuous-time case. Here $E(\text{On}) = 1/\lambda_i$. The Off subcycle has duration $T_{ii} - 1/\lambda_i$. Note again that T_{ii} is measured from the time we enter state i once until the time we enter it again. We then have

$$\pi_i = \frac{1/\lambda_i}{E(T_{ii})} \tag{8.86}$$

Thus positive recurrence means that $\pi_i > 0$. For a null recurrent chain, the limits in Equation (8.3) are 0, which means that there may be rather little one can say of interest regarding the long-run behavior of the chain.

We are often interested in finding quantities of the form $E(T_{ij})$. We can do so by setting up systems of equations similar to the balance equations used for finding stationary distributions.

First consider the discrete case. Conditioning on the first step we take after being at state i, and using the Law of Total Expectation, we have

$$E(T_{ij}) = \sum_{k \neq j} p_{ik}[1 + E(T_{kj})] + p_{ij} \cdot 1 \tag{8.87}$$

By varying i and j in (8.87), we get a system of linear equations which we can solve to find the ET_{ij}. Note that (8.84) gives us equations we can use here too.

The continuous version uses the same reasoning:

$$E(T_{ij}) = \sum_{k \neq j} p_{ik} \left[\frac{1}{\lambda_i} + E(T_{kj}) \right] + p_{ij} \cdot \frac{1}{\lambda_i} \tag{8.88}$$

One can use a similar analysis to determine the probability of ever reaching a state, in chains which have transient or **absorbing** states, i.e. states i such that $p_{ij} = 0$ whenever $j \neq i$. For fixed j define

$$\alpha_i = P(T_{ij} < \infty) \tag{8.89}$$

Then

$$\alpha_i = \sum_{k \neq j} p_{ik} \alpha_k + p_{ij} \tag{8.90}$$

8.4.4 Example: Finite Random Walk

Let's go back to the example in Section 8.1.1.

Suppose we start our random walk at 2. How long will it take to reach state 4? Set $b_i = E(T_{i4} | \text{start at i})$. From (8.87) we could set up equations like

$$b_2 = \frac{1}{3}(1 + b_1) + \frac{1}{3}(1 + b_2) + \frac{1}{3}(1 + b_3) \tag{8.91}$$

Now change the model a little, and make states 1 and 6 absorbing. Suppose we start at position 3. What is the probability that we eventually are absorbed at 6 rather than 1? We could set up equations like (8.90) to find this.

8.4.5 Example: Tree-Searching

Consider the following Markov chain with infinite state space $\{0,1,2,3,...\}$.[10] The transition matrix is defined by $p_{i,i+1} = q_i$ and $p_{i0} = 1 - q_i$. This kind of model has many different applications, including in computer science tree-searching algorithms. (The state represents the level in the tree where the search is currently, and a return to 0 represents a backtrack. More general backtracking can be modeled similarly.)

[10] Adapted from *Performance Modelling of Communication Networks and Computer Architectures*, by P. Harrison and N. Patel, pub. by Addison-Wesley, 1993.

The question at hand is, What conditions on the q_i will give us a positive recurrent chain?

Assuming $0 < q_i < 1$ for all i, the chain is clearly irreducible. Thus, to check for recurrence, we need check only one state, say state 0.

For state 0 (and thus the entire chain) to be recurrent, we need to show that $P(T_{00} < \infty) = 1$. But

$$P(T_{00} > n) = \Pi_{i=0}^{n-1} q_i \tag{8.92}$$

Therefore, the chain is recurrent if and only if

$$\lim_{n \to \infty} \Pi_{i=0}^{n-1} q_i = 0 \tag{8.93}$$

For positive recurrence, we need $E(T_{00}) < \infty$. Now, one can show that for any nonnegative integer-valued random variable Y

$$E(Y) = \sum_{n=0}^{\infty} P(Y > n) \tag{8.94}$$

Thus for positive recurrence, our condition on the q_i is

$$\sum_{n=0}^{\infty} \Pi_{i=0}^{n-1} q_i < \infty \tag{8.95}$$

Exercises

1. Consider a “wraparound” variant of the random walk in Section 8.1.1. We still have a reflecting barrier at 1, but at 5, we go back to 4, stay at 5 or “wrap around” to 1, each with probability 1/3. Find the new set of stationary probabilities.

2. Consider the Markov model of the shared-memory multiprocessor system in our PLN. In each part below, your answer will be a function of $q_1, ..., q_m$.

(a) For the case m = 3, find $p_{(2,0,1),(1,1,1)}$.

(b) For the case m = 6, give a compact expression for $p_{(1,1,1,1,1,1),(i,j,k,l,m,n)}$.

Hint: We have an instance of a famous parametric distribution family here.

3. This problem involves the analysis of call centers. This is a subject of much interest in the business world, with there being commercial simulators sold to analyze various scenarios. Here are our assumptions:

- Calls come in according to a Poisson process with intensity parameter λ.
- Call duration is exponentially distributed with parameter η.
- There are always at least b operators in service, and at most b+r.
- Operators work from home, and can be brought into or out of service instantly when needed. They are paid only for the time in service.
- If a call comes in when the current number of operators is larger than b but smaller than b+r, another operator is brought into service to process the call.
- If a call comes in when the current number of operators is b+r, the call is rejected.
- When an operator completes processing a call, and the current number of operators (including this one) is greater than b, then that operator is taken out of service.

Note that this is a birth/death process, with the state being the number of calls currently in the system.

(a) Find approximate closed-form expressions for the π_i for large b+r, in terms of b, r, λ and η. (You should not have any summation symbols.)

(b) Find the proportion of rejected calls, in terms of π_i and b, r, λ and η.

(c) An operator is paid while in service, even if he/she is idle, in which case the wages are "wasted." Express the proportion of wasted time in terms of the π_i and b, r, λ and η.

(d) Suppose b = r = 2, and $\lambda = \eta = 1.0$. When a call completes while we are in state b+1, an operator is sent away. Find the mean time until we make our next summons to the reserve pool.

4. The **bin-packing problem** arises in many computer science applications. Items of various sizes must be placed into fixed-sized bins. The goal is to find a packing arrangement that minimizes unused space. Toward that end, work the following problem.

We are working in one dimension, and have a continuing stream of items arriving, of lengths $L_1, L_2, L_3, ...$. We place the items in the bins in the order of arrival, i.e. without optimizing. We continue to place items in a bin until we encounter an item that will not fit in the remaining space, in which case we go to the next bin.

Suppose the bins are of length 5, and an item has length 1, 2, 3 or 4, with probability 0.25 each. Find the long-run proportion of wasted space.

Hint: Set up a discrete-time Markov chain, with "time" being the number of items packed so far, and the state being the amount of occupied space in the current bin. Define T_n to be 1 or 0, according to whether the n^{th} item causes us to begin packing a new bin, so that the number of bins used by "time" n is $T_1 + ... + T_n$.

5. Suppose we keep rolling a die. Find the mean number of rolls needed to get three consecutive 4s.

Hint: Use the material in Section 8.4.

6. A system consists of two machines, with exponentially distributed lifetimes having mean 25.0. There is a single repairperson, but he is not usually on site. When a breakdown occurs, he is summoned (unless he is already on his way or on site), and it takes him a random amount of time to reach the site, exponentially distributed with mean 2.0. Repair time is exponentially distributed with mean 8.0. If after completing a repair the repairperson finds that the other machine needs fixing, he will repair it; otherwise he will leave. Repair is performed on a First Come, First Served schedule. Find the following:

(a) The long-run proportion of the time that the repairperson is on site.

(b) The rate per unit time of calls to the repairperson.

(c) The mean time to repair, i.e. the mean time between a breakdown of a machine and completion of repair of that machine.

(d) The probability that, when two machines are up and one of them goes down, the second machine fails before the repairperson arrives.

7. Consider again the random walk in Section 8.1.1. Find

$$\lim_{n \to \infty} \rho(X_n, X_{n+1}) \tag{8.96}$$

Hint: Apply the Law of Total Expectation to $E(X_n X_{n+1})$.

8. Consider a random variable X that has a continuous density. That implies that $G(u) = P(X > u)$ has a derivative. Differentiate (8.48) with respect to r, then set r = 0, resulting in a differential equation for G. Solve that equation to show that the only continuous densities that produce the memoryless property are those in the exponential family.

9. Suppose we model a certain database as follows. New items arrive according to a Poisson process with intensity parameter α. Each item stays in the database for an exponentially distributed amount of time with parameter σ, independently of the other items. Our state at time t is the number of items in the database at that time. Find closed-form expressions for the stationary distribution π and the long-run average size of the database.

10. Consider our machine repair example in Section 8.3.5, with the following change: The repairperson is offsite, and will not be summoned unless both machines are down. Once the repairperson arrives, she will not leave until both machines are up. So for example, if she arrives and repairs machine B, then while repairing A finds that B has gone down again, she will start work on B immediately after finishing with A. Travel time to the site from the maintenance office is 0. Repair is performed on a First Come, First Served schedule. The time a machine is in working order has an exponential distribution with rate ω, and repair is exponentially distributed with rate ρ. Find the following in terms of ω and ρ:

(a) The long-run proportion of the time that the repairperson is on site.

(b) The rate per unit time of calls to the repairperson.

(c) The mean time to repair, i.e. the mean time between a breakdown of a machine and completion of repair of that machine. (Hint: The best approach is to look at rates. First, find the number of breakdowns per unit time. Then, ask how many of these occur during a time when both machines are up, etc. In each case, what is the mean time to repair for the machine that breaks?)

11. There is a town with two social groups, with the following dynamics:

- Everyone is in exactly one group at a time.
- People arrive from outside town, with exponentially distributed interarrival times at rate α, and join one of the groups with probability 0.5 each.
- Each person will occasionally switch groups, with one possible "group" being to leave town entirely (never to return). A person's time before switching groups is exponentially distributed with rate σ. The switch will either be to the other group or to the outside world, with probabilities q and 1-q, respectively.

Let the state of the system be (i,j), where i and j are the number of current members in groups 1 and 2, respectively. Answer in terms of α, λ, τ and π:

(a) Give the balance equation for the state (8,8).

(b) Fill in the blank: The president of Group 1 tells reporter, "We've found over the years that ______________________% of entries into our group come as transfers from the other group."

Chapter 9

Renewal Theory and Some Applications

9.1 Introduction

9.1.1 The Light Bulb Example, Generalized

Supposre a certain lamp is used continuously, and that whenever its bulb burns out, it is immediately replaced by a new one. Let N(t) denote the number of replacements, called **renewals** here, that have occurred up through time t. Assume the lifetimes of the bulbs, $L_1,\ L_{2,}\ L_{3,}$ ... are independent and identically distributed. The collection of random variables $N(t), t \geq 0$ is called a **renewal process.** The quantities $R_i = L_1 + ... + L_i$ are called the **renewal points**.

We will see that most renewal processes are not Markovian. However, time does "start over" at the renewal points. We say the process is **regenerative** at these points.

Note that although we are motivating this with the lightbulb example, in which the L_i measure time, the theory we will present here is not limited to such a context. All that is needed is that the L_i be i.i.d. and nonnegative.

There is a very rich collection of mathematical material on renewal processes, and there are myriad applications to a variety of fields.

9.1.2 Duality Between "Lifetime Domain" and "Counts Domain"

A very important property of renewal processes is that

$$N(t) \geq k \text{ if and only if } R_k \leq t \tag{9.1}$$

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 (9.1) might be described as relating the “counts domain” (left-hand side of the equation) to the “lifetimes domain” (right-hand side).

9.2 Where We Are Going

Some of the material in Sections 9.3 and 9.4 of this chapter may seem a little theoetical. However, it all does have practical value, and it will also exercise some of the concepts you’ve learned in earlier chapters.

After those two sections, the focus will mainly be on concepts which apply the theory, and on specific examples.

9.3 Properties of Poisson Processes

9.3.1 Definition

A renewal process N(t) is called a **Poisson process** if each N(t) has a Poisson distribution, i.e. for some λ we have

$$P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!},\ k = 0, 1, 2, 3, ... \tag{9.2}$$

The parameter λ is called the **intensity parameter** of the process, and since $E[N(t)] = \lambda t$, it has the natural interpretation of the (average) rate at which renewal events occur per unit time.

9.3.2 Alternate Characterizations of Poisson Processes

9.3.2.1 Exponential Interrenewal Times

Theorem 15 *A renewal process N(t) is a Poisson process if and only if the interrenewal times L_i have an exponential distribution with parameter λ.*

Note that this shows that a Poisson process is also a Markov chain (with state being the current number of renewals), due to the memoryless property of the exponential distribution.

Proof

For the "only if" part of the claim, note that from the "domain switching" discussed in Section (9.1.2),

$$F_{L_1}(t) = 1 - P(L_1 > t) = 1 - P(N(t) = 0) = 1 - e^{-\lambda t} \tag{9.3}$$

Differentiating with respect to t, we find that $f_{L_1}(t) = \lambda e^{-\lambda t}$. Since the L_i are i.i.d. by definition of renewal proceses, this shows that all the L_i have an exponential distribution.

■

9.3.2.2 Stationary, Independent Increments

Theorem 16 *Suppose N(t) is a renewal process N(t) for which the L_i are continuous random variables.*

Then N(t) is a Poisson process if and only if it has ***stationary, independent increments****. The latter term means that for all $0 < r < s < t < u$ we have the following properties:*

(a) Independent increments: N(s)-N(r) and N(u)-N(t) are independent random variables.

(b) Stationarity: The distribution of N(s)-N(r) is the same as that of N(s+z)-N(r+z) for any z > 0.

Proof

(Sketch.)

For the "only if" part: As noted above, Poisson processes have the Markov property, which immediately implies independent increments. Also, since "time starts over," say at time r, then for s > r we will have that N(s) - N(r) has the same distribution as N(s-r).

For the "if" part: The independence of the increments implies the Markov property, and since the only continuous distribution which is memoryless is the exponential, that implies the the L_i have exponential distributions, which we saw above implies that N(t) is Poisson.

■

Note too that for these reasons N(s)-N(r) will have the same distribution as N(s-r).

We can use these properties to find the **autocorrelation function** of the Poisson process, which shows the correlation of the process with itself at different times $s < t$:

$$c(s,t) = \frac{Cov[N(s), N(t)]}{\sqrt{Var[N(s)]Var[N(t)]}} \tag{9.4}$$

To derive this quantity, note first that

$$
\begin{aligned}
E[N(s)N(t)] &= E\left[N(s)\{N(t)-N(s)\}+N^2(s)\right] && (9.5)\\
&= E\left[N(s)\{N(t)-N(s)\}\right]+E\left[N^2(s)\right] && (9.6)\\
&= E[N(s)]E[N(t)-N(s)]+E\left[N^2(s)\right] && (9.7)\\
&= [\lambda s\cdot\lambda(t-s)]+[\lambda s+(\lambda s)^2] && (9.8)
\end{aligned}
$$

Here we used the property of independent increments in going from the second to third line. In the last line we have used the fact that N(u) has mean and variance both equal to λu for all $u > 0$.

After performing the remaining calculations, we find that

$$c(s,t) = \sqrt{\frac{s}{t}} \qquad (9.9)$$

9.3.3 Conditional Distribution of Renewal Times

Theorem 17 *Suppose N(t) is a Poisson process. Let $R_i = L_1 + ... + L_i$ be the renewal times, i = 1,2,3,...*

Given N(t) = k, then $M_1, ..., M_k$ are i.i.d. U(0,t) (uniform distribution on (0,t)), where the $(M_1, ..., M_k)$ is a random permutation of $(R_1, ..., R_k)$.

In other words, conditional on there being k renewals within (0,t), the unordered renewal times are independent and uniformly distributed on (0,t).

This fact often plays a key role in analyses of Poisson models.

Let's look at the intuitive meaning of this, using our "notebook" view, taking k = 3 and t = 12.0 for concreteness.

Each line of the notebook would consist of the results of our observing the process up to time 12.0. The first column would show the R_1, the second R_2 and so on.

We would also have columns for M_1, M_2 and M_3. These would be filled with NAs ("not applicable") in rows for which N(12.0) is not equal 3, while in rows for which N(12.0) *is* equal to 3, (M_1, M_2, M_3) would be a random scrambling of (R_1, R_2, R_3).

The uniform distribution part of the theorem is saying that among the rows in which N(12,0) = 3, 1/2 of them will have, for instance, $M_1 < 6.0$. That will NOT be true for R_1; considerably fewer than 1/2 of the rows will have $R_1 < 6.0$. The independence part of the theorem is saying, for example, that 1/4 of these rows will have *both* $M_1 < 6.0$ *and* $M_2 < 6.0$.

Here's the proof.

Proof

Let Y denote the number of $M_i \leq s$. Then

$$\begin{aligned}
P(Y = b|N(t) = k) &= \frac{P(Y = b \text{ and } N(t) = k)}{P[N(t) = k]} && (9.10)\\
&= \frac{P[N(s) = b \text{ and } N(t) - N(s) = k - b]}{P[N(t) = k]} && (9.11)\\
&= \frac{\frac{e^{-\lambda s}(\lambda s)^b}{b!} \cdot \frac{e^{-\lambda(t-s)}(\lambda(t-s))^{k-b}}{(k-b)!}}{\frac{e^{\lambda t}(\lambda t)^k}{k!}} && (9.12)\\
&= \binom{k}{b}(\frac{s}{t})^b(1 - \frac{s}{t})^{k-b} && (9.13)
\end{aligned}$$

This is the probability that would arise if the M_i were i.i.d. U(0,t) as claimed in the theorem.

■

9.3.3.1 Example: Message Buildup at a Broken Network Link

Suppose messages arrive to a network link according to a Poisson process. Unfortunately, the link is down, and the messages just keep piling up. Suppose at time 60.0 we see 10 messages in the buffer. Find the variance of the total wait time of those messages.

Solution: Number the messages 1 through 10, not necessarily in the order of their arrival. Say for instance that these messages came from mouse clicks of 10 Web users, and we order the messages by the surnames of the users. That gives us a random permutation of the 10 messages, and since we are only analyzing the total of the 10 wait times, it doesn't matter which permutation we have.

Let $W_1, ..., W_{10}$ be the wait times of our 10 messages. Then, conditional on N(60.0), the W_i are independent and i.i.d. U(0,60.0). Then from Section 2.3.1.1, $EW_i = 6.0$ and $Var(W_i) = 60^2/12 = 300$. The mean and variance of the total wait time are thus 60 and 3000.

9.3.4 Decomposition and Superposition of Poisson Processes

Theorem 18 *Poisson processes can be split and combined:*

(a) *Let N(t) be a Poisson process with intensity parameter λ. Say we* ***decompose*** *N(t) into $N_1(t)$ and $N_2(t)$ by assigning each renewal in N(t) to either $N_1(t)$ or $N_2(t)$ with probability p and 1-p respectively. Then the two resulting processes are again Poisson processes. They are independent of each other, and have intensity parameters $p\lambda$ and $(1-p)\lambda$.*

(b) *If we* ***superimpose*** *two independent Poisson processes $N_1(t)$ and $N_2(t)$, the result $N(t) = N_1(t) + N_2(t)$ will be a Poisson process, with intensity parameter equal to the sum of the two original parameters.*

These properties are often useful in queuing models, where an arrival process is subdivided in two processes corresponding to two job classes.

9.3.5 Nonhomogeneous Poisson Processes

A useful variant of Poisson processes is the **nonhomogeneous Poisson process**. The key here is that the intensity parameter λ varies over time. We will write it as $\lambda(t)$. I'll define it this way:

Definition 19 *Let $N(t), t \geq 0$ be a counting process with independent increments, and let*

$$m(t) = E[N(t)] \tag{9.14}$$

If for all $0 < s < t$ we have

$$P[N(t) - N(s) = k] = \frac{e^{-m(t)}[m(t)[^k}{k!}, k = 0, 1, 2, ... \tag{9.15}$$

then we say that $N(t), t \geq 0$ is a nonhomogeneous Poisson process with intensity function

$$\lambda(t) = \frac{d}{dt} m(t) \tag{9.16}$$

Intuitively, $N(t), t \geq 0$ is a Markov process with states of the form [t,N(t)]. In that state, the probability that there will be a renewal in the next Δt amount of time is approximately $\Delta t \cdot \lambda(t)$, for small increments of time Δt.

9.3.5.1 Example: Software Reliability

Nonhomogeneous Poisson process 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 m(t) from that data.

See for example Estimating the Parameters of a Non-homogeneous Poisson-Process Model for Software Reliability, *IEEE Transactions on Reliability, Vol. 42, No. 4, 1993.*

9.4 Properties of General Renewal Processes

We now turn our attention to the general case, in which the L_i are not necessarily exponentially distributed. We will still assume the L_i to be continuous random variables, though.

9.4.1 The Regenerative Nature of Renewal Processes

Recall that Markov chains are “memoryless.” If we are now at time t, “time starts over,” and the probabilities of events after time t do not depend on what happened before time t.

If the L_i are not exponentially distributed, N(t) is not Markovian, since the exponential distribution is the only continuous memoryless distribution. However, it is true that “time starts over” at each renewal epoch $R_i = L_1 + ... + L_i$. Note the difference: The definition of the Markov property concerns time starting over at a fixed time, t. Here, in the context of renewal processes, “time starts over” at random times, of the form R_i.

9.4.2 Some of the Main Theorems

Let m(t) = E[N(t)]. Many of the results concern this function m. Please forgive a bit of abuse of notation:

$$F = F_L \tag{9.17}$$

$$f = f_L \tag{9.18}$$

$$F_n = F_{R_n} \tag{9.19}$$

9.4.2.1 The Functions F_n Sum to m

First, we need a general property of means of nonnegative random variables (this was an exercise in Chapter 1):

Lemma 20 *For any nonnegative-integer valued random variable Z,*

$$E(Z) = \sum_{j=1}^{\infty} P(Z \geq j) = \sum_{j=0}^{\infty} [1 - F_Z(j)] \tag{9.20}$$

If Z is a nonnegative continuous random variable, then

$$E(Z) = \int_0^{\infty} [1 - F_Z(t)]\, dt \tag{9.21}$$

Proof

In the discrete case,

$$\begin{aligned}
E(Z) &= \sum_{i=1}^{\infty} i\, P(Z = i) && (9.22)\\
&= \sum_{i=1}^{\infty} P(Z = i) \sum_{j=1}^{i} 1 && (9.23)\\
&= \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} P(Z = i) && (9.24)\\
&= \sum_{j=1}^{\infty} P(Z \geq j) && (9.25)\\
&= \sum_{j=0}^{\infty} P(Z > j) && (9.26)\\
&= \sum_{j=0}^{\infty} [1 - F_Z(j)] && (9.27)
\end{aligned}$$

The continuous case is similar.

■

Now, let's apply that:

Theorem 21

$$m(t) = \sum_{n=1}^{\infty} F_n(t) \tag{9.28}$$

Proof

Let Z = N(t). Then yet again switching domains as in Section (9.1.2),

$$m(t) = \sum_{j=1}^{\infty} P(N(t) \geq j) = \sum_{j=1}^{\infty} P(R_j \leq t) = \sum_{j=1}^{\infty} F_j(t) \tag{9.29}$$

■

9.4.2.2 The Renewal Equation

Theorem 22 *The function m satisfies the equation*

$$m(t) = F(t) + \int_0^t m(t-w)\, f(w)\, dw \tag{9.30}$$

Proof

Using the Law of Total Expectation, we have

$$m(t) = E[N(t)] = E\{E[N(t)|L_1]\} \tag{9.31}$$

But at time L_1 "time starts over again," by the regenerative property of renewal processes. So,

$$E[N(t)|L_1) = \begin{cases} 1 + m(t - L_1), & \text{if } L_1 < t \\ 0, & \text{otherwise} \end{cases} \tag{9.32}$$

Remember, $E[N(t)|L_1)$ is a function of L_1. Thus its expected value, which we want to find now (see (9.31)), is the integral of that function times the density of L_1:

$$m(t) = \int_0^t [1 + m(t-w)] \, f(w) \, dw = F(t) + \int_0^t m(t-w) \, f(w) \, dw \tag{9.33}$$

■

If f is known, then F is known too, then Equation (9.30) can be solved for m. This is known as an **integral equation** for m, analogous to a differential equation. In fact, it can be converted to a differential equation by taking the derivative of both sides. There are various techniques for solving integral equations, including numerical approximations, but we will not pursue those here.

The function m is useful in a number of contexts (one of which will be seen in Section 9.5.2), so we will now look at some of its properties.

9.4.2.3 The Function m(t) Uniquely Determines F(t)

Remember, each different distribution for the L_i gives rise to a different distribution for the N(t), thus a different m(t). The converse is also true:

Theorem 23 *The function m(t) uniquely determines F(t).*

Proof

For any function h (for which the absolute value has a finite integral on $(0, \infty)$), let Λ_h define the Laplace transform of h,

$$\Lambda_h(s) = \int_0^\infty e^{-st} h(t) \, dt \tag{9.34}$$

Now, we will take the Laplace transform of both sides of (9.30). To do so, we will need to state some facts about Laplace transforms.

Remember, Laplace transforms are, except for a change of variable, just like generating functions or moment generating functions, which you probably studied in your undergraduate probability course. Thus Laplace transforms have the same properties as generating functions.

Now recall that the generating function of the sum of two independent nonnegative random variables is the product of their individual moment generating functions. Well, the density of such a sum has the same integral form as in (9.30), which we call a **convolution.** Even though the m part of the integral is not a density function, it is still true that the Laplace transform of a convolution of two functions is the product of their individual Laplace transforms.

So, taking the Laplace transform of both sides of (9.30) yields:

$$\Lambda_m(s) = \Lambda_F(s) + \Lambda_m(s)\Lambda_f(s) \tag{9.35}$$

i.e.

$$\Lambda_m(s) = \frac{\Lambda_F(s)}{1 - \Lambda_f(s)} \tag{9.36}$$

This says that m uniquely determines the Laplace transforms of F and f, and since there is a one-to-one correspondence between distributions and Laplace transforms, we see that m uniquely determines the interrenewal distribution. In other words, there is only one possible interrenewal distribution for any given mean function.

■

In fact, some similar analysis, which we will not present here, yields:

$$\Lambda_f(s) = \frac{\Lambda_r(s)}{1 + \Lambda_r(s)} \tag{9.37}$$

where $r(t) = \frac{d}{dt}m(t)$. So, we can recover f from m.

As an example, we earlier saw that there are three equivalent definitions of a Poisson process, i.e. each implies the other. We can use the above result to show one of those equivalences:

Suppose a renewal process has stationary, independent increments. This would imply that m(t) = ct for some c $>$ 0, and thus r(t) = c. Then

$$\Lambda_r(s) = \int_0^\infty e^{-st} c\, dt = \frac{c}{s} \tag{9.38}$$

so (9.37) gives us

$$\Lambda_f(s) = \frac{c}{s+c} \tag{9.39}$$

This is the Laplace transform for the exponential distribution with parameter c. So, we see how the stationary, independent increments property implies exponential interrenewal times.

9.4.2.4 Asymptotic Behavior of m(t)

Theorem 24

$$\lim_{t \to \infty} \frac{m(t)}{t} = \frac{1}{E(L)} \tag{9.40}$$

This should make good intuitive sense to you. By the way, this (and some other things we'll find in this chapter) can be used to formally prove some assertions we made on only an intuitive basis in early chapters.

9.5 Alternating Renewal Processes

9.5.1 Definition and Main Result

Suppose we have a sequence of pairs of random variables (Y_n, Z_n) which are i.i.d. *as pairs*. In other words, for instance the pair (Y_1, Z_1) is independent of, but has the same distribution as, (Y_2, Z_2) and so on, but on the other hand Y_n and Z_n are allowed dependency.

Imagine a machine being busy, then idle, then busy, then idle, and so on, with Y_n being the amount of "on" time in the n^{th} cycle, and similarly Z_n being the "off" time. Each time an on/off pair finishes, call that a "renewal." The sequence is called an **alternating renewal process**.

It is intuitively clear, and can be proven, that

$$\lim_{t \to \infty} P(\text{on at time t}) = \frac{E(Y)}{E(Y) + E(Z)} \tag{9.41}$$

and

$$\lim_{t \to \infty} P(\text{off at time t}) = \frac{E(Z)}{E(Y) + E(Z)} \tag{9.42}$$

Again, these results can be used to formally prove some assertions we have made in earlier chapters on just an intuitive basis.

9.5.2 Example: Inventory Problem (difficult)

(Adapted from *Stochastic Processes*, by Sheldon Ross, Wiley, 1996.)

Consider a vendor which uses an **(s,S) policy** for replenishing its inventory.[1] What this means is that after filling a customer order, if inventory of the item falls below level s, the inventory is replenished to level S.

Suppose customer orders arrive as a renewal process, with the L_n in this case being i.i.d. interarrival times. Let O_n denote the size of the n^{th} order, and let I(t) denote the amount of inventory on hand at time t. Take I(0) to be S. We wish to find the long-run distribution of I(t), i.e.

$$\lim_{t \to \infty} P(I(t) \geq u) \tag{9.43}$$

for all $s < u < S$.

To do this, define the following alternating renewal process: The system is "on" when the inventory level is at least u, and "off" otherwise. So, we start out with inventory S, begin losing it during this "on" period, and eventually it falls below u, which is a transition from on to off. The inventory continues to fall, during this off time, and it eventually falls below s, at which time there is a replenishment back up to S. At that point the next "on" period starts, etc.

Note that each on/off cycle begins with the inventory being at level S, and that each transition from on to off and vice versa coincides with some renewal time R_n in the customer arrival process.

Then Equation (9.41) says that

$$\lim_{t \to \infty} P(I(t) \geq u) = \frac{\text{E(amount of time inventory } \geq \text{ u in an on/off cycle)}}{\text{E(time of an on/off cycle)}} \tag{9.44}$$

To evaluate numerator and denominator in (9.44), consider the first cycle, and let

$$N_x = min\{n : O_1 + ... + O_n > S - x\} \tag{9.45}$$

In other words, N_x is the index of the first order which makes the inventory fall below x. All the time prior to this order, the inventory is at least x. Then our earlier point that all transitions into the on and off states coincide with some R_i can be refined to say that

- Customer number N_u triggers the end of the on cycle. Thus the length of the on cycle is R_{N_u}.
- Customer number N_s ends the full cycle. Thus the length of the full cycle is R_{N_s}.

[1] I generally try to reserve capital letters for names of random variables, using lower-case letters (or Greek) to denote constants, but am using Ross' notation here.

Therefore the fraction in (9.44) is equal to

$$\frac{E(R_{N_u})}{E(R_{N_s})} = \frac{E(\sum_{i=1}^{N_u} L_i)}{E(\sum_{i=1}^{N_s} L_i)} = \frac{E(N_u)E(L)}{E(N_s)E(L)} = \frac{E(N_u)}{E(N_s)} \tag{9.46}$$

where L is a random variable having the common distribution of the L_i. Here we have used the fact that the O_i are independent of the L_j.

Now consider a new "renewal process," with the O_i playing the role of "lightbulb lifetimes." This is a bit difficult to get used to, since the O_i are of course not times, but they are nonnegative random variables, and thus from the considerations of (9.1) we see that

$$\tilde{N}(t) = max\{n : O_1 + ... + O_n \leq t\} \tag{9.47}$$

is indeed a renewal process. The importance of that renewal process is that

$$\tilde{N}(t) + 1 = N_{S-t} \tag{9.48}$$

Now let

$$m_O(x) = E\tilde{N}(x) \tag{9.49}$$

Then (9.44) and 9.46) imply that

$$\lim_{t \to \infty} P(I(t) \geq u) = \frac{m_O(S-u)+1}{m_O(S-s)+1} \tag{9.50}$$

Though Ross' treatment ends at this point, we can also extend it by using (9.40) if S-s is large, yielding

$$\lim_{t \to \infty} P(I(t) \geq u) \approx \frac{\frac{S-u}{EO}+1}{\frac{S-s}{EO}+1} = \frac{S-u+EO}{S-s+EO} \tag{9.51}$$

9.6 Residual-Life Distribution

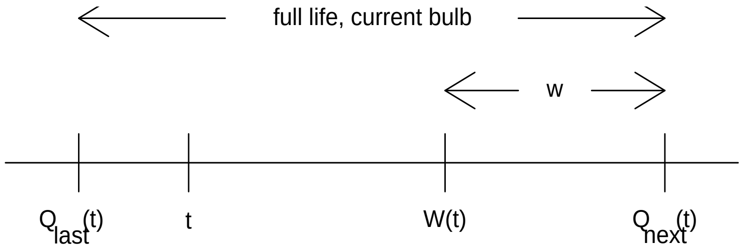
(It is assumed here that you know about the "bus paradox," described in Section 2.5 of our chapter on continuous distributions.)

9.6.1 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. So, in general, the time from a fixed time point t until the next renewal, is known as the **residual life**. (Another name for it is the **forward recurrence time.**)

Here is a derivation for the continuous case. Consider a renewal process, which for concreteness we will describe in the "light bulb" context but is fully general. Define the following for any time $t, w \geq 0$:

$$D(t) = \text{remaining life of the current bulb} \tag{9.52}$$
$$Q_{last}(t) = \text{time of the most recent burnout} \tag{9.53}$$
$$Q_{next}(t) = \text{time of the most recent burnout} \tag{9.54}$$
$$W(t) = \max[Q_{next}(t) - w, Q_{last}(t)] \tag{9.55}$$



The definition of W(t) sounds complicated, but it is merely saying this: Looking at the picture, start at $Q_{next}(t)$ and move your eyes leftward a distance w, but no further leftward than $Q_{last}(t)$. The place where you end up is defined to be W(t).

Though W(t) is shown in the picture as being to the right of t, the opposite could be true. In fact,

$$D(t) \leq w \text{ if and only if } W(t) \leq t \tag{9.56}$$

Our goal is to find

$$\lim_{t\to\infty} P[D(t) \leq w] \tag{9.57}$$

But (9.56) shows that

$$\lim_{t\to\infty} P[D(t) \leq w] = \lim_{t\to\infty} P[W(t) \leq t] = \text{ long-run fraction of the time } W(t) \leq t \text{ as } t \to \infty \tag{9.58}$$

(Existence of that long-run fraction as a constant can be shown mathematically to imply that the limiting probability exists and is equal to that constant. But the result is intuitively clear.)

Now let's evaluate the far right-hand side of (9.58). For each light bulb, let Z(t) the length of the portion of its lifetime to the right of W(t), i.e.

$$Z(t) = Q_{next}(t) - W(t) \tag{9.59}$$

and let Y(t) be the portion to the left of W(t). Then:

$$W(t) \leq t \text{ if and only if t is in the interval covered by Z(t)} \tag{9.60}$$

The Ys and Z form an alternating renewal process as we move from bulb to bulb. Thus from Equations (9.42), (9.58) and (9.60) we have that

$$\begin{aligned} \lim_{t\to\infty} P(D(t) \leq w) &= \frac{E(Z)}{E(Y)+E(Z)} && (9.61) \\ &= \frac{E[\min(L,w)]}{E(L)} && (9.62) \end{aligned}$$

where L is the lifetime of a bulb.

By Lemma 20 we have

$$E[\min(L,w)] = \int_0^\infty P[\min(L,w) > u]\,du = \int_0^w P(L > u)\,du = \int_0^w [1 - F_L(u)]\,du \tag{9.63}$$

since $P[\min(L,w) > u] = 0$ whenever $u > w$.

Substituting this into (9.61), and taking derivatives with respect to w, we have that

$$\lim_{t\to\infty} f_{D(t)}(w) = \frac{1-F_L(w)}{E(L)} \tag{9.64}$$

This is a classic result, of central importance and usefulness, as seen in our upcoming examples later in this section.

9.6.2 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

$$\lim_{t\to\infty} f_{A(t)}(w) = \frac{1-F_L(w)}{E(L)} \tag{9.65}$$

In other words, A(t) has the same long-run distribution as D(t)!

Here is a derivation for the case in which the L_i are discrete. 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. Then we have a Markov chain in which our state at any time is the value of W. 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.

Let's find the transition probabilities. 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,

$$p_{i,0} = P(L = i+1 | L > i) = \frac{p_L(i+1)}{1-F_L(i)} \tag{9.66}$$

$$p_{i,i+1} = \frac{1-F_L(i+1)}{1-F_L(i)} \tag{9.67}$$

Define

$$q_i = \frac{1-F_L(i+1)}{1-F_L(i)} \tag{9.68}$$

and write

$$\pi_{i+1} = \pi_i q_i \tag{9.69}$$

Applying (9.69) recursively, we have

$$\pi_{i+1} = \pi_0 q_i q_{i-1}) \cdots q_0 \tag{9.70}$$

But the right-hand side of (9.70) telescopes down to

$$\pi_{i+1} = \pi_0 [1 - F_L(i+1)] \tag{9.71}$$

Then

$$1 = \sum_{i=0}^{\infty} \pi_i = \pi_0 \sum_{i=0}^{\infty} [1 - F_L(i)] = \pi_0 E(L) \tag{9.72}$$

by (20).

Thus

$$\pi_i = \frac{1 - F_L(i+1)}{EL} \tag{9.73}$$

in analogy to (9.65).

9.6.3 Mean of the Residual and Age Distributions

Taking the expected value of (9.64) or (9.65), we get a double integral. Reversing the order of integration, we find that the mean residual life or age is given by

$$\frac{E(L^2)}{2EL} \tag{9.74}$$

9.6.4 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.

9.6.5 Example: The (S,s) Inventory Model Again

Here I extend Ross' example that we saw in Section 9.5.2.

When an order causes the inventory to go below s, we must dip into our reserves to fill it. Let R be the amount of reserves we must draw upon. Assuming that we always have sufficient reserves, what is the distribution of R?

Recall the renewal process $\tilde{N}(t)$ in Section 9.5.2. Then the distribution of R is that of the residual life for the process $\tilde{N}(t)$, given approximately by (9.64).

Suppose for instance that S = 20.0, s = 2.5 and $f_O(x) = 2x$ for $0 < x < 1$.[2] Then from (9.74),

$$ER = \frac{E(O^2)}{2EO} = \frac{3}{8} \tag{9.75}$$

9.6.6 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.

[2] Actually, the values of S and s do not matter here, though of course the larger S-s is, the better the approximation.

Then from Equation (9.65), we have

$$f_A(w) = \frac{1-\frac{w}{3}}{1.5} = \frac{2}{3} - \frac{2}{9}w \tag{9.76}$$

Then

$$P(A > 1) = \int_1^3 \left(\frac{2}{3} - \frac{2}{9}w \right) \, dw = \frac{4}{9} \tag{9.77}$$

9.6.7 Example: Event Sets in Discrete Event Simulation (difficult)

Discrete event simulation involves systems whose states change in a discrete rather than a continuous manner. For example, the number of packets currently waiting in a network router changes discretely, while the air temperature in a weather model changes continuously.

A discrete event simulation program must maintain an **event set**, which is a data structure containing all the pending events. To make this concrete, suppose we are simulating a k-server queue.[3] There are two kinds of events—job service completions and customer arrivals. Since we have k servers, at any time in the simulation we could have as many as k+1 pending events. If k = 2, for instance, our event set could be, say, consist of a service completion for Machine 1 at time 124.3, one for Machine 2 at 95.4, and a customer arrival at time 99.0.

The core of the program consists of a main loop that repeatedly loops through the following:

(a) Delete the earliest member of the event set.

(b) Update simulated time to the time for that event.

(c) Generate a new event triggered by the one just executed.

(d) Add the new event to the event set.

In our k-server queue example, say the event in (a) is a service completion. Then if there are jobs waiting in the queue, this will trigger the start of service for the new job at the head of the queue, in (c).

Due to (a), the event list is a priority queue, and thus any of the wealth of data structures for priority queues could be used to implement it. Here we will assume the simplest one, a linear linked list, which is always maintained in sorted order.

[3] For the details, using the SimPy language, see my introduction to discrete event simulation, at `http://heather.cs.ucdavis.edu/~matloff/156/PLN/DESimIntro.pdf`.

The question is this: In (d) above, we need to search the list to determine where to insert our new event in order to enhance our program's run speed. Should we start our search at the head of the list or at its tail?

An answer to this question was provided in an old paper: On the Distribution of Event Times for the Notices in a Simulation Event List, Jean Vaucher, *INFOR*, June 1977. Vaucher realized that this problem was right up renewal theory's alley. Our presentation here is adapted from that paper.

First, think of all events that are ever generated during the entire simulation. They comprise the renewal process. Let L_i denote the (simulated) time between the $(i-1)^{st}$ and i^{th} events.

Let t_c denote the time of the event in (a), and let t_d denote earliest time in the event list *after* step (a) is performed. Let L denote the duration of the event generated in (c), so that that event's simulated occurrence time will be $t_c + L$.

We will assume that the the size of the event list stays constant at r. (This is true for many simulations, and approximately true for many others.) Now, the question of where in the event list to place our new event is equivalent to asking how many events in the list have their times t_e after $t_c + L$. That in turn is the same as asking how many of the events have a residual life Z of greater than L. Call that number K.

The quantity of interest is

$$\gamma = \frac{E(K)}{r} \tag{9.78}$$

If $\gamma < 0.5$ we should start our search at the head of the list; otherwise we should start at the tail.

So, let's derive E(K). First write

$$E(K) = E[E(K|L)] \tag{9.79}$$

To simplify notation, let g denote the residual life density (9.64). Then, using the point above concerning residual life and keeping in mind that L is a "constant' in our conditional computation here, we have

$$E(K|L) = rP(Z > L) = r \cdot \int_L^\infty g(t)\ dt \tag{9.80}$$

Now use this in (9.79):

$$E(K) = r \int_0^\infty \left(\int_s^\infty g(t)\ dt \right) f_L(s)\ ds \tag{9.81}$$

One then does an integration by parts, making use of the fact that $g\prime(z) = -\frac{f(x)}{E(L)}$. After all the dust clears, it turns out that

$$\gamma = 1 - E(L) \int_0^\infty g^2(t)\, dt \tag{9.82}$$

In Vaucher's paper, he then evaluated this expression for various distributions of L, and found that for most of the distributions he tried, γ was in the 0.6 to 0.8 range. In other words, one typically should start the search from the tail end.

9.6.8 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.[4] 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

$$s(\tau) = \lim_{t \to \infty} E[S(t, \tau)] \tag{9.83}$$

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 (9.73).

Accordingly, let $A_i(t)$ be the age at time t for page i. Then

> Page i is in the working set if and only if it has been accessed after time $t - \tau$., i.e. $A_i(t) < \tau$.

Thus, using (9.73) and letting 1_i be 1 or 0 according to whether or not $A_i(t) < \tau$, we have that

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

$$\begin{aligned}
s(\tau) &= \lim_{t\to\infty} E(\sum_{i=1}^{n} 1_i) \\
&= \lim_{t\to\infty} \sum_{i=1}^{n} P(A_i(t) < \tau) \\
&= \sum_{i=1}^{n} \sum_{j=0}^{\tau-1} \frac{1-F_i(j)}{E(L_i)} \qquad (9.84)
\end{aligned}$$

Exercises

1. Consider a renewal process in which lifetimes have the values 1, 2, 3 or 4, with probability 1/4 each.

(a) Find P[N(3) = 2].

(b) For large integer t, find the probability that the current renewal period began at t-2.

2. Suppose the burn time for light bulbs is normally distributed with mean 100.0 and variance 225, and consider the associated renewal process (the lamp's bulb is replaced as soon as it burns out). Find $P[N(500) \geq 4]$.

3. Consider a renewal process in which lifetimes have the values 1, 2, 3 or 4, with probability 1/4 each. Express your answers as single fractions, e.g. 25/6 (not as decimal numbers or numerical algebraic expressions).

(a) Find P[N(3) = 2].

(b) For large integer t, find the probability that the current renewal period began at t-2.

Chapter 10

Introduction to Queuing Models

10.1 Introduction

Like other areas of applied stochastic processes, queuing theory has a vast literature, covering a huge number of variations on different types of queues. Our tutorial here can only just scratch the surface to this field.

Here is a rough overview of a few of the large categories of queuing theory:

- Single-server queues.
- Networks of queues, including **open** networks (in which jobs arrive from outside the network, visit some of the servers in the network, then leave) and **closed** networks (in which jobs continually circulate within the network, never leaving).
- Non-First Come, First Served (FCFS) service orderings. For example, there are Last Come, First Served (i.e. stacks) and Processor Sharing (which models CPU timesharing).

In this brief introduction, we will not discuss non-FCFS queues, and will only scratch the surface on the other tops.

10.2 M/M/1

The first M here stands for "Markov" or "memoryless," alluding to the fact that arrivals to the queue are Markovian, i.e. interarrivals are i.i.d. exponentially distributed. The second M means that the service times are also i.i.d. exponential. Denote the reciprocal-mean interarrival and service times by λ and μ.

The 1 in M/M/1 refers to the fact that there is a single server. We will assume FCFS job scheduling here, but close inspection of the derivation will show that it applies to some other kinds of scheduling too.

This system is a continuous-time Markov chain, with the state X_t at time t being the number of jobs in the system (not just in the queue but also including the one currently being served, if any).

10.2.1 Steady-State Probabilities

Intuitively the steady-state probabilities π_i will exist only if $\lambda < \mu$. Otherwise jobs would come in faster than they could be served, and the queue would become infinite. So, we assume that $u < 1$, where $u = \frac{\lambda}{\mu}$.

Clearly this is a birth-and-death chain. For state k, the birth rate $\rho_{k,k+1}$ is λ and the death rate $\rho_{k,k-1}$ is μ, k = 0,1,2,... (except that the death rate at state 0 is 0). Using the formula derived for birth/death chains, we have that

$$\pi_i = u^i \pi_0, \ i \geq 0 \tag{10.1}$$

and

$$\pi_0 = \frac{1}{\sum_{j=0}^{\infty} u^j} = 1 - u \tag{10.2}$$

In other words,

$$\pi_i = u^i (1-u), \ i \geq 0 \tag{10.3}$$

Note by the way that since $\pi_0 = 1 - u$, then u is the *utilization* of the server, i.e. the proportion of the time the server is busy. In fact, this can be seen intuitively: Think of a very long period of time of length t. During this time approximately λt jobs having arrived, keeping the server busy for approximately $\lambda t \cdot \frac{1}{\mu}$ time. Thus the fraction of time during which the server is busy is approximantely

$$\frac{\lambda t \cdot \frac{1}{\mu}}{t} = \frac{\lambda}{\mu} \tag{10.4}$$

10.2.2 Mean Queue Length

Another way to look at Equation (10.3) is as follows. Let the random variable N have the long-run distribution of X_t, so that

$$P(N = i) = u^i(1-u), \ i \geq 0 \tag{10.5}$$

Then this says that N+1 has a geometric distribution, with "success" probability 1-u. (N itself is not quite geometrically distributed, since N's values begin at 0 while a geometric distribution begins at 1.)

Thus the long-run average value E(N) for X_t will be the mean of that geometric distribution, minus 1, i.e.

$$EN = \frac{1}{1-u} - 1 = \frac{u}{1-u} \tag{10.6}$$

The long-run mean queue length E(Q) will be this value minus the mean number of jobs being served. The latter quantity is $1 - \pi_0 = u$, so

$$EQ = \frac{u^2}{1-u} \tag{10.7}$$

10.2.3 Distribution of Residence Time/Little's Rule

Let R denote the **residence time** of a job, i.e. the time elapsed from the job's arrival to its exit from the system. Little's Rule says that

$$EN = \lambda ER \tag{10.8}$$

This property holds for a variety of queuing systems, including this one. It can be proved formally, but here is the intuition:

Think of a particular job (in the literature of queuing theory, it is called a "tagged job") at the time it has just exited the system. If this is an "average" job, then it has been in the system for ER amount of time, during which an average of λER new jobs have arrived behind it. These jobs now comprise the total number of jobs in the system, which in the average case is EN.

Applying Little's Rule here, we know EN from Equation (10.6), so we can solve for ER:

$$ER = \frac{1}{\lambda}\frac{u}{1-u} = \frac{1/\mu}{1-u} \tag{10.9}$$

With a little more work, we can find the actual distribution of R, not just its mean. This will enable us to obtain quantities such as Var(R) and P(R > z). Here is our approach:

When a job arrives, say there are N jobs ahead of it, including one in service. Then this job's value of R can be expressed as

$$R = S_{self} + S_{1,resid} + S_2 + ... + S_N \tag{10.10}$$

where S_{self} is the service time for this job, $S_{1,resid}$ is the remaining time for the job now being served (i.e. the residual life), and for $i > 1$ the random variable S_i is the service time for the i^{th} waiting job.

Then the Laplace transform of R, evaluated at say w, is

$$\begin{aligned} E(e^{-wR}) &= E[e^{-w(S_{self}+S_{1,resid}+S_2+...+S_N)}] && (10.11)\\ &= E\left(E[e^{-w(S_{self}+S_{1,resid}+S_2+...+S_N)}|N]\right) && (10.12)\\ &= E[\{E(e^{-wS})\}^{N+1}] && (10.13)\\ &= E[g(w)^{N+1}] && (10.14) \end{aligned}$$

where

$$g(w) = E(e^{-wS}) \tag{10.15}$$

is the Laplace transform of the service variable, i.e. of an exponential distribution with parameter equal to the service rate μ. Here we have made use of these facts:

- The Laplace transform of a sum of independent random variables is the product of their individual Laplace transforms.
- Due to the memoryless property, $S_{1,resid}$ has the same distribution as do the other S_i.
- The distribution of the service times S_i and queue length N observed by our tagged job is the same as the distributions of those quantities at all times, not just at arrival times of tagged jobs. This property can be proven for this kind of queue and many others, and is called PASTA—Poisson Arrivals See Time Averages.

  (Note that the PASTA property is not obvious. On the contrary, given our experience with the Bus Paradox and length-biased sampling in Section 2.5, we should be wary of such things. But the PASTA property does hold and can be proven.)

But that last term in (10.14), $E[g(w)^{N+1}]$, is the generating function of N+1, evaluated at g(w). And we know from Section 10.2.2 that N+1 has a geometric distribution. The generating function for a nonnegative-integer valued random variable K with success probability p is

$$g_K(s) = E(s^K) = \sum_{i=1}^{\infty} s^i (1-p)^{i-1} p = \frac{ps}{1-s(1-p)} \tag{10.16}$$

In (10.14), we have p = 1 -u and s = g(w). So,

$$E(v^{N+1}) = \frac{g(w)(1-u)}{1-u[g(w)]} \tag{10.17}$$

Finally, by definition of Laplace transform,

$$g(w) = E(e^{-wS}) = \int_0^\infty e^{-wt} \mu e^{-\mu t} dt = \frac{\mu}{w+\mu} \tag{10.18}$$

So, from (10.11), (10.17) and (10.18), the Laplace transform of R is

$$\frac{\mu(1-u)}{w+\mu(1-u)} \tag{10.19}$$

In principle, Laplace transforms can be inverted, and we could use numerical methods to retrieve the distribution of R from (10.19). But hey, look at that! Equation (10.19) has the same form as (10.18). In other words, we have discovered that R has an exponential distribution too, only with parameter $\mu(1-u)$ instead of μ.

This is quite remarkable. The fact that the service and interarrival times are exponential doesn't mean that everything else will be exponential too, so it is surprising that R does turn out to have an exponential distribution.

It is even more surprising in that R is a sum of independent exponential random variables, as we saw in (10.10), and we know that such sums have Erland distributions. The resolution of this seeming paradox is that the number of terms N in (10.10) is itself random. Conditional on N, R has an Erlang distribution, but unconditionally R has an exponential distribution.

10.3 Multi-Server Models

Here we have c servers, with a common queue. There are many variations.

10.3.1 M/M/c

Here the servers are homogeneous. When a job gets to the head of the queue, it is served by the first available server.

The state is again the number of jobs in the system, including any jobs at the servers. Again it is a birth/death chain, with $u_{i,i+1} = \lambda$ and

$$u_{i,i-1} = \begin{cases} i\mu, & \text{if } 0 < \text{i} < \text{c} \\ c\mu, & \text{if i} \geq \text{c} \end{cases} \tag{10.20}$$

The solution turns out to be

$$\pi_k = \begin{cases} \pi_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}, & \text{k} < c \\ \pi_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{c! c^{k-c}}, & \text{k} \geq c \end{cases} \tag{10.21}$$

where

$$\pi_0 = \left[\sum_{k=0}^{c-1} \frac{(cu)^k}{k!} + \frac{(cu)^c}{c!} \frac{1}{1-u} \right]^{-1} \tag{10.22}$$

and

$$u = \frac{\lambda}{c\mu} \tag{10.23}$$

Note that the latter quantity is still the utilization per server, using an argument similar to that which led to (10.4).

Recalling that the Taylor series for e^z is $\sum_{k=0}^{\infty} z^k/k!$ we see that

$$\pi_0 \approx e^{-cu} \tag{10.24}$$

10.3.2 M/M/2 with Heterogeneous Servers

Here the servers have different rates. We'll treat the case in which c = 2. Assume $\mu_1 < \mu_2$. When a job reaches the head of the queue, it chooses machine 2 if that machine is idle, and otherwise waits for the first available machine. Once it starts on a machine, it cannot be switched to the other.

Denote the state by (i,j,k), where

- i is the number of jobs at server 1
- j is the number of jobs at server 2
- k is the number of jobs in the queue

The key is to notice that states 111, 112, 113, ... act like the M/M/k queue. This will reduce finding the solution of the balance equations to solving a finite system of linear equations.

For $k \geq 1$ we have

$$(\lambda + \mu_1 + \mu_2)\pi_{11k} = (\mu_1 + \mu_2)\pi_{11,k+1} + \lambda\pi_{11,k-1} \tag{10.25}$$

Collecting terms as in the derivation of the stationary distribution for birth/death processes, (10.25) becomes

$$\lambda(\pi_{11k} - \pi_{11,k-1}) = (\mu_1 + \mu_2)(\pi_{11,k+1} - \pi_{11k}),\ k = 1, 2, ... \tag{10.26}$$

Then we have

$$(\mu_1 + \mu_2)\pi_{11,k+1} - \lambda\pi_{11k} = (\mu_1 + \mu_2)\pi_{11k} - \lambda\pi_{11,k-1} \tag{10.27}$$

So, we now have all the π_{11i}, i = 2,3,... in terms of π_{111} and π_{110}, thus reducing our task to solving a finite set of linear equations, as promised. Here are the rest of the equations:

$$\lambda\pi_{000} = \mu_2\pi_{010} + \mu_1\pi_{100} \tag{10.28}$$

$$(\lambda + \mu_2)\pi_{010} = \lambda\pi_{000} + \mu_1\pi_{110} \tag{10.29}$$

$$(\lambda + \mu_1)\pi_{100} = \mu_2\pi_{110} \tag{10.30}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{110} = \lambda\pi_{010} + \lambda\pi_{100} + (\mu_1 + \mu_2)\pi_{111} \tag{10.31}$$

From (1.62), we have

$$(\mu_1 + \mu_2)\pi_{111} - \lambda\pi_{110} = (\mu_1 + \mu_2)\pi_{110} - \lambda(\pi_{010} + \pi_{100}) \tag{10.32}$$

Look at that last term, $\lambda(\pi_{010} + \pi_{100})$. By adding (10.29) and (10.30), we have that

$$\lambda(\pi_{010} + \pi_{100}) = \lambda\pi_{000} + \mu_1\pi_{110} + \mu_2\pi_{110} - \mu_1\pi_{100} - \mu_2\pi_{010} \tag{10.33}$$

Substituting (10.28) changes (10.33) to

$$\lambda(\pi_{010} + \pi_{100}) = \mu_1\pi_{110} + \mu_2\pi_{110} \tag{10.34}$$

So...(10.32) becomes

$$(\mu_1 + \mu_2)\pi_{111} - \lambda\pi_{110} = 0 \tag{10.35}$$

By induction in (10.27), we have

$$(\mu_1 + \mu_2)\pi_{11,k+1} - \lambda\pi_{11k} = 0,\ k = 1, 2, ... \tag{10.36}$$

and

$$\pi_{11i} = \delta^i\pi_{110},\ i = 0, 1, 2, ... \tag{10.37}$$

where

$$\delta = \frac{\lambda}{\mu_1 + \mu_2} \tag{10.38}$$

$$\begin{aligned}
1 &= \sum_{i,j,k} \pi_{ijk} && (10.39)\\
&= \pi_{000} + \pi_{010} + \pi_{100} + \sum_{i=0}^{\infty} \pi_{11i} && (10.40)\\
&= \pi_{000} + \pi_{010} + \pi_{100} + \pi_{110} \sum_{i=0}^{\infty} \delta^i && (10.41)\\
&= \pi_{000} + \pi_{010} + \pi_{100} + \pi_{110} \cdot \frac{1}{1-\delta} && (10.42)
\end{aligned}$$

Finding close-form expressions for the π_i is then straightforward.

10.4 Loss Models

One of the earliest queuing models was M/M/c/c: Markovian interarrival and service times, c servers and a buffer space of c jobs. Any job which arrives when c jobs are already in the system is lost. This was used by telephone companies to find the proportion of lost calls for a bank of c trunk lines.

10.4.1 Cell Communications Model

Let's consider a more modern example of this sort, involving cellular phone systems. (This is an extension of the example treated in K.S. Trivedi, *Probability and Statistics, with Reliability and Computer Science Applications* (second edition), Wiley, 2002, Sec. 8.2.3.2, which is in turn is based on two papers in the *IEEE Transactions on Vehicular Technology*.)

We consider one particular cell in the system. Mobile phone users drift in and out of the cell as they move around the city. A call can either be a **new call**, i.e. a call which someone has just dialed, or a **handoff call**, i.e. a call which had already been in progress in a neighboring cell but now has moved to this cell.

Each call in a cell needs a **channel**.[1] There are n channels available in the cell. We wish to give handoff calls priority over new calls.[2] This is accomplished as follows.

The system always reserves g channels for handoff calls. When a request for a new call (i.e. a non-handoff call) arrives, the system looks at X_t, the current number of calls in the cell. If that number is less than n-g, so that there are more than g idle channels available, the new call is accepted; otherwise it is rejected.

We assume that new calls originate from within the cells according to a Poisson process with rate λ_1, while handoff calls drift in from neighboring cells at rate λ_2. Meanwhile, call durations are exponential with rate μ_1, while the time that a call remains within the cell is exponential with rate μ_2.

10.4.1.1 Stationary Distribution

We again have a birth/death process, though a bit more complicated than our earlier ones. Let $\lambda = \lambda_1 + \lambda_2$ and $\mu = \mu_1 + \mu_2$. Then here is a sample balance equation, focused on transitions into (left-hand side in the equation) and out of (right-hand side) state 1:

$$\pi_0 \lambda + \pi_2 2\mu = \pi_1(\lambda + \mu) \tag{10.43}$$

Here's why: How can we enter state 1? Well, we could do so from state 0, where there are no calls; this

[1] This could be a certain frequency or a certain time slot position.

[2] We would rather give the caller of a new call a polite rejection message, e.g. "No lines available at this time, than suddenly terminate an existing conversation.

occurs if we get a new call (rate λ_1) or a handoff call (rate λ_2. In state 2, we enter state 1 if one of the two calls ends (rate μ_1) or one of the two calls leaves the cell (rate μ_2). The same kind of reasoning shows that we leave state 1 at rate $\lambda + \mu$.

As another example, here is the equation for state n-g:

$$\pi_{n-g}[\lambda_2 + (n-g)\mu] = \pi_{n-g+1} \cdot (n-g+1)\mu + \pi_{n-g-1}\lambda \tag{10.44}$$

Note the term λ_2 in (10.44), rather than λ as in (10.43).

Using our birth/death formula for the π_i, we find that

$$\pi_k = \begin{cases} \pi_0 \frac{A^k}{k!}, & \text{k} \leq \text{n-g} \\ \pi_0 \frac{A^{n-g}}{k!} A_1^{k-(n-g)}, & \text{k} \geq \text{n-g} \end{cases} \tag{10.45}$$

where $A = \lambda/\mu$, $A_1 = \lambda_2/\mu$ and

$$\pi_0 = \left[\sum_{k=0}^{n-g-1} \frac{A^k}{k!} + \sum_{k=n-g}^{n} \frac{A^{n-g}}{k!} A_1^{k-(n-g)} \right]^{-1} \tag{10.46}$$

10.4.1.2 Going Beyond Finding the π

One can calculate a number of interesting quantities from the π_i:

- The probability of a handoff call being rejected is π_n.
- The probability of a new call being dropped is

$$\sum_{k=n-g}^{n} \pi_k \tag{10.47}$$

- Since the per-channel utilization in state i is i/n, the overall long-run per-channel utilization is

$$\sum_{i=0}^{n} \pi_i \frac{i}{n} \tag{10.48}$$

- The long-run proportion of accepted calls which are handoff calls is the rate at which handoff calls are accepted, divided by the rate at which calls are accepted:

$$\frac{\lambda_2 \sum_{i=0}^{n-1} \pi_i}{\lambda_1 \sum_{i=0}^{n-g-1} \pi_i + \lambda_2 \sum_{i=0}^{n-1} \pi_i} \tag{10.49}$$

10.5 Nonexponential Service Times

The Markov property is of course crucial to the analyses we made above. Thus dropping the exponential assumption presents a major analytical challenge.

One queuing model which has been found tractable is M/G/1: Exponential interarrival times, general service times, one server. In fact, the mean queue length and related quantities can be obtained fairly easily, as follows.

Consider the residence time R for a tagged job. R is the time that our tagged job must first wait for completion of service of all jobs, if any, which are ahead of it—queued or now in service—plus the tagged job's own service time. Let $T_1, T_2, ...$ be i.i.d. with the distribution of a generic service time random variable S. T_1 represents the service time of the tagged job itself. $T_2, ..., T_N$ represent the service times of the queued jobs, if any.

Let N be the number of jobs in the system, either being served or queued; B be either 1 or 0, depending on whether the system is busy (i.e. N $>$ 0) or not; and $S_{1,resid}$ be the remaining service time of the job currently being served, if any. Finally, we define, as before, $u = \frac{\lambda}{1/ES}$, the utilization. Note that that implies the EB = u.

Then the distribution of R is that of

$$BS_{1,resid} + (T_1 + ... + T_N) + (1 - B)T_1 \tag{10.50}$$

Note that if N = 0, then $T_1 + ... + T_N$ is considered to be 0, i.e. not present in (10.50).

Then

$$\begin{aligned} E(R) &= uE(S_{1,resid}) + E(T_1 + ... + T_N) + (1 - u)ET_1 && (10.51) \\ &= uE(S_{1,resid}) + E(N)E(S) + (1 - u)ES && (10.52) \\ &= uE(S_{1,resid}) + \lambda E(R)E(S) + (1 - u)ES && (10.53) \end{aligned}$$

The last equality is due to Little's Rule. Note also that we have made use of the PASTA property here, so that the distribution of N is the same at arrival times as general times.

Then

$$E(R) = \frac{uE(S_{1,resid})}{1-u} + ES \tag{10.54}$$

Note that the two terms here represent the mean residence time as the mean queuing time plus the mean service time.

So we must find $E(S_{1,resid})$. This is just the mean of the remaining-life distribution which we saw in Section 9.6 of our unit on renewal theory. Then

$$\begin{aligned}
E(S_{1,resid}) &= \int_0^\infty t \frac{1-F_S(t)}{ES} \, dt & (10.55)\\
&= \frac{1}{ES} \int_0^\infty t \int_t^\infty f_S(u) \, du \, dt & (10.56)\\
&= \frac{1}{ES} \int_0^\infty f_S(u) \int_0^u t \, dt \, du & (10.57)\\
&= \frac{1}{2ES} E(S^2) & (10.58)
\end{aligned}$$

So,

$$E(R) = \frac{uE(S^2)}{2ES(1-u)} + ES \tag{10.59}$$

What is remarkable about this famous formula is that E(R) depends not only on the mean service time but also on the variance. This result, which is not so intuitively obvious at first glance, shows the power of modeling. We might observe the dependency of E(R) on the variance of service time empirically if we do simulation, but here is a compact formula that shows it for us.

10.6 Reversed Markov Chains

We can get insight into some kinds of queuing systems by making use of the concepts of **reversed** Markov chains, which involve "playing the Markov chain backward," just as we could play a movie backward.

Consider a continuous-time, irreducible, positive recurrent Markov chain X(t).[3] For any fixed time τ (typ-

[3] Recall that a Markov chain is irreducible if it is possible to get from each state to each other state in a finite number of steps, and that the term *positive recurrent* means that the chain has a long-run state distribution π. Also, concerning our assumption here of continuous time, we should note that there are discrete-time analogs of the various points we'll make below.

ically thought of as large), define the **reversed** version of X(t) as Y(t) = X(τ-t), for $0 \leq t \leq \tau$. We will discuss a number of properties of reversed chains. These properties will enable what mathematicians call "soft analysis" of some Markov chains, especially those related to queues. This term refers to short, simple, elegant proofs or derivations.

10.6.1 Markov Property

The first property to note is that Y(t) is a Markov chain! Here is our first chance for soft analysis.

The "hard analysis" approach would be to start with the definition, which in continuous time would be that

$$P\left(Y(t)=k|Y(u), u \leq s\right)=P\left(Y(t)=k|Y(s)\right) \tag{10.60}$$

for all $0 < s < t$ and all k, using the fact that X(t) has the same property. That would involve making substitutions in Equation (10.60) like Y(t) = X(τ-t), etc.

But it is much easier to simply observe that the Markov property holds if and only if, conditional on the present, the past and the future are independent. Since that property holds for X(t), it also holds for Y(t) (with the roles of the "past" and the "future" interchanged).

10.6.2 Long-Run State Proportions

Clearly, if the long-run proportion of the time X(t) = k is π_i, the same long-run proportion will hold for Y(t). This of course only makes sense if you think of larger and larger τ.

10.6.3 Form of the Transition Rates of the Reversed Chain

Let $\tilde{\rho}_{ij}$ denote the number of transitions from state i to state j per unit time in the reversed chain. That number must be equal to the number of transitions from j to i in the original chain. Therefore,

$$\pi_i \tilde{\rho}_{ij}=\pi_j \rho_{ji} \tag{10.61}$$

This gives us a formula for the $\tilde{\rho}_{ij}$:

$$\tilde{\rho}_{ij}=\frac{\pi_j}{\pi_i} \rho_{ji} \tag{10.62}$$

10.6.4 Reversible Markov Chains

In some cases, the reversed chain has the same probabilistic structure as the original one! Note carefully what that would mean. In the continuous-time case, it would mean that $\tilde{\rho}_{ij} = \rho_{ij}$ for all i and j, where the $\tilde{\rho}_{ij}$ are the transition rates of Y(t).[4] If this is the case, we say that X(t) is **reversible**.

That is a very strong property. An example of a chain which is not reversible is the tree-search model in Section 8.4.5.[5] There the state space consists of all the nonnegative integers, and transitions were possible from states n to n+1 and from n to 0. Clearly this chain is not reversible, since we can go from n to 0 in one step but not vice versa.

10.6.4.1 Conditions for Checking Reversibility

Equation (10.61) shows that the original chain X(t) is reversible if and only if

$$\pi_i \rho_{ij} = \pi_j \rho_{ji} \tag{10.63}$$

for all i and j. These equations are called the **detailed balance equations**, as opposed to the general **balance equations**,

$$\sum_{j \neq i} \pi_j \rho_{ji} = \pi_i \lambda_i \tag{10.64}$$

which are used to find the π values. Recall that (10.64) arises from equating the flow into state i with the flow out of it. By contrast, Equation (10.63) equates the flow into i from a particular state j to the flow from i to j. Again, that is a much stronger condition, so we can see that most chains are not reversible. However, a number of important ones are reversible, as we'll see.

For example, consider birth/death chains. Here, the only cases in which ρ_{rs} is nonzero are those in which $|i-j| = 1$. Now, Equation (8.72) in our derivation of π for birth/death chains is exactly (10.63)! So we see that birth/death chains are reversible.

More generally, equations (10.63) may not be so easy to check, since for complex chains we may not be able to find closed-form expressions for the π values. Thus it is desirable to have another test available for reversibility. One such test is **Kolmogorov's Criterion**:

[4] Note that for a continuous-time Markov chain, the transition rates do indeed uniquely determined the probabilistic structure of the chain, not just the long-run state proportions. The short-run behavior of the chain is also determined by the transition rates, and at least in theory can be calculated by solving differential equations whose coefficients make use of those rates.

[5] That is a discrete-time example, but the principle here is the same.

> The chain is reversible if and only if for any **loop** of states, the product of the transition rates is the same in both the forward and backward directions.

For example, consider the loop $i \rightarrow j \rightarrow k \rightarrow i$. Then we would check whether $\rho_{ij}\rho_{jk}\rho_{ki} = \rho_{ik}\rho_{kj}\rho_{ji}$.

Technically, we do have to check *all* loops. However, in many cases it should be clear that just a few loops are representative, as the other loops have the same structure.

Again consider birth/death chains. Kolmogorov's Criterion trivially shows that they are reversible, since any loop involves a path which is the same path when traversed in reverse.

10.6.4.2 Making New Reversible Chains from Old Ones

Since reversible chains are so useful (when we are lucky enough to have them), a very useful trick is to be able to form new reversible chains from old ones. The following two properties are very handy in that regard:

(a) Suppose U(t) and V(t) are reversible Markov chains, and define W(t) to be the tuple [U(t),V(t)]. Then W(t) is reversible.

(b) Suppose X(t) is a reversible Markov chain, and A is an irreducible subset of the state space of the chain, with long-run state distribution π. Define a chain W(t) with transition rates $\rho\prime_{ij}$ for $i \in A$, where $\rho\prime_{ij} = \rho_{ij}$ if $j \in A$ and $\rho\prime_{ij} = 0$ otherwise. Then W(t) is reversible, with long-run state distribution given by

$$\pi\prime_i = \frac{\pi_i}{\sum_{j \in A} \pi_j} \tag{10.65}$$

10.6.4.3 Example: Distribution of Residual Life

In Section 9.6.2, we used Markov chain methods to derive the age distribution at a fixed observation point in a renewal process. From remarks made there, we know that residual life has the same distribution. This could be proved similarly, at some effort, but it comes almost immediately from reversibility considerations. After all, the residual life in the reversed process is the age in the original process.

10.6.4.4 Example: Queues with a Common Waiting Area

Consider two M/M/1 queues, with chains G(t) and H(t), with independent arrival streams but having a common waiting area, with jobs arriving to a full waiting area simply being lost.[6]

First consider the case of an infinite waiting area. Let u_1 and u_2 be the utilizations of the two queues, as in (10.3). G(t) and H(t), being birth/death processes, are reversible. Then by property (a) above, the chain [G(t),H(t)] is also reversible. Long-run proportion of the time that there are m jobs in the first queue and n jobs in the second is

$$\pi_{mn} = (1-u_1)^m u_1 (1-u_2)^n u_2 \tag{10.66}$$

for m,n = 0,1,2,3,...

Now consider what would happen if these two queues were to have a common, finite waiting area. Denote the amount of space in the waiting area by w. The new process is the restriction of the original process to a subset of states A as in (b) above. (The set A will be precisely defined below.) It is easily verified from the Kolmogorov Criterion that the new process is also reversible.

Recall that the state m in the original queue U(t) is the number of jobs, including the one in service if any. That means the number of jobs waiting is $(m-1)^+$, where $x^+ = \max(x,0)$. That means that for our new system, with the common waiting area, we should take our subset A to be

$$\{(m,n) : m,n \geq 0, (m-1)^+ + (n-1)^+ \leq w\} \tag{10.67}$$

So, by property (b) above, we know that the long-run state distribution for the queue with the finite common waiting area is

$$\pi_{mn} = \frac{1}{a}(1-u_1)^m u_1 (1-u_2)^n u_2 \tag{10.68}$$

where

$$a = \sum_{(i,j) \in A} (1-u_1)^i u_1 (1-u_2)^j u_2 \tag{10.69}$$

In this example, reversibility was quite useful. It would have been essentially impossible to derive (10.68) algebraically. And even if intuition had suggested that solution as a guess, it would have been quite messy to verify the guess.

[6] Adapted from Ronald Wolff, *Stochastic Modeling and the Theory of Queues* Prentice Hall, 1989.

10.6.4.5 Closed-Form Expression for π for Any Reversible Markov Chain

(Adapted from Ronald Nelson, *Probability, Stochastic Processes and Queuing Theory*, Springer-Verlag, 1995.)

Recall that most Markov chains, especially those with infinite state spaces, do not have closed-form expressions for the steady-state probabilities. But we can always get such expressions for reversible chains, as follows.

Choose a fixed state s, and find paths from s to all other states. Denote the path to i by

$$s = j_{i1} \rightarrow j_{i2} \rightarrow ... \rightarrow j_{im_i} = i \tag{10.70}$$

Define

$$\psi_i = \begin{cases} 1, & \text{i = s} \\ \Pi_{k=1}^{m_i} r_{ik}, & \text{i} \neq \text{s} \end{cases} \tag{10.71}$$

where

$$r_{ik} = \frac{\rho(j_{ik}, j_{i,k+1})}{\rho(j_{i,k+1}, j_{i,k})} \tag{10.72}$$

Then the steady-state probabilities are

$$\pi_i = \frac{\psi_i}{\sum_k \psi_k} \tag{10.73}$$

You may notice that this looks similar to the derivation for birth/death processes, which as has been pointed out, are reversible.

10.7 Networks of Queues

10.7.1 Tandem Queues

Let's first consider an M/M/1 queue. As mentioned earlier, this is a birth/death process, thus reversible. This has an interesting and very useful application, as follows.

Think of the times at which jobs *depart* this system, i.e. the times at which jobs finish service. In the reversed process, these times are *arrivals*. Due to the reversibility, that means that the distribution of departure times is the same as that of arrival times. In other words:

- Departures from this system behave as a Poisson process with rate λ.

Also, let the initial state X(0) be distributed according to the steady-state probabilities π.[7] Due to the PASTA property of Poisson arrivals, the distribution of the system state at arrival times is the same as the distribution of the system state at nonrandom times t. Then by reversibility, we have that:

- The state distribution at departure times is the same as at nonrandom times.

And finally, noting as in Section 10.6.1 that, given X(t), the states $\{X(s), s \leq t\}$ of the queue before time t are statistically independent of the arrival process after time t, reversibility gives us that:

- Given t, the departure process before time t is statistically independent of the states $\{X(s), s \geq t\}$ of the queue after time t.

Let's apply that to **tandem** queues, which are queues acting in series. Suppose we have two such queues, with the first, $X_1(t)$ feeding its output to the second one, $X_2(t)$, as input. Suppose the input into $X_1(t)$ is a Poisson process with rate λ, and service times at both queues are exponentially distributed, with rates μ_1 and μ_2.

$X_1(t)$ is an M/M/1 queue, so its steady-state probabilities for $X_1(t)$ are given by Equation (10.3), with $u = \lambda/\mu_1$.

By the first bulleted item above, we know that the input into $X_2(t)$ is also Poisson. Therefore, $X_2(t)$ also is an M/M/1 queue, with steady-state probabilities as in Equation (10.3), with $u = \lambda/\mu_2$.

Now, what about the joint distribution of $[X_1(t), X_2(t)]$? The third bulleted item above says that the input to $X_2(t)$ up to time t is independent of $\{X_1(s), s \geq t\}$. So, using the fact that we are assuming that $X_1(0)$ has the steady-state distribution, we have that

$$P[X_1(t) = i, X_2(t) = j] = (1 - u_1)u_1^i P[X_2(t) = j] \tag{10.74}$$

Now letting $t \to \infty$, we get that the long-run probability of the vector $[X_1(t), X_2(t)]$ being equal to (i,j) is

$$(1 - u_1)u_1^j (1 - u_2)u_2^j \tag{10.75}$$

[7] Recall Section 8.1.2.4.

In other words, the steady-state distribution for the vector has the two components of the vector being independent.

Equation (10.75) is called a **product form solution** to the balance equations for steady-state probabilities.

By the way, the vector $[X_1(t), X_2(t)]$ is *not* reversible.

10.7.2 Jackson Networks

The tandem queues discussed in the last section comprise a special case of what are known as **Jackson networks**. Once again, there exists an enormous literature of Jackson and other kinds of queuing networks. The material can become very complicated (even the notation is very complex), and we will only present an introduction here. Our presentation is adapted from I. Mitrani, *Modelling of Computer and Communcation Systems*, Cambridge University Press, 1987.

Our network consists of N nodes, and jobs move from node to node. There is a queue at each node, and service time at node i is exponentially distributed with mean $1/\mu_i$.

10.7.2.1 Open Networks

Each job originally arrives externally to the network, with the arrival rate at node i being γ_i. After moving among various nodes, the job will eventually leave the network. Specifically, after a job completes service at node i, it moves to node j with probability q_{ij}, where

$$\sum_j q_{ij} < 1 \tag{10.76}$$

reflecting the fact that the job will leave the network altogether with probability $1 - \sum_j q_{ij}$.[8] It is assumed that the movement from node to node is memoryless.

As an example, you may wish to think of movement of packets among routers in a computer network, with the packets being jobs and the routers being nodes.

Let λ_i denote the total traffic rate into node i. By the usual equating of flow in and flow out, we have

$$\lambda_i = \gamma_i + \sum_{j=1}^{N} \lambda_j q_{ji} \tag{10.77}$$

[8] By the way, q_{ii} can be nonzero, allowing for feedback loops at nodes.

Note the in Equations (10.77), the knowns are γ_i and the q_{ji}. We can solve this system of linear equations for the unknowns, λ_i.

The utilization at node i is then $u_i = \lambda_i/\mu_i$, as before. Jackson's Theorem then says that in the long run, node i acts as an M/M/1 queue with that utilization, and that the nodes are independent in the long run:[9]

$$\lim_{t\to\infty} P[X_1(t) = i_1, ..., X_N(t) = i_N] = \Pi_{i=1}^N (1-u_i)u^i \tag{10.78}$$

So, again we have a product form solution.

Let L_i denote the average number of jobs at node i. From Equation (10.6), we have $L_i = u_i/(1-u_i)$. Thus the mean number of jobs in the system is

$$L = \sum_{i=1}^N \frac{u_i}{1-u_i} \tag{10.79}$$

From this we can get the mean time that jobs stay in the network, W: From Little's Rule, $L = \gamma W$, so

$$W = \frac{1}{\gamma} \sum_{i=1}^N \frac{u_i}{1-u_i} \tag{10.80}$$

where $\gamma = \gamma_1 + ... + \gamma_N$ is the total external arrival rate.

Jackson networks are not generally reversible. The reversed versions of Jackson networks are worth studying for other reasons, but we cannot pursue them here.

10.7.3 Closed Networks

In a closed Jackson network, we have for all i, $\gamma_i = 0$ and

$$\sum_j q_{ij} = 1 \tag{10.81}$$

In other words, jobs never enter or leave the network. There have been many models like this in the computer performance modeling literature. For instance, a model might consist of some nodes representing CPUs, some representing disk drives, and some representing users at terminals.

[9] We do not present the proof here, but it really is just a matter of showing that the distribution here satisfies the balance equations.

It turns out that we again get a product form solution.[10] The notation is more involved, so we will not present it here.

Exercises

1. Investigate the robustness of the M/M/1 queue model with respect to the assumption of exponential service times, as follows. Suppose the service time is actually uniformly distributed on (0,c), so that the mean service time would be c/2. Assume that arrivals do follow the exponential model, with mean interarrival time 1.0. Find the mean residence time, using (10.9), and compare it to the true value obtained from (10.59). Do this for various values of c, and graph the two curves using R.

2. Many mathematical analyses of queuing systems use **finite source** models. There are always a fixed number j of jobs in the system. A job queues up for the server, gets served in time S, then waits a random time W before queuing up for the server again.

A typical example would be a file server with j clients. The time W would be the time a client does work before it needs to access the file server again.

(a) Use Little's Rule, on two or more appropriately chosen boxes, to derive the following relation:

$$ER = \frac{jES}{U} - EW \tag{10.82}$$

where R is residence time (time spent in the queue plus service time) in one cycle for a job and U is the utilization fraction of the server.

(b) Set up a continuous time Markov chain, assuming exponential distributions for S and W, with state being the number of jobs currently at the server. Derive closed-form expressions for the π_i.

3. Consider the following variant of an M/M/1 queue: Each customer has a certain amount of patience, varying from one customer to another, exponentially distributed with rate η. When a customer's patience wears out while the customer is in the queue, he/she leaves (but not if his/her job is now in service). Arrival and service rates are λ and ν, respectively.

(a) Express the π_i in terms of λ, ν and η.

(b) Express the proportion of lost jobs as a function of the π_i, λ, ν and η.

4. A shop has two machines, with service time in machine i being exponentially distributed with rate μ_i, i = 1,2. Here $\mu_1 > \mu_2$. When a job reaches the head of the queue, it chooses machine 1 if that machine is

[10] This is confusing, since the different nodes are now not independent, due to the fact that the number of jobs in the overall system is constant.

idle, and otherwise waits for the first available machine. If when a job finishes on machine 1 there is a job in progress at machine 2, the latter job will be transferred to machine 1, getting priority over any queued jobs. Arrivals follow the usual Poisson process, parameter λ.

(a) Find the mean residence time.

(b) Find the proportion of jobs that are originally assigned to machine 2.