Basic Statistical Thought

Michael Lavine

June 12, 2019

Contents

List of Figures

List of Tables

Preface

Basic Statistical Thought is an introductory undergraduate textbook on statistical thought with a likelihood emphasis. By "statistical thought" is meant a focus on ideas that statisticians care about as opposed to technical details of putting those ideas into practice. By "likelihood emphasis" is meant that the likelihood function and likelihood principle are unifying ideas throughout the text. In particular, models are compared by how well they describe the data, not by the usual frequentist criteria.

This book makes heavy use of statistical software for two purposes. One is pedagogical: algorithms and simulations can help readers understand concepts. The other is graphical. It is important to display data graphically to look for what models might be appropriate and the ways in which they might fail.

Our software of choice is R (R Core Team, 2017). R and accompanying manuals are available for free download from HTTP://WWW.R-PROJECT.ORG. You may wish to download *An Introduction to R* from that site to keep as a reference. In addition, there are many gentle introductions to R on the web.

It is highly recommended that you try all the examples in R. They will help you understand concepts, give you a little programming experience, and give you facility with a very flexible statistical software package. And don't just try the examples as written. Vary them a little; play around with them; experiment. You won't hurt anything and you'll learn a lot.

Throughout the text are notes, like this one, with a grey background and smaller font. They are set apart visually to indicate they are not part of the main flow of ideas. Often they are about R and how to accomplish various tasks in R. You can skip them and come back to them later without losing the central ideas.

The text is created with HYPERLINKS IN SMALL CAPS FONT. Hovering your mouse over them shows a bit of other text and clicking them takes you there.

Chapter 1

Probability

1.1 Basic Probability

To study statistics, we have to know a little probability. In this book, probability is denoted by “Pr”. We write expressions like

$$\Pr[\text{a coin lands Heads}]$$

or

$$\Pr[\text{patient 12 develops cancer}].$$

For example, when a die is tossed there are six basic outcomes — 1, 2, . . . , 6 — with probabilities Pr[1], Pr[2], . . . , Pr[6]. A model for a fair die is

$$\Pr[1] = \Pr[2] = \cdots = \Pr[6] = 1/6. \tag{1.1}$$

Example 1.1 shows how (1.1) can be used to calculate probabilities in the game of craps.

Mathematical writing uses notation like “(1.1)” to mean “Equation (1.1)” without having to write “Equation.”

Example 1.1 (The Game of Craps)
Craps is a gambling game played with two dice. Here are the rules, as explained at `WWW.ONLINE-CRAPS-GAMBLING.COM/CRAPS-RULES.HTML`.

> "For the dice thrower (shooter) the object of the game is to throw a 7 or an 11 on the first roll (a win) and avoid throwing a 2, 3 or 12 (a loss). If none of these numbers (2, 3, 7, 11 or 12) is thrown on the first throw (the Come-out roll) then a Point is established (the point is the number rolled) against which the shooter plays. The shooter continues to throw until one of two numbers is thrown, the Point number or a Seven. If the shooter rolls the Point before rolling a Seven he/she wins, however if the shooter throws a Seven before rolling the Point he/she loses."

Craps was popular in the US in the first part of the 20^{th} century. It was central to several Damon Runyon stories, which formed the basis of the musical *Guys and Dolls* by Frank Loesser, Jo Swerling, and Abe Burrows which was later adapted into a movie of the same name starring Marlon Brando, Jean Simmons, Frank Sinatra, and Vivian Blaine.

Ultimately we would like to calculate $\Pr($shooter wins$)$, But for now, let's calculate

$$\Pr(\text{shooter wins on Come-out roll}) = \Pr(7 \text{ or } 11) = \Pr(7) + \Pr(11).$$

Let D_1 be the number on the first die and D_2 the number on the second. D_1 and D_2 are integers from 1 to 6. The possibilities for (D_1, D_2) are the ordered pairs $(1,1)$, $(1,2)$, ..., $(6,5)$, $(6,6)$. More formally we could write

$$(D_1, D_2) \in \left\{ \begin{matrix} (6,6) & (6,5) & (6,4) & (6,3) & (6,2) & (6,1) \\ (5,6) & (5,5) & (5,4) & (5,3) & (5,2) & (5,1) \\ (4,6) & (4,5) & (4,4) & (4,3) & (4,2) & (4,1) \\ (3,6) & (3,5) & (3,4) & (3,3) & (3,2) & (3,1) \\ (2,6) & (2,5) & (2,4) & (2,3) & (2,2) & (2,1) \\ (1,6) & (1,5) & (1,4) & (1,3) & (1,2) & (1,1) \end{matrix} \right\}.$$

The notation $A \in \{B\}$ means that A is an element of the set B.

If the dice are fair, then all pairs are equally likely. Since there are 36 of them, we assign $\Pr(d_1, d_2) = 1/36$ for any combination (d_1, d_2). Finally, we can calculate

$$\begin{aligned} \Pr(7 \text{ or } 11) = {} & \Pr(6,5) + \Pr(5,6) + \Pr(6,1) + \Pr(5,2) \\ & + \Pr(4,3) + \Pr(3,4) + \Pr(2,5) + \Pr(1,6) = 8/36 = 2/9. \end{aligned}$$

Not all dice are fair, and (1.1) is a model that might not accurately represent a particular die. If we suspect a die is loaded so that 6 appears more often (See

HTTPS://WWW.WIKIHOW.COM/LOAD-DICE for how to load dice.) we might consider a model in which

$$
\begin{aligned}
\Pr[1] &< 1/6 \\
\Pr[2] &< 1/6 \\
\Pr[3] &< 1/6 \\
\Pr[4] &< 1/6 \\
\Pr[5] &< 1/6 \\
\Pr[6] &> 1/6.
\end{aligned}
\tag{1.2}
$$

If we cared, we could toss the die many times and compare how well (1.1) and (1.2) describe the data.

Comparing (1.1) and (1.2) is a prototypical statistics problem: assessing several models of the real world according to how well they describe data. Example 1.2 is another.

Example 1.2 (The Slater School)
The Slater school is an elementary school in Fresno, California. Brodeur, 1992 reports that teachers and staff were "concerned about the presence of two high-voltage transmission lines that ran past the school" and whether they contribute to the "high incidence of cancer at Slater."

At the time, some people worried that prolonged exposure to the magnetic fields induced by high-voltage transmission lines increased the risk of cancer. More recent studies have not shown any association between cancer and exposure. According to the National Cancer Institute, (HTTPS://WWW.CANCER.GOV/ABOUT-CANCER/CAUSES-PREVENTION/RISK/RADIATION/ELECTROMAGNETIC-FIELDS-FACT-SHEET) "No mechanism by which ELF-EMFs or radiofrequency radiation could cause cancer has been identified. . . . Studies of animals have not provided any indications that exposure to ELF-EMFs is associated with cancer . . . No consistent evidence for an association between any source of non-ionizing EMF and cancer has been found."

To address their concern, Dr. Raymond Neutra of the California Department of Health Services' Special Epidemiological Studies Program conducted a statistical analysis on the

> "eight cases of invasive cancer, . . . , the total years of employment of the hundred and forty-five teachers, teachers' aides, and staff members, . . . , [and] the number of person-years in terms of National Cancer Institute statistics showing the annual rate of invasive cancer in American women between the ages of forty and forty-four — the age group encompassing the average age of the teachers and staff at Slater — [which] enabled him to calculate that 4.2 cases of cancer could have been expected to occur among the Slater teachers and staff members . . . ."

One model of the Slater school says that all employees have the same probability of developing invasive cancer and that they develop cancer independently of each other, i.e., that whether employee A develops cancer is not affected by whether employee B develops cancer. That model says

$$\begin{aligned}&\mathrm{Pr}[\text{employee 1 develops cancer}] = \mathrm{Pr}[\text{employee 2 develops cancer}] = \cdots \\ &\quad \cdots = \mathrm{Pr}[\text{employee 144 develops cancer}] = \mathrm{Pr}[\text{employee 145 develops cancer}] \end{aligned} \tag{1.3}$$

Eq. (1.3) is an oversimplified model because different employees likely have different probabilities of developing invasive cancer due to their different ages, sexes, years working at the school, etc. Nevertheless, for the purpose of this example, we will continue working with (1.3). However, (1.3) is not yet a fully specified model because it doesn't specify $\mathrm{Pr}[\text{employee } i \text{ develops cancer}]$ for every $i \in \{1, 2, \ldots, 145\}$.

Because "4.2 cases of cancer could have been expected to occur" we estimate the national rate of invasive cancer to be about $4.2/145 \approx 0.03$. Therefore,

$$\begin{aligned}&\mathrm{Pr}[\text{employee 1 develops cancer}] = \mathrm{Pr}[\text{employee 2 develops cancer}] = \cdots \\ &\qquad \cdots = \mathrm{Pr}[\text{employee 145 develops cancer}] = 0.03 \end{aligned} \tag{1.3a}$$

is a fully specified model of interest. Further, since 8 cases did occur and $8/145 \approx 0.055$,

$$\begin{aligned}&\mathrm{Pr}[\text{employee 1 develops cancer}] = \mathrm{Pr}[\text{employee 2 develops cancer}] = \cdots \\ &\qquad \cdots = \mathrm{Pr}[\text{employee 145 develops cancer}] = 0.055 \end{aligned} \tag{1.3b}$$

is another fully specified model of interest.

The next step is to compare how well (1.3A) and (1.3B) describe the data. Figure 1.1 depicts the setting. Because there are 145 employees, the number who develop invasive cancer will be some integer from 0 to 145. Those numbers are on the x-axis. The y-axis shows

$$\begin{aligned}
&\mathrm{Pr}_{(1.3a)}[\text{0 employees develop cancer}]\\
&\mathrm{Pr}_{(1.3a)}[\text{1 employee develops cancer}]\\
&\qquad\qquad\vdots\\
&\mathrm{Pr}_{(1.3a)}[\text{144 employees develop cancer}]\\
&\mathrm{Pr}_{(1.3a)}[\text{145 employees develop cancer}]
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Pr}_{(1.3b)}[\text{0 employees develop cancer}]\\
&\mathrm{Pr}_{(1.3b)}[\text{1 employee develops cancer}]\\
&\qquad\qquad\vdots\\
&\mathrm{Pr}_{(1.3b)}[\text{144 employees develop cancer}]\\
&\mathrm{Pr}_{(1.3b)}[\text{145 employees develop cancer}]
\end{aligned}$$

where $\mathrm{Pr}_{(1.3a)}[\dots]$ and $\mathrm{Pr}_{(1.3b)}[\dots]$ mean probabilities calculated according to models $(1.3\textsc{a})$ and $(1.3\textsc{b})$, respectively. Section 1.2 shows how to calculate those probabilities. Figure 1.1a shows all possible values of x from 0 to 145. Figure 1.1b zooms in on the values of x from 0 to 20 and shows the particular value of x that occurred: $x = 8$.

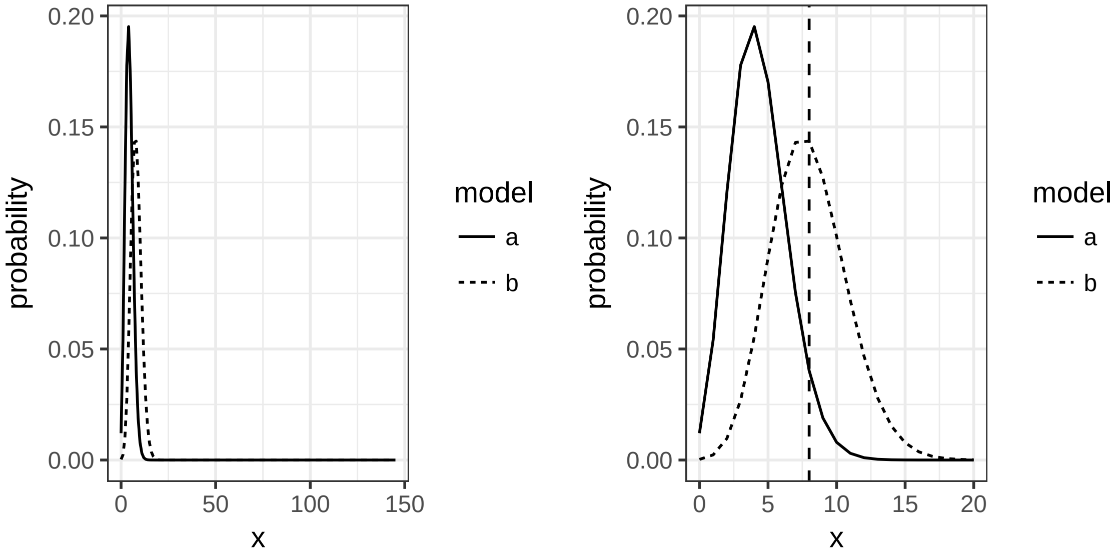
The key idea is that the ratio

$$\frac{\mathrm{Pr}_{(1.3b)}[\text{8 employees develop cancer}]}{\mathrm{Pr}_{(1.3a)}[\text{8 employees develop cancer}]} \approx \frac{0.144}{0.040} = 3.6 \tag{1.4}$$

says how much better $(1.3\textsc{b})$ describes the data than $(1.3\textsc{a})$. According to (1.4), $(1.3\textsc{b})$ describes the data about three-and-a-half times better than $(1.3\textsc{a})$. The numerator and denominator of the middle term in (1.4) can be read on the vertical line in Figure $1.1\textsc{b}$. Section 2.2 will discuss whether three-and-a-half times is a lot better or just a little better.

We'll use Example 1.2 to begin our study of R. When you start R for the first time it opens a console window. Mine looks something like the following.

```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)
```



(a) x ranges from 0 to 145

(b) x ranges from 0 to 20. The vertical dashed line indicates $x = 8$.

Figure 1.1: Pr[x employees develop cancer]

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.69 (7328) x86_64-apple-darwin13.4.0]

[History restored from /Users/michael/.Rapp.history]

>
```

It begins with information about R and how to use R. After the introductory material it ends with the line

```
>
```

The “>” is called the prompt and is R’s sign that it’s ready to receive your commands. You can type arithmetic commands directly in the console. For instance, in Example 1.2 I typed `4.2/145` and the return key, then I typed `8/145` and the return key. Then the console looked like this.

```
> 4.2/145
[1] 0.02896552
> 8/145
[1] 0.05517241
>
```

Each time, R responded with `[1]` and the answer. You can ignore the `[1]`. I rounded `0.02896552` to 0.03 and `0.05517241` to 0.055 to make pretty numbers for Example 1.2.

R knows the following arithmetic commands.

```
+ x
- x
x + y
x - y
x * y
x / y
x ^ y
x %% y
x %/% y
```

You can probably guess what most of them do. Try them out. If any are not obvious, try to figure out what they do. Use the web for help if you need it.

Making Figure 1.1 is more complicated. It requires x values from 0 to 145 and the corresponding y values from (1.3a) and (1.3b). The x values can be created by the R command

```
x <- 0:145
```

The “`<-`” is two characters — “`<`” and “`-`” — and means “assign”. `x <- 0:145` assigns the integers from 0 to 145 to a vector called `x`.

The y values were generated by the commands `dbinom ( 0:145, 145, 0.03 )` and `dbinom ( 0:145, 145, 0.055 )`. `dbinom` calculates probabilities; we’ll learn more about it in Section 1.2. The `0.03` and `0.055` tell `dbinom` whether to use (1.3a) or (1.3b) in its calculations.

We combined the x values and y values into a dataframe. “Dataframe” is R’s name for data arranged in rows and columns. Our dataframe is called `slater1.m` and looks like this.

```
  > slater1.m
      x model   probability
1     0     a  1.207540e-02
2     1     a  5.415256e-02
3     2     a  1.205871e-01
4     3     a  1.777728e-01
5     4     a  1.951835e-01
               ⋮
142 141     a 2.939866e-208
143 142     a 2.561230e-211
144 143     a 1.661818e-214
145 144     a 7.138392e-218
146 145     a 1.522586e-221
```

```
147   0     b  2.739127e-04
148   1     b  2.311592e-03
149   2     b  9.686670e-03
150   3     b  2.687325e-02
151   4     b  5.552383e-02
               ⋮
288 141     b 3.467449e-171
289 142     b 5.684766e-174
290 143     b 6.941107e-177
291 144     b 5.610830e-180
292 145     b 2.252112e-183
>
```

Row numbers are shown on the left. `slater1.m` has 146 rows for model (1.3a) and 146 rows for model (1.3b), for a total of 292 rows.

Column names are shown at the top. `slater1.m` has three columns called `x`, `model`, and `probability`. `x` goes from 0 to 145 twice, once for model (1.3a) and once for model (1.3b). `probability` shows probabilities calculated by `dbinom`. Notation like `1.207540e-02` means 1.207540×10^{-2}, or 0.0120754. The first line of `slater1.m` says $\Pr_{(1.3a)}[0 \text{ employees develop invasive cancer}] = 0.0120754$. The second line says $\Pr_{(1.3a)}[1 \text{ employee develops invasive cancer}] = 0.05415256$, and so on. The last line says $\Pr_{(1.3b)}[145 \text{ employees develop invasive cancer}] = 2.252112 \times 10^{-183}$, a *very* small number. Those probabilities can be read on the y axes of Figure 1.1.

Once we have `slater1.m` it can be used to make Figure 1.1. The commands are

```
slater.p1 <- ggplot ( slater1.m, aes ( x = x, y = probability, lty = model ) )
pdf ( "slater1a.pdf", height = 3, width = 3 )
slater.p1 + geom_line() +
            xlim ( 0, 145 ) +
            theme_bw()
dev.off()
pdf ( "slater1b.pdf", height = 3, width = 3 )
slater.p1 + geom_line() +
            geom_vline ( xintercept = 8, lty=2 ) +
            scale_x_continuous ( limits = c ( 0, 20 ),
                                 breaks = seq ( 0, 20, by = 5 )
                               ) +
            theme_bw()
dev.off()
```

`slater.p1` is a plot created with the `ggplot` command. We could have given it any name. For example, we could have written `myfirstplot <- ggplot ...` and then used `myfirstplot` in the subsequent commands. `ggplot ( slater1.m, aes ( x = x, y = probability, lty = model ) )` tells R to create a plot from the `slater1.m` dataframe. The x axis of the plot should use the column named `x` and the y axis should use the column named `probability`. `lty` stands for line type and should use the column named `model`. That's why we get a solid line for (1.3a) and a dashed line for (1.3b). The command `slater.p1 + geom_line() + xlim ( 0, 145 ) + theme_bw()` says draw the `slater.p1` plot; add a line connecting the points (We actually get two lines because we have two different line types); let the x axis run from 0 to 145; and use a black-and-white color scheme.

That plotting command is enclosed between two other commands:

```
pdf ( "slater1a.pdf", height = 3, width = 3 )
⋮
dev.off()
```

`pdf ( ... )` says to put the next plot in a pdf file named `slater1a.pdf` and make the plot 3 inches tall and 3 inches wide. `dev.off()` tells R we're finished making that pdf file and won't add any more plots to it. Running `dev.off()` creates a new pdf file on your computer.

After making the first pdf file, we opened a second pdf file with a slightly different name. Then we remade the plot with a few changes. We add a vertical line with `geom_vline ( ... )` whose x intercept is 8 and whose line type is 2. We used `scale_x_continuous ( ... )` to let the x axis run from 0 to 20 and to make tick marks every 5 units from 0 to 20.

When running a complicated sequence of R commands it is better not to type them into the console. Instead, it is better to put them in a separate file called an R script and tell R to read commands from the script. That way, we can build our command sequence bit by bit. If there is an error, we can easily correct the script without having to start all over. If we want to make minor changes, we can make them in the script without having to start all over. My script for making Figure 1.1 is

```
library ( ggplot2 )
library ( reshape2 )

slater1 <- data.frame (
  x = 0:145,
  a = dbinom ( 0:145, 145, 0.03 ),
  b = dbinom ( 0:145, 145, 0.055 )
)

slater1.m <- melt ( slater1,
                    id.vars = "x",
                    measure.vars = c ( "a", "b" ),
                    variable.name = "model",
                    value.name = "probability"
                  )

slater.p1 <- ggplot ( slater1.m, aes ( x = x, y = probability, lty = model ) )
pdf ( "slater1a.pdf", height = 3, width = 3 )
slater.p1 + geom_line() +
            xlim ( 0, 145 ) +
            theme_bw()
dev.off()
pdf ( "slater1b.pdf", height = 3, width = 3 )
slater.p1 + geom_line() +
            geom_vline ( xintercept = 8, lty=2 ) +
            scale_x_continuous ( limits = c ( 0, 20 ),
                                 breaks = seq ( 0, 20, by = 5 )
                               ) +
            theme_bw()
dev.off()
```

We've talked about most of these commands already, but a few are new. The two `library ( ... )` commands tell R to load external packages. Many people have written packages with special R commands to perform specific tasks. My script uses two of them. It uses `ggplot2` because that package has the `ggplot` command. It uses `reshape2` because that package has the `melt` command. Packages must be downloaded and installed before being used. The commands are `install.packages ( "ggplot2" )` and `install.packages ( "reshape2" )`. When you type those commands into R they should open a dialog box asking from which site you want to download the packages. Choose a site near you. Once you install the packages, then you can use them with the `library ( ... )` command. You need to install them only the first time you use them. After that, you can use `library ( ... )` without having to reinstall the packages.

Try my script. See what `slater1` is and how it gets reshaped into `slater1.m`. Start learning how to figure out R on your own by running a few commands, investigating their results, using R's `help`, and getting help on the web.

1.2 The Binomial Distribution

Model 1.3 exemplifies a common situation.

- A repeatable event results in either a success or a failure.
- Many repetitions are observed.
- Successes and failures are counted.
- The number of successes helps us learn about the probability of success.

Such observations are called *binomial.* Some examples are

Medical trials A new treatment is given to many patients. Each is either cured or not.

Free throws A basketball player shoots many free throws. Each is either successful or not.

Toxicity tests Many laboratory animals are exposed to a potential carcinogen. Each either develops cancer or not.

Ecology Many seeds are planted. Each either germinates or not.

Quality control Many supposedly identical items are subjected to a test. Each either passes or not.

Because binomial experiments are so prevalent there is specialized language to describe them. Each repetition is called a *trial*; the number of trials is usually denoted N; the unknown probability of success is usually denoted either p or θ; the number of successes is usually denoted X. We write $X \sim \text{Bin}(N, p)$. The symbol "$\sim$" is read *is distributed as*; we would say *"X is distributed as Binomial N, p"* or *"X has the Binomial N, p distribution"*. Some assumptions about binomial experiments are that N is fixed in advance, p is the same for every trial, and the outcome of any trial does not influence the outcome of any other trial. When $N = 1$ we say X has a Bernoulli(p) distribution and write $X \sim \text{Bern}(p)$; the individual trials in a binomial experiment are called Bernoulli trials.

Model (1.3) describes a Binomial distribution with $N = 145$ because there are 145 employees at Slater school, all having the same p. Model (1.3A) describes the $\text{Bin}(145, 0.03)$ distribution while Model (1.3B) describes the $\text{Bin}(145, 0.055)$ distribution.

When a binomial experiment is performed, X will turn out to be one of the integers from 0 to N. Equation (1.5) gives the formula for computing the probabilities; i.e., $\Pr_p[X = k]$ for each value of k from 0 to N.

$$\Pr_p[X = k] = \binom{N}{k} p^k (1-p)^{N-k} \tag{1.5}$$

Equation (1.5) illustrates a common notational device. When we write $\Pr_{\text{something}}[\ldots]$, the subscript indicates conditions under which the probability is calculated.

The term $p^k(1-p)^{N-k}$ is the probability of any particular sequence of k 1's and $(N-k)$ 0's. In Example 1.2, where $N = 145$ and X turned out to be 8, the probability of any particular sequence of 8 1's and 137 0's is $p^8(1-p)^{137}$. That is,

$$\begin{aligned}
&\Pr_{(1.3\mathrm{a})}[1,1,1,1,1,1,1,1,0,0,\ldots,0] \\
&= \Pr_{(1.3\mathrm{a})}[1,1,1,1,1,1,1,0,1,0,\ldots,0] \\
&= \Pr_{(1.3\mathrm{a})}[1,1,1,1,1,1,0,1,1,0,\ldots,0] \\
&= \ldots \\
&= \Pr_{(1.3\mathrm{a})}[0,\ldots,0,1,1,1,1,1,1,1,1] \\
&= 0.03^8 \times 0.97^{137} = 1.010875 \times 10^{-14}.
\end{aligned} \tag{1.6}$$

Each sequence in (1.6) has 8 1's and 137 0's.

The term $\binom{N}{k}$ in (1.5) is called a *binomial coefficient* and is read "N choose k". $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ and is equal to the number of sequences of k 1's and $(N-k)$ 0's. For Example 1.2,

$$\binom{145}{8} = \frac{145!}{8!137!} = 3.981763 \times 10^{12} \tag{1.7}$$

The R command is `choose(145,8)`. Try it.

Combining (1.6) and (1.7) gives

$$\begin{aligned}
\Pr_{(1.3\mathrm{a})}[X = 8] &= \binom{145}{8} 0.03^8\, 0.97^{137} \\
&= 3.981763 \times 10^{12} \times 1.010875 \times 10^{-14} = 0.04025065.
\end{aligned}$$

That's the denominator of (1.4). The numerator is

$$\begin{aligned}\Pr_{(1.3b)}[X = 8] = \binom{145}{8} 0.055^8 \, 0.945^{137} \\ = 3.981763 \times 10^{12} \times 3.606279 \times 10^{-14} = 0.1435935.\end{aligned}$$

Binomial probabilities are readily computed in R. The command is `dbinom ( x, N, p )`. For example, the numerator and denominator of (1.4) can be calculated by `dbinom ( 8, 145, .055 )` and `dbinom ( 8, 145, .03 )`. The ratio can be calculated by `dbinom ( 8, 145, .055 ) / dbinom ( 8, 145, .03 )`. Try it.

1.3 The Poisson Distribution

Another common type of observation occurs in the following situation.

- There is a domain of study, usually a block of space or time.
- Events arise seemingly at random in the domain.
- There is an underlying rate at which events arise.

Such observations are called *Poisson*. The number of events in the domain of study helps us learn about the rate. Some examples are

Ecology Tree seedlings emerge from the forest floor.

Computer programming Bugs occur in computer code.

Quality control Defects occur along a strand of yarn.

Genetics Mutations occur in a genome.

Traffic flow Cars arrive at an intersection.

Customer service Customers arrive at a service counter.

Neurobiology Neurons fire.

The rate at which events occur is often called λ, the Greek letter lambda. The number of events that occur in the domain of study is often called X; we write $X \sim \text{Poi}(\lambda)$. Any particular value of λ specifies a particular Poisson distribution for the random variable X. Assumptions about Poisson distributions are that two events cannot occur at exactly the same location in space or time, that the occurrence of an event at one location does not influence whether an event occurs at any other location, and the rate at which events arise is constant over the entire domain of study.

When a Poisson experiment is observed, X will turn out to be a nonnegative integer. The probabilities are given by (1.8).

$$\text{Pr}_\lambda[X = x] = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{1.8}$$

We do not know in advance which value of the emergence rate λ will best describe our data. The data are collected to help us learn λ.

The expression $\text{Pr}[X = x]$ may seem odd. It certainly sounds odd to say "the probability that ex equals ex." However, it's correct. The upper case X is a random variable, which is the as-yet unknown value that will be observed. The lower case x is a specific integer. In Example 1.3, X is the number of seedlings that will emerge in a meter-square forest quadrat. Before the experiment is performed, or before the data are observed, the number of seedlings X is unknown and may turn out to be any integer 0 or bigger. Equation (1.8) describes $\text{Pr}_\lambda[X = 0]$, $\text{Pr}_\lambda[X = 1]$, $\text{Pr}_\lambda[X = 2]$, and so on. The lower case x stands for whatever integer we're plugging into the equation.

Figure 1.2 plots $\text{Pr}_\lambda[X = x]$ for different values of x (on the x axis) and different values of λ (different types of points and lines). Figure 1.2 illustrates some things that are true about all Poisson distributions.

1. For each value of λ, that is for each specific Poisson distribution, the plot of $\text{Pr}_\lambda[X = x]$ has a single peak. The probabilities are highest for x's at or near the peak and gradually decrease as x moves away from the peak.

2. For each λ, the value of x with the highest probability is a value of x near λ.

3. As λ increases, the peak becomes lower and the probability distribution becomes more spread out. When $\lambda = 1$, there are only about 5 values of x with appreciable probability. As λ increases, the number of x's with appreciable probability also increases.

To summarize, when the rate λ is small, we can be fairly certain that the number of events X we observe will also be small, and we can predict X fairly accurately. When λ is large we can still be fairly certain that X will be large, but we cannot predict it with as much accuracy.

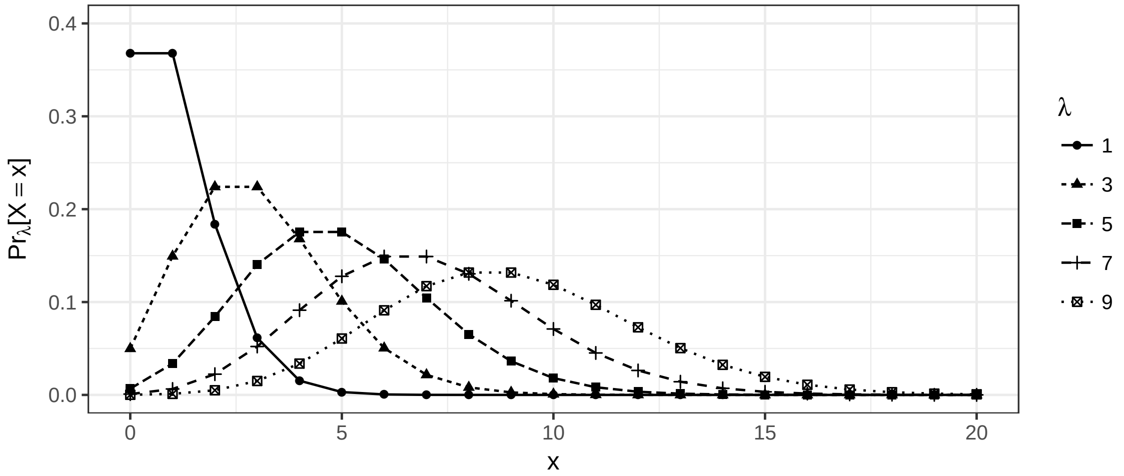


Figure 1.2: $\Pr_\lambda[X = x]$. Each line is for a different value of λ.

Figure 1.2 was created by the following R script.

```
library ( ggplot2 )

# Illustrate the Poisson distribution
df.Pois <- expand.grid ( x = 0:20, lambda = seq ( 1, 9, by = 2 ) )
df.Pois$prob <- dpois ( df.Pois$x, df.Pois$lambda )

pdf ( "PoissonProbs.pdf", height = 3, width = 7 )
p <- ggplot ( df.Pois,
              aes ( x = x,
                    y = prob,
                    lty = as.factor(lambda),
                    shape = as.factor(lambda)
                  )
            )
p + geom_point() +
    geom_line() +
    ylim ( 0, 0.4 ) +
    ylab ( expression ( paste ( Pr[lambda], "[", X == x, "]" ) ) ) +
    scale_linetype ( name = expression(lambda) ) +
    scale_shape ( name = expression(lambda) ) +
    theme_bw()
dev.off()
```

`#` is R's comment character. It is used to write comments in your script to help you remember later what you were doing when you wrote the script. Everything after the `#` on a line is ignored by R. After the comment, I created a dataframe called `df.Pois`. The command `seq ( 1, 9, by = 2 )` creates a sequence of `lambda` values starting at 1, going to 9, and taking steps of size 2. In other words, it yields the sequence 1, 3, 5, 7, 9. The command `0:20` creates a sequence of x values: 0, 1, 2, . . . , 20. In R, `a:b` creates a sequence of integers from `a` to `b` in steps of size 1. It is shorthand for `seq(a,b,by=1)`. `expand.grid ( x = ..., lambda = ... )` creates a dataframe that pairs each value of x with each value of λ. If we were to examine `df.Pois` we would see

```
> df.Pois
     x lambda
1    0      1
2    1      1
3    2      1
4    3      1
⋮
102 17      9
103 18      9
104 19      9
105 20      9
>
```

There are 21 values of x and 5 values of λ, so `df.Pois` has $21 \times 5 = 105$ rows. The R notation `df.Pois$prob` shows how to use the "`$`" notation to refer to a specific column in a dataframe by name. The command `df.Pois$prob <- ...` creates a column called `prob`. The command `dpois` calculates probabilities according to (1.8). If we were to examine `df.Pois` again we would now see

```
> head ( df.Pois )
  x lambda        prob
1 0      1 0.367879441
2 1      1 0.367879441
3 2      1 0.183939721
4 3      1 0.061313240
5 4      1 0.015328310
6 5      1 0.003065662
> tail ( df.Pois )
     x lambda         prob
100 15      9 0.0194306663
101 16      9 0.0109297498
102 17      9 0.0057863381
103 18      9 0.0028931691
104 19      9 0.0013704485
105 20      9 0.0006167018
```

I purposely used a different way of displaying `df.Pois` so I could show you a few more R commands. `head(...)` and `tail(...)` show the first several rows and last several rows of a dataframe. We see that `df.Pois` has a new column called `prob`. The plotting commands are similar to ones we've seen before, but with a few additions you may explore.

As usual, you should try these commands and play with them until you understand what they do and how to use them.

One of the main themes of statistics is the way in which data help us learn about the phenomenon we are studying. Example 1.3 shows how this works when we want to learn about which values of λ in a Poisson distribution give a good description of the rate at which tree seedlings emerge in a forest.

Example 1.3 (Seedlings in a Forest)
Tree populations move by dispersing their seeds. Seeds become seedlings, seedlings become saplings, and saplings become adults which eventually produce more seeds. Over time, whole populations may migrate in response to climate change. One instance oc-

curred at the end of the Ice Age when species that had been sequestered in the south were free to move north. Another instance may be occurring today in response to global warming. One critical feature of migration is its speed. Some of the factors determining the speed are the typical distances of long range seed dispersal, the proportion of seeds that germinate and emerge from the forest floor to become seedlings, and the proportion of seedlings that survive each year.

To learn about emergence and survival, ecologists return annually to forest quadrats (square meter sites) to count seedlings that have emerged since the previous year. One such study was conducted at the Coweeta Long Term Ecological Research station in western North Carolina and reported in Lavine, Beckage, and Clark, 2002. A fundamental quantity of interest is the rate λ at which seedlings emerge. In one quadrat, three new seedlings were observed in one year. What does that quadrat tell us about λ?

Different values of λ yield different values of $\mathrm{Pr}_\lambda[X = 3]$. To compare different values of λ we see how well each one describes the data $X = 3$; i.e., we compare $\mathrm{Pr}_\lambda[X = 3]$ for different values of λ. For example,

$$\begin{aligned}
\mathrm{Pr}_{\lambda=1}[X = 3] &= \frac{1^3 e^{-1}}{3!} \approx 0.06 \\
\mathrm{Pr}_{\lambda=2}[X = 3] &= \frac{2^3 e^{-2}}{3!} \approx 0.18 \\
\mathrm{Pr}_{\lambda=3}[X = 3] &= \frac{3^3 e^{-3}}{3!} \approx 0.22 \\
\mathrm{Pr}_{\lambda=4}[X = 3] &= \frac{4^3 e^{-4}}{3!} \approx 0.19
\end{aligned}$$

The R commands `dpois(3,1); dpois(3,2); dpois(3,3); dpois(3,4)` do the calculations. You can put them all on one line, separated by semicolons. Try it.

In other words, the value $\lambda = 3$ explains the data almost four times as well as the value $\lambda = 1$ and just a little bit better than the values $\lambda = 2$ and $\lambda = 4$. Figure 1.3 shows $\mathrm{Pr}_\lambda[X = 3]$ plotted as a function of λ. The figure suggests that $\mathrm{Pr}_\lambda[X = 3]$ is largest for values of λ near 3. In other words, values of λ near 3 describe the data better than values of λ far from 3. The figure also shows that any value of λ from about 0.5 to about 9 describes the data not too much worse than $\lambda = 3$, but values of λ less than about 0.5 or bigger than about 9 describe the data much worse than $\lambda \approx 3$. That's the inference from just one quadrat with $X = 3$. There are many other quadrats and each has its own value of X. Later we'll see how to combine the data from all quadrats to make better inference for λ.

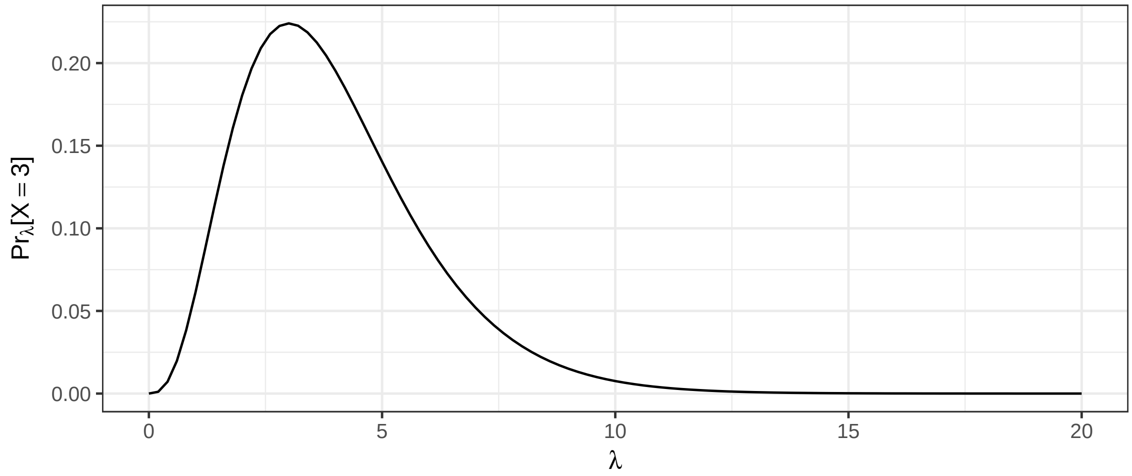


Figure 1.3: $\Pr_\lambda[X = 3]$

1.4 Probability Density

Section 1.4 uses Example 1.4 to illustrate the idea of *probability density*.

Example 1.4 (Galton's Heights)
There is much better understanding of heredity today than there was in the 1890's when Sir Francis Galton collected several famous data sets to study the topic. This example uses one of them, `GaltonFamilies`, which can be found in the R package `HistData`. As usual, you must install `HistData` before you use it. The R command `install.packages("HistData")` does that. When you type that command into your R console, it should open a window asking from which R repository you want to download the package. Choose a repository near you. You have to install the package only once. Then you use `library ( HistData )` in each R session you want to use it.

Once you have loaded `HistData` you can type `?GaltonFamilies`, which will open a Help page. (Typing `help.start()` opens an extensive help page in your web browser.) You can use the `dim`, `names`, and `head` commands to learn about the data. When I used those commands my console looked like this.

```
> dim ( GaltonFamilies )
[1] 934   8
> names ( GaltonFamilies )
[1] "family"          "father"          "mother"          "midparentHeight"
[5] "children"        "childNum"        "gender"          "childHeight"
> head ( GaltonFamilies )
```

```
  family father mother midparentHeight children childNum gender childHeight
1    001   78.5   67.0           75.43        4        1   male        73.2
2    001   78.5   67.0           75.43        4        2 female        69.2
3    001   78.5   67.0           75.43        4        3 female        69.0
4    001   78.5   67.0           75.43        4        4 female        69.0
5    002   75.5   66.5           73.66        4        1   male        73.5
6    002   75.5   66.5           73.66        4        2   male        72.5
>
```

dim told me that GaltonFamilies has 934 rows and 8 columns; names told me the names of the columns; and head showed me the first several rows.

We see the first four rows are all for family #1. That family had four children, first a boy, then three girls. The father's height is 78.5 inches and the mother's height is 67.0 inches. Galton studied heredity by studying the relationship between the heights of parents and children. We'll get to that later, but for now we'll ignore the children and work with just the parents. It's convenient to make a dataframe that has just one row for each family and that has columns just for fathers' and mothers' heights. The following R command does the trick.

```
GaltonParents <- GaltonFamilies [ GaltonFamilies$childNum == 1, c ( "father", "mother" ) ]
```

(I wrote that in a small font so it would fit on one line.) The square brackets "[" and "]" make subsets. When we write GaltonFamilies [something, something] the first something says which rows we want and the second something says which columns. We told R we wanted all the rows for which childNum is equal to 1. That's how we got one row for each family. We also told R we wanted just the columns named father and mother. Then we saved the result in a new dataframe called GaltonParents. We can check we got something sensible like this:

```
> head(GaltonParents)
   father mother
1    78.5   67.0
5    75.5   66.5
9    75.0   64.0
11   75.0   64.0
16   75.0   58.5
22   74.0   68.0
```

Can you figure out what the 1, 5, 9, 11, 16, and 22 mean?

I made `GaltonParents` without using external R packages. Another way to accomplish the same task is with the `dplyr` package. The commands would be something like this:

```
library ( dplyr )
GaltonParents <-
  GaltonFamilies %>%
    filter ( childNum == 1 ) %>%
    select ( father, mother )
```

`%>%` is a pipe. It pipes `GaltonFamilies` into the `filter` command and pipes the output of `filter` into the `select` command. `filter` selects the desired rows and `select` selects the desired columns. `%>%`, `filter`, and `select` are all in the `dplyr` package. The `dplyr` package simplifies many data-handling tasks and is worth learning if you have to manipulate many data sets.

```
> range ( GaltonParents$mother )  # 12 inches
[1] 58.0 70.5
> range ( GaltonParents$father )  # 16.5 inches
[1] 62.0 78.5
```

shows us the shortest and tallest mothers and fathers. We're going to combine all the mothers and fathers into one large group of people for what comes next, so we modify the dataframe as follows:

```
GaltonParents <- data.frame ( height = c ( GaltonParents$mother, GaltonParents$father ) )
```

`c` combines the two vectors into one. The result is that `GaltonParents` is now a dataframe with just one column, `height`. There are 205 families altogether, so there are 410 people in `GaltonParents`.

Figure 1.4 displays the heights of all the parents in Galton's data. The points have been jittered a bit in both directions so they don't fall on top of each other. The x axis is height. The y axis is meaningless because it's just random jitter.

Figure 1.4 was produced by the following R code.

```
p <- ggplot ( GaltonParents, aes ( x = height, y = 0 ) )
p + geom_jitter ( height = .5, width = .5 ) +
    scale_x_continuous ( breaks = seq ( 56, 80, by = 4 ) ) +
    scale_y_continuous ( name = "", breaks = NULL ) +
    theme_bw()
```

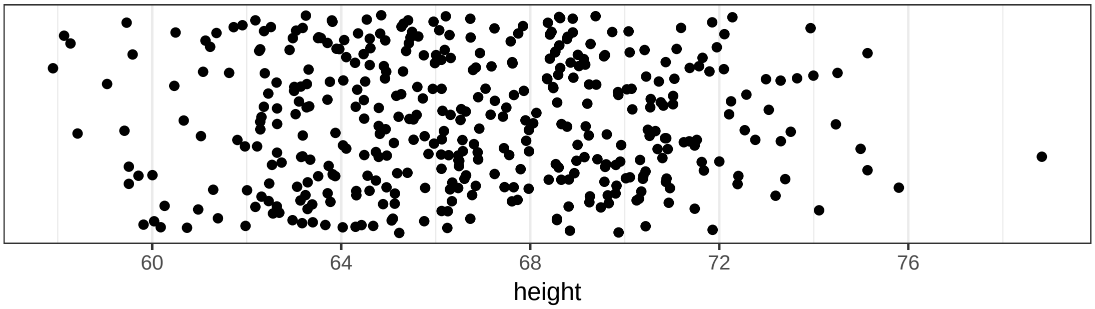


Figure 1.4: Heights of parents in Galton's data. Points are jittered in both directions for legibility.

`geom_jitter` does the jittering.

Notice that the points are denser in the middle of the plot and sparser at both ends. That's the idea we'll try to express with *probability density*. To begin, we can use the following code to get R to count the number of people with heights between 56 and 60, 60 and 64, 64 and 68, 68 and 72, 72 and 76, and 76 and 80.

```
# density
tmp <- cut ( GaltonParents$height, breaks = seq ( 56, 80, by = 4 ) )
table ( tmp )
tmp
(56,60] (60,64] (64,68] (68,72] (72,76] (76,80]
     17     102     148     121      21       1
```

`# density` is a comment in my script to remind me what I'm doing. `cut` divides `GaltonParents$height` into groups which are saved in the variable called `tmp`. Finally, the `table` command shows how many people are in each group. It shows there are 17 parents whose heights are bigger than 56 inches but less than or equal to 60 inches; 102 parents whose heights are bigger than 60 inches but less than or equal to 64 inches; and so on. Naturally, the counts add up to 410, the number of people in the dataframe.

By generalizing the idea of population density we could say there are 17 people in four inches, or $17/4 = 4.25$ people per inch, near the height of 58 inches. There are 102 people in four inches, or $102/4 = 25.5$ people per inch, near the height of 62 inches, ..., and 1 person in four inches, or $1/4 = .25$ people per inch near the height of 78 inches. That gives us a population density, but our population density is people per inch

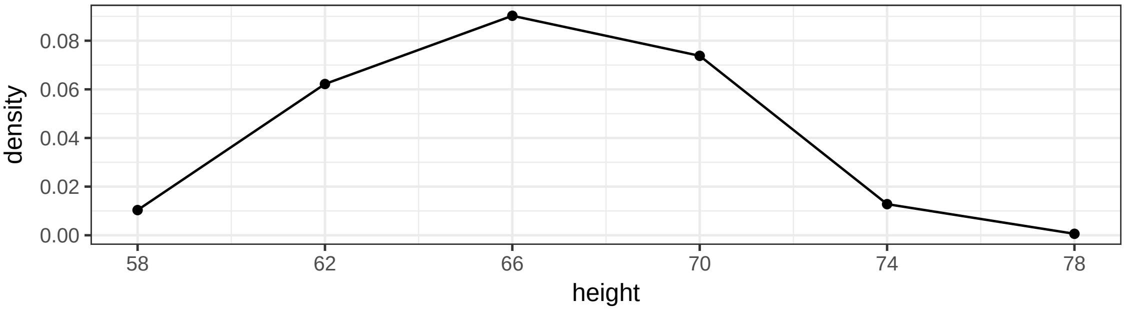


Figure 1.5: Probability density of parents' heights in Galton's data, estimated from bins of width 4 inches.

of height, whereas the usual population density is people per square kilometer of area. Our population density of heights varies from height to height — i.e., from the height of 58 to the height of 78 — just as ordinary population density varies from place to place.

The next step is to recognize that we don't care as much about the number of people in each group as about the fraction of people in each group. So instead of saying there are 17 people between 56 and 60, or $17/4 = 4.25$ people per inch we say that $17/410 \approx 0.04$, or about 4% of the population falls between 56 and 60 inches. That's 0.04 in four inches, or a density of $0.04/4 = 0.01$, or about 1% of the population per inch, near the height of 58 inches. Similarly, the density is about $(102/410)/4 \approx 0.06$, or about 6% of the population per inch near the height of 62 inches. The highest density occurs around 66 inches and is about $(148/410)/4 \approx 0.09$, or about 9% of the population per inch. Figure 1.5 plots the densities we just calculated.

Figure 1.5 was created with the following code.

```
  p <- ggplot ( mapping = aes ( x = seq ( 58, 78, by = 4 ),
                                y = as.vector ( table ( tmp ) ) / (410*4)
                              )
              )
  p + geom_point() +
      geom_line() +
      scale_x_continuous ( name = "height", breaks = seq ( 58, 78, by = 4 ) ) +
      scale_y_continuous ( name = "density", breaks = seq ( 0, .1, by = .02 ) ) +
      theme_bw()
```

The new trick is that we didn't tell `ggplot` to get its data from a dataframe. Instead we put the data directly in `aes ( x = ..., y = ... )`.

We could just as well estimate probability density from bins of size 2 inches or 1 inch.

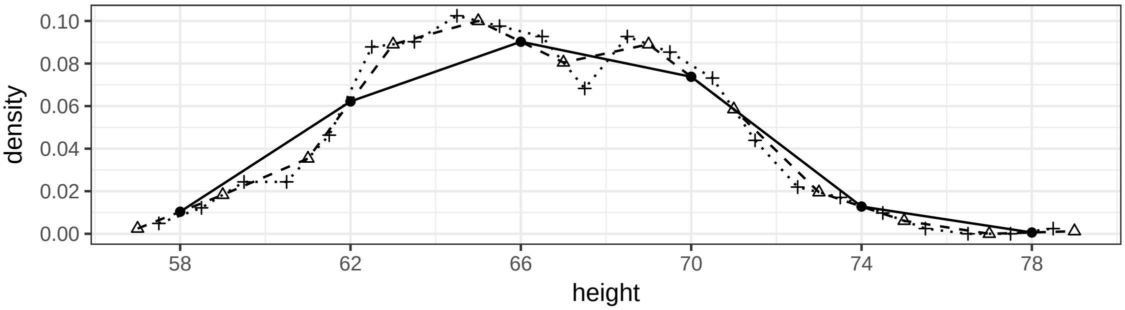


Figure 1.6: Estimated probability density of parents' heights in Galton's data. Solid line: bins of 4 inches. Dashed line: bins of 2 inches. Dotted line: bins of 1 inch.

(Bins smaller than 1 inch wouldn't make sense for Galton's heights because they were recorded only to the nearest half-inch or so.) Figure 1.6 shows the density estimated from bins of size 4 inches (same as Figure 1.5), 2 inches, and 1 inch. The three estimated densities are about the same. They all show that heights are densest between about 63 and 69 inches or so and that the density drops gradually below 63 and above 69 inches.

Figure 1.6 was created with the following code.

```
  tmp2in <- data.frame ( height = seq ( 57, 79, by = 2 ),
                         den = as.vector ( table ( cut ( GaltonParents$height,
                                                         breaks = seq ( 56, 80, by = 2 )
                                                       )
                                                 )
                                         ) / ( 410 * 2 )
                       )
  tmp1in <- data.frame ( height = seq ( 57.5, 78.5, by = 1 ),
                         den = as.vector ( table ( cut ( GaltonParents$height,
                                                         breaks = seq ( 57, 79, by = 1 )
                                                       )
                                                 )
                                         ) / 410
                       )

  p + geom_point() +
      geom_line() +
      scale_x_continuous ( name = "height", breaks = seq ( 58, 78, by = 4 ) ) +
      scale_y_continuous ( name = "density", breaks = seq ( 0, .1, by = .02 ) ) +
      geom_point ( data = tmp2in,
                   mapping = aes ( x = height, y = den ),
                   shape = 2
                 ) +
      geom_line ( data = tmp2in,
                  mapping = aes ( x = height, y = den ),
                  lty = 2
                ) +
      geom_point ( data = tmp1in,
                   mapping = aes ( x = height, y = den ),
                   shape = 3
                 ) +
      geom_line ( data = tmp1in,
                  mapping = aes ( x = height, y = den ),
                  lty = 3
                 ) +
      theme_bw()
```

We created two new dataframes, `tmp2in` and `tmp1in`. Then we repeated Figure 1.5 and added to it one `geom_point` and one `geom_line` for `tmp2in` and another for `tmp1in`.

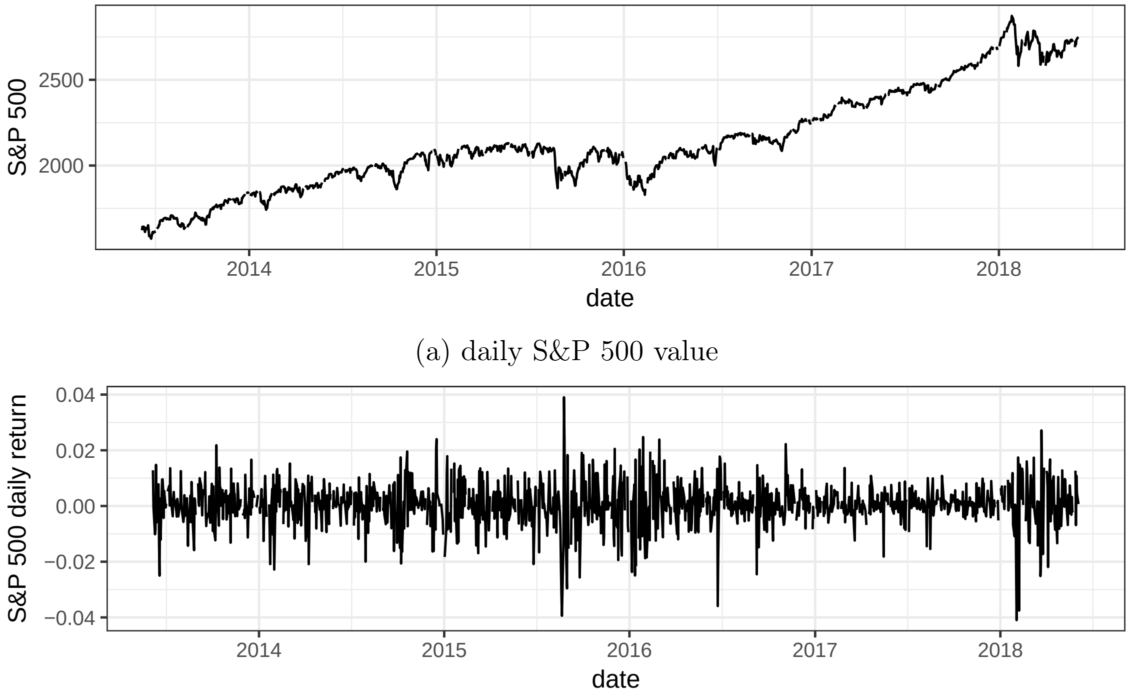
Example 1.4 introduced probability density for one aspect of a human population — heights. But human populations aren't the only things that can have probability density. Example 1.5 illustrates probability density for stock returns.

Example 1.5 (Standard and Poor's 500)
Standard and Poor's 500, or the S&P 500, is a composite index of prices on US stock exchanges. According to the US FRED (Federal Reserve Economic Data (FRED) is a database maintained by the Research division of the Federal Reserve Bank of St. Louis.), "The S&P 500 is regarded as a gauge of the large cap U.S. equities market. The index includes 500 leading companies in leading industries of the U.S. economy, which are publicly held on either the NYSE or NASDAQ, and covers 75% of U.S. equities." The daily level of the S&P 500 can be downloaded from FRED at HTTPS://FRED.STLOUISFED.ORG/SERIES/SP500. This book's website has five years of daily data in the file SP500.csv which we shall explore in this example. Figure 1.7A displays the data. We see that the S&P 500 has been steadily rising from around 1600 in mid-2013 to around 2750 in mid-2018.

Figure 1.7A was produced with the following code.

```
  sp500 <- read.csv ( "data/SP500.csv",
                      na.strings = ".",
                      colClasses = c ( "Date", "numeric" ),
                      col.names = c ( "date", "sp" ),
                    )
  p <- ggplot ( sp500, aes ( x = date, y = sp ) )
  p + geom_line() + ylab ( "S&P 500" ) + theme_bw()
```

`read.csv` is for reading .csv files. We told it the name of the file, that lone dots are to be treated as missing values, that the first column is a date and the second column is a number, and that the columns of the resulting dataframe are to be called `date` and `sp`. The rest of the code should be familiar.



(a) daily S&P 500 value

(b) daily S&P 500 return

Figure 1.7: Standard and Poor's 500, 2013–2018

The gain or loss of the S&P 500 varies from day to day. The daily rate of return on day d is defined to be

$$\text{return}_d \equiv \frac{\text{sp}_{d+1} - \text{sp}_d}{\text{sp}_d}$$

where sp_d is the S&P 500 on day d and sp_{d+1} is the S&P 500 on day $d+1$. Figure 1.7B shows the daily returns. Daily returns are mostly near 0, but with an occasional excursion to around ±4%. There is no obvious pattern, so we shall treat these data as a sample from a potentially infinite population of returns and we shall calculate a probability density that describes the population fairly well.

Figure 1.7B was produced by the following code.

```
sp500$ret <- NA
sp500$ret[-1304] <- ( sp500$sp[-1] - sp500$sp[-1304] ) / sp500$sp[-1304]
p <- ggplot ( sp500, aes ( x = date, y = ret ) )
p + geom_line() + ylab ( "S&P 500 daily return" ) + theme_bw()
```

The first two lines define a new variable `ret`. Can you figure out how they do it? The last two lines should be familiar.

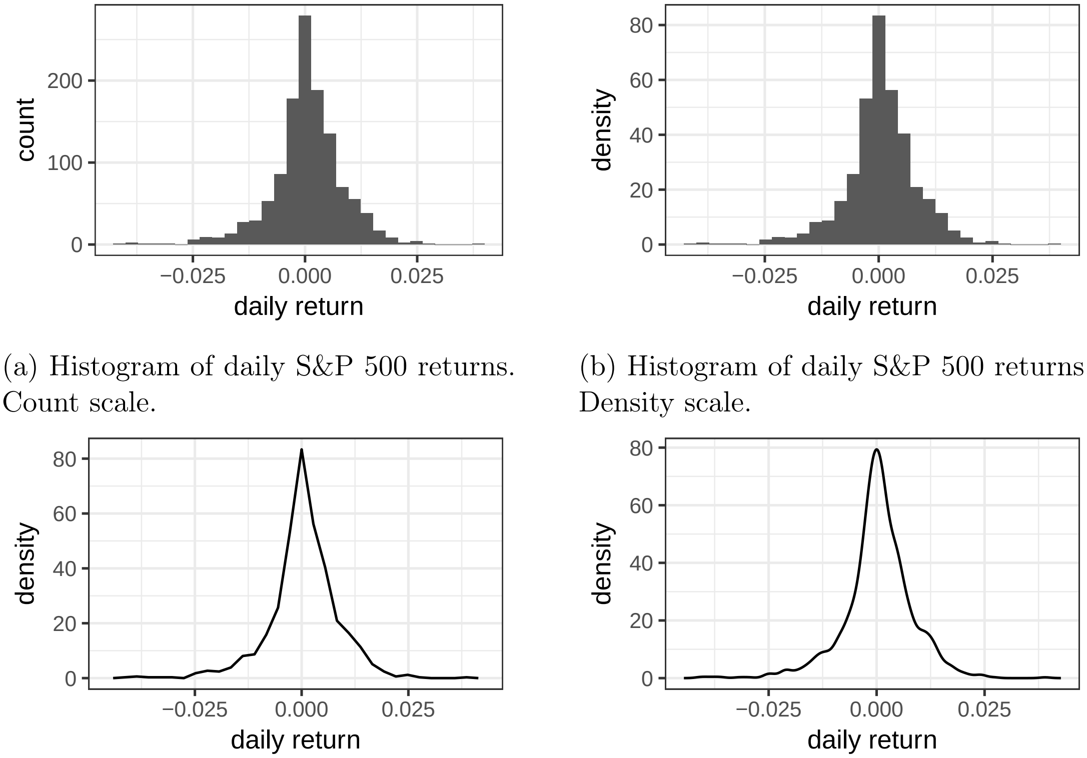
Figure 1.8 shows four different ways of looking at the S&P 500 return density. Figure 1.8A is an ordinary histogram, not really a density; the y-axis is the number of days in each bin. Figure 1.8B is the same histogram but with the y-axis rescaled so the figure is a density. Figure 1.8C is the same as 1.8B but we told `ggplot` to connect the dots instead of fill in the rectangles. Figure 1.8D is the same as 1.8C but we estimated the density with R's `density` function instead of asking `ggplot` to estimate it. In each panel of Figure 1.8 the return density is centered near 0, which means that more days have returns closer to 0 than to any other number. However, the density falls slightly more steeply to the left of 0 than to the right, which means there are more days with positive than with negative returns. That's what makes the stock market go up, on average.

Figure 1.8 was drawn with the following code.

```
p <- ggplot ( sp500, aes ( x = ret ) )
p + geom_histogram() +
    xlab ( "daily return" ) +
    theme_bw()  # Fig 1.8a

p + geom_histogram ( mapping = aes ( y = ..density.. ) ) +
    xlab ( "daily return" ) +
    theme_bw()  # Fig 1.8b

p + geom_freqpoly ( mapping = aes ( y = ..density.. ) ) +
```



(a) Histogram of daily S&P 500 returns. Count scale.

(b) Histogram of daily S&P 500 returns. Density scale.

(c) Density of daily S&P 500 returns estimated by `ggplot2`.

(d) Density of daily S&P 500 returns estimated by R's `density` function.

Figure 1.8: Standard and Poor's 500 daily returns, 2013–2018

```
    xlab ( "daily return" ) +
    theme_bw()  # Fig 1.8c

tmp <- density ( sp500$ret, na.rm = TRUE )
df <- data.frame ( ret = tmp$x, den = tmp$y )
p <- ggplot ( df, aes ( x = ret, y = den ) )
p + geom_line() +
    xlab ( "daily return" ) +
    ylab ( "density" ) +
    theme_bw()  # Fig 1.8d
```

See how the R code changes slightly in each part of the figure.

Probability density has units of fraction-of-population per unit-of-x. For example, with the heights of parents in Galton's data (Example 1.4) there were 148 parents, or $148/410 \approx 0.37$, or about 37%, of the population with heights between 64 and 68 inches. So the density near 66 inches is $\frac{.37 \text{ of the population}}{4 \text{ inches}} \approx 0.09 \frac{\text{of the population}}{\text{inch}}$.

Unlike probability, probability density can be greater than 1. For example, the middle bar of the density histogram in Figure 1.8B has a height of a little more than 80, or 8000% of the population per inch. The width of that bar is only about 0.0028 inches, so that bar represents about $\frac{8000\% \text{ of the population}}{\text{inch}} \times 0.0028 \text{ inches} \approx$ 22% of the population. We can verify the estimate with the R command

```
sum ( sp500$ret > -.0014 & sp500$ret <= .0014, na.rm = TRUE ) / sum ( !is.na ( sp500$ret ) )
```

which yields 0.231657 or about 23%. Try to understand the command.

Figure 1.9 shows what the density would be if Galton had recorded heights in centimeters instead of inches. $1\,\text{in} = 2.54\,\text{cm}$ so $66\,\text{in} = 66 \times 2.54\,\text{cm} \approx 168\,\text{cm}$ and the density near 66 inches is about

$$\begin{aligned} 0.09 \frac{\text{of the population}}{\text{in}} &= 0.09 \frac{\text{of the population}}{\text{in}} \times \frac{1\,\text{in}}{2.54\,\text{cm}} \\ &\approx .035 \frac{\text{of the population}}{\text{cm}} \end{aligned}$$

near 168 centimeters.

The area under any probability density curve is 1, or 100% of the population, as illustrated by Figure 1.10. Figure 1.10A repeats Figure 1.5 but outlines the trapezoids under the density and their mid-heights (dotted lines). The area under the curve is the area of the trapezoids. The area of each trapezoid is width × mid-height. Reading the mid-heights from Figure 1.10A we approximate the area to be $4 \times (.035 + .075 + .08 + .04 + .005) = 4 \times 0.235 = 0.94$. The difference between 0.94 and 1 is due to the error in reading the mid-heights from the figure instead of calculating them more accurately. Figure 1.10B repeats Figure 1.8D but with an

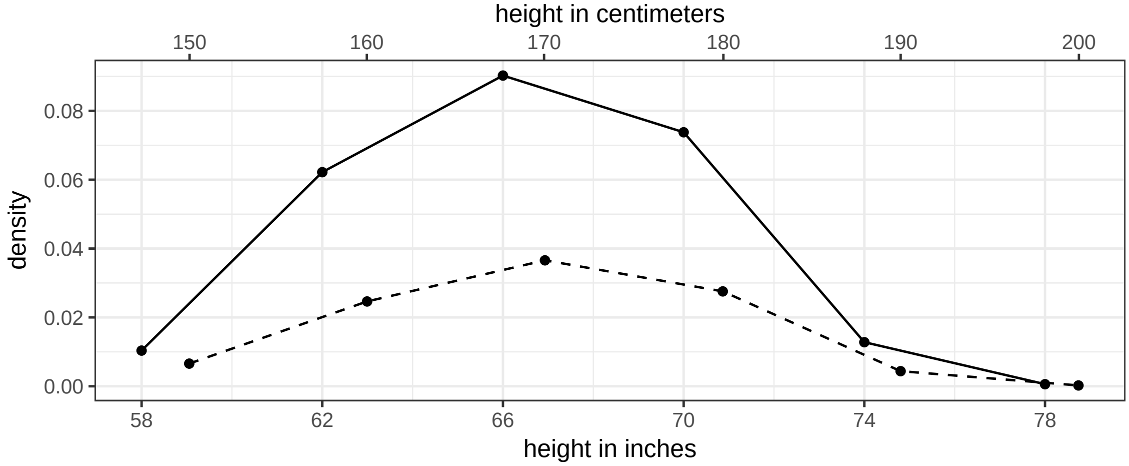


Figure 1.9: Density of Galton's parent's heights. Solid line: per inch. Dashed line: per centimeter.

approximating triangle formed by the points $(-0.0125, 0)$, $(0, 80)$, and $(0, 0.0125)$. The area under the curve can be approximated by the area of the triangle, which is $\frac{1}{2} \times 0.025 \times 80 = 1$.

In this section we started with observations — either heights of parents or rates of return — from which we estimated densities. It is also useful to have preexisting families of densities already worked out and then, when we work with a particular data set, see whether any of the preexisting families is an adequate model for the data. Sections 1.5 and 1.6 introduce two useful families.

1.5 The Exponential Distribution

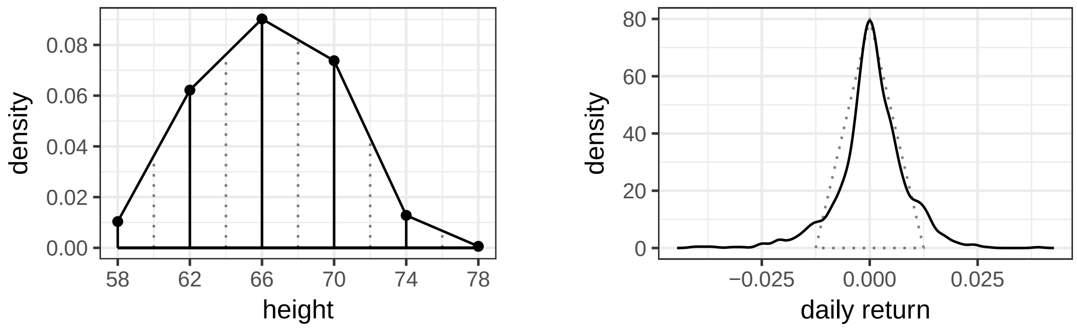
We often deal with positive random variables X that are densely concentrated on values of x near 0 and become gradually sparser as x increases. Some examples are

Customer service time on hold at a help line

Neurobiology time until the next neuron fires

Seismology time until the next earthquake

Medicine remaining years of life for a cancer patient



(a) Probability density of parents' heights in Galton's data, with trapezoids and mid-heights (dotted lines).

(b) Probability density of S&P 500 daily returns with approximating triangle.

Figure 1.10: Area under density curves

Ecology dispersal distance of a seed

In these examples it is expected that most calls, times, or distances will be short and a few will be long. So a probability density for these examples should be large near $x = 0$ and decreasing as x increases. A useful density for such situations is the *Exponential density*

$$p_\lambda(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \qquad \text{for } x > 0 \tag{1.9}$$

where p_λ is the particular exponential density with parameter λ. There is one exponential density for each value of $\lambda > 0$. We say X *has an exponential density (or distribution) with parameter* λ and write $X \sim \text{Exp}(\lambda)$. Figure 1.11 shows exponential densities for several values of λ.

An R command for evaluating (1.9) is `dexp`. Figure 1.11 was produced by the following code.

```
df <- expand.grid ( x = seq ( 0, 4, length = 50 ),
                    lambda = c ( .1, 1, 5 )
                  )
df$den <- dexp ( df$x, rate = df$lambda )

p <- ggplot ( df, aes ( x = x, y = den, lty = as.factor ( lambda ) ) )
p + geom_line() +
    ylab ( "density" ) +
    scale_linetype ( name = expression ( lambda ) ) +

    theme_bw()
```

We chose three values of λ: .1, 1, and 5. We chose 50 values of x from 0 to 4 at which to evaluate the density (1.9). Then we used `expand.grid` to make a dataframe with all 150 combinations of x and λ. After we created the dataframe we used `dexp` to add another column called `den`. Then we made the plot. The plot uses two new R commands. `scale_linetype` allows us to specify various aspects of the scale on the right-hand side of the plot that shows the different linetypes. The only aspect we specified is the name of the scale. See R's help pages or the web for other aspects we could have specified. The other new command is `expression`, which allows us to put mathematical expressions into titles, axis labels, and other places where we normally put text. That's how we got the symbol λ as the name of the scale, instead of the word `lambda`.

Example 1.6 illustrates with the firing times of a neuron.

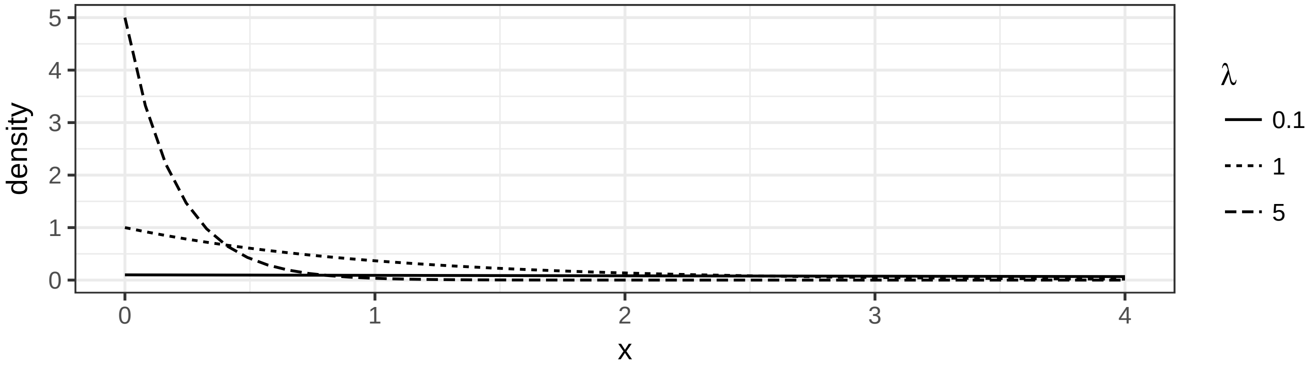
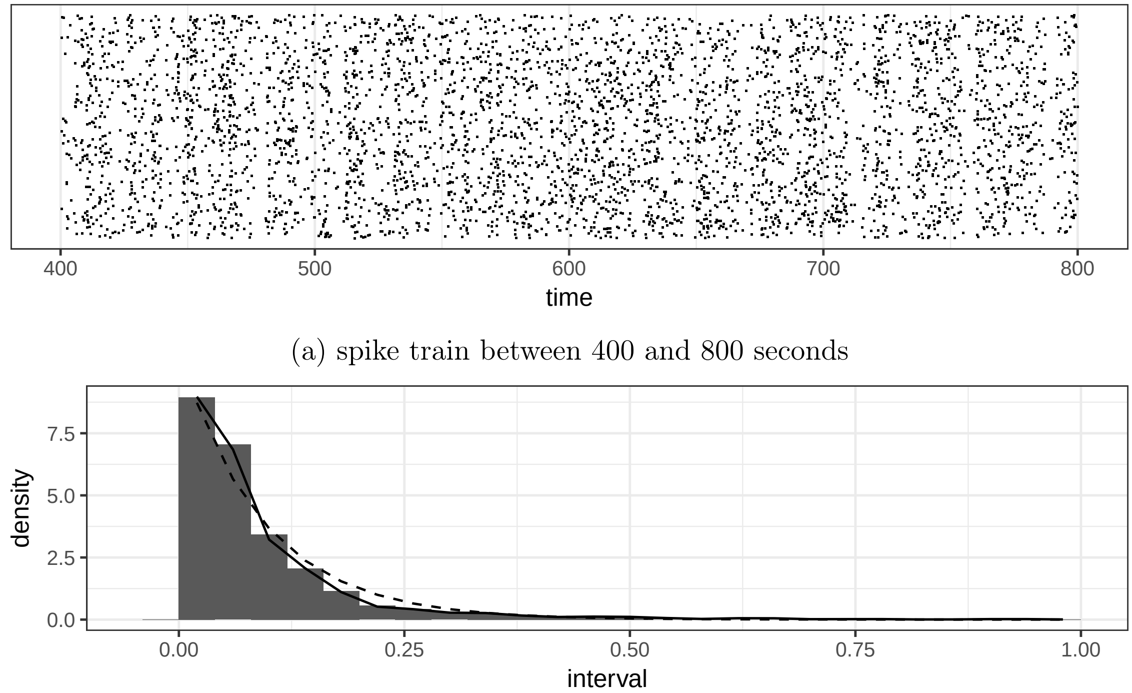


Figure 1.11: Exponential densities

Example 1.6 (neuron firing)
To learn about brain function, neuroscientists can record electrical signals from individual neurons. Most of the time, a neuron's electrical signal is small, but every so often it spikes, and we say the neuron fires. Example 1.6 is about the spike times of one neuron in one experiment. A sequence of spike times is called a spike train. Figure 1.12 shows a spike train from 400 to 800 seconds. There were about 4300 spikes. Points have been jittered vertically to avoid overplotting. For the present, we're interested in the time between one spike and the next, the so-called interspike interval. Figure 1.12B shows three ways of representing the intervals. One is a histogram, rescaled to the density scale. The solid line is the density estimated from the approximately 4300 interspike intervals. The dashed line is the Exp(10.84) density. The Exponential density with $\lambda = 10.84$ represents the interspike intervals reasonably well.

There are at least two reasons why Exp(10.84) is not a perfect model for this neuron's spike train. Figure 1.13 shows one of them by displaying the neuron's spike train from 0 to a little more than 5000 seconds. The figure has light and dark



(a) spike train between 400 and 800 seconds

(b) density of interspike intervals. Solid line: density estimated from data. Dashed line: the Exp(10.84) density.

Figure 1.12: Spike train of neuron sig002a from 400 to 800 seconds.

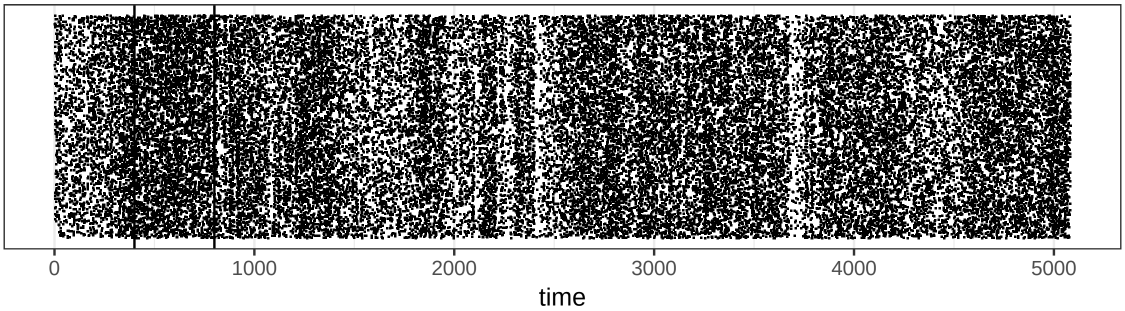


Figure 1.13: Spike train of neuron sig002a from 0 to around 5000 seconds with vertical lines at 400 and 800 seconds.

vertical strips that indicate times when the neuron is firing less or more often than usual. Example 1.6 examined only the time between the vertical lines at 400 and 800 seconds, when the neuron was firing at approximately a constant rate. An exponential density is good at describing interspike intervals when the firing rate is roughly constant. But when the firing rate is varying, more complicated models are needed.

Another reason Exp(10.84) is not a perfect model is that neurons have a refractory period, a short period of time after a spike during which they cannot have another spike. You can read about spikes and refractory periods at HTTP://WWW.PHYSIOLOGYWEB.COM/LECTURE_NOTES/NEURONAL_ACTION_POTENTIAL/NEURONAL_ACTION_POTENTIAL_REFRACTORY_PERIODS.HTML.

Example 1.6 raises a few questions, including the following.

1. How did we settle on $\lambda = 10.84$ as the value of λ to display?
2. Are there other values of λ that would also describe the data reasonably well? What are they? How can we find them?
3. The neuron in Example 1.6 was recorded in a live rat while the rat was licking a spout. The rat is sometimes given pure water to lick and sometimes other substances, like salty water. Neuron sig002a is in the rat's gustatory cortex, the brain's taste center, so might be expected to respond differently to different tastes. How can we assess the evidence that the value of λ changes when the rat licks different substances?

These questions will be addressed later in the book. **Section references**

Example 1.7 (VA lung cancer trial)
The US Veteran's Administration, or VA, conducted a study of a new treatment for lung cancer. Data for 137 patients can be uploaded from the file `VA.csv` on this book's

website. They are also in the R package `MASS`, which you can install and use, and were famously analyzed in the textbook by Kalbfleisch and Prentice (2011). The dataframe has 137 rows, for 137 patients, and 8 columns, for the variables

stime survival or follow-up time in days

status dead or censored

treat treatment: standard or test

age patient's age in years

Karn Karnofsky score of patient's performance on a scale of 0 to 100

diag.time times since diagnosis in months at entry to trial

cell one of four cell types

prior prior therapy?

The first several lines of the dataframe are

```
> head ( VA )
  stime status treat age Karn diag.time cell prior
1    72      1     1  69   60         7    1     0
2   411      1     1  64   70         5    1    10
3   228      1     1  38   60         3    1     0
4   126      1     1  63   60         9    1    10
5   118      1     1  65   70        11    1    10
6    10      1     1  49   20         5    1     0
```

Here we'll begin to explore whether the distribution of `stime` seems to vary according to which treatment is given. Figure 1.14 displays `stime` separately for patients receiving the standard and new treatments. For both treatments, the points are densest near `stime` $= 0$ and sparser as `stime` increases, so exponential distributions might model the data well. It is hard to tell from the figure whether the distribution of `stime` is appreciably different for the two groups. Figure 1.15 shows the `stime` density separately for each treatment. The histograms are on the density scale. The tick marks on the x-axis are called a *rug* and are the observed values of `stime`, one mark for each patient. The solid line is the Exp($1/121.6277$) density, the best fitting Exponential density when the patients from both treatments are combined. The dashed lines are the best fitting Exponential

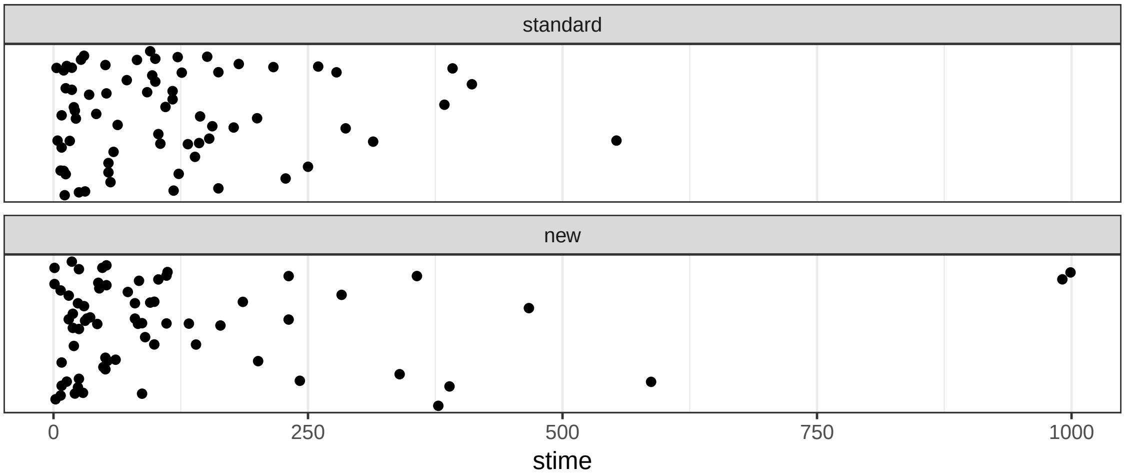


Figure 1.14: Survival time of patients in the VA lung cancer study under standard and new treatment

densities for each group of patients on their own. There is not much difference between the solid and dashed lines, so there does not appear to be much difference between the two treatments. We'll see in Exercise 12 of Chapter 2 how to quantify the improvement of the two-exponential description over the one-exponential description.

Figure 1.15 was made with the following code.

```
p + geom_histogram ( aes ( y = ..density.. ), binwidth = 50, center = 25 ) +
    geom_line ( data = df, aes ( x = stime, y = den.both ) ) +
    geom_line ( data = df, aes ( x = stime, y = den.each ), lty = 2 ) +
    geom_rug() +
    facet_wrap ( ~ treat,
                 ncol = 1,
                 labeller = labeller ( treat = c ( "1" = "standard", "2" = "new" ) )
               ) +
    theme_bw()
```

There are three new commands. One is `geom_rug`, which adds a rug to the plot. A *rug* is a set of tick marks on the x-axis that indicate actual data points. Each tick mark is supposedly like a fiber in a rug. Another is `facet_wrap`, which makes one facet of the plot for each value of `treat`. The other new command is `labeller`. We use it here because, in the `treat` column of the dataframe, the standard treatment is called "1" and the new treatment is called "2". We want them to be called "standard" and "new" in the plot. Those become the labels of the facets.

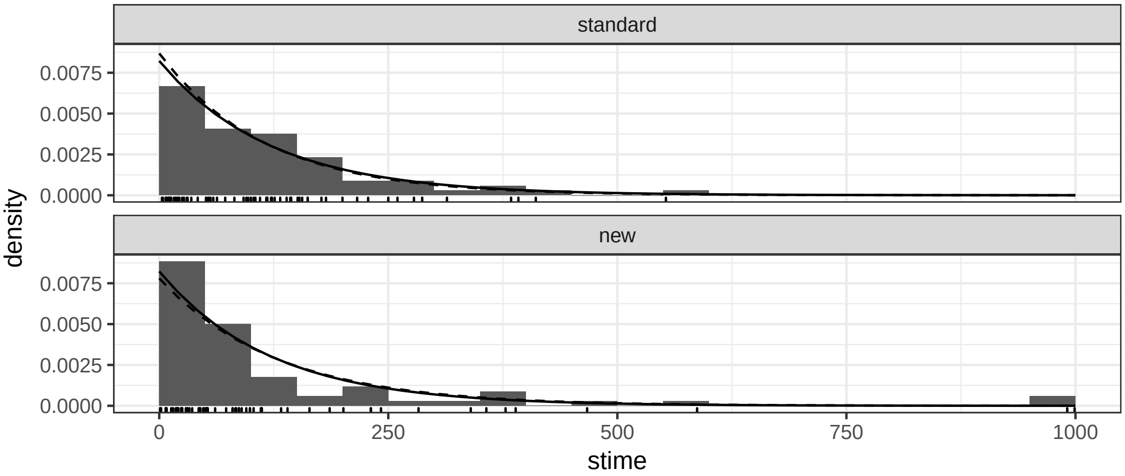


Figure 1.15: Survival time density of patients in the VA lung cancer study under standard and new treatment. Histograms on the density scale. Solid line: best Exponential density for both treatments combined. Dashed line: best Exponential density for each treatment on its own. Tick marks: observed `stimes`.

1.6 The Normal Density

We often deal with random variables X that are densest near some central value and become sparser as x moves away from the central value in either direction. Some examples are

Biological Anthropology Heights of people

Oceanography Ocean temperatures at a particular location

Quality Control Diameters of ball bearings

Education SAT scores

In each case the random variable is expected to have a central value around which most of the observations cluster. Fewer and fewer observations are farther and farther away from the center. So the density should be unimodal — large in the center and decreasing in both directions away from the center. A useful density for such situations is the *Normal density*

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{1.10}$$

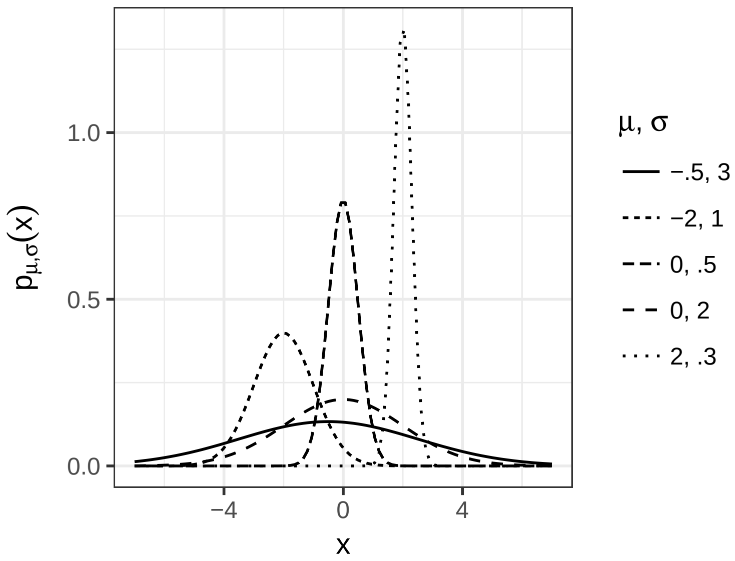


Figure 1.16: Normal densities

where μ is the central value and σ describes how quickly the density falls away from the center. We say X *has a Normal distribution with mean* μ *and standard deviation* σ and write $X \sim \mathrm{N}(\mu, \sigma)$. There is a family of Normal densities, one for combination (μ, σ). Figure 1.16 shows Normal densities for several different values of (μ, σ). As illustrated by the figure, the mean μ controls the center of the density; each density is centered over its own value of μ. The standard deviation σ controls the spread. Densities with larger values of σ are more spread out; densities with smaller σ are tighter. In every case, the Normal density is symmetric around its mean.

Figure 1.16 was created with the following code.

```
mu <- c ( -2, 0, 0, 2, -.5 )
sigma <- c ( 1, 2, .5, .3, 3 )
x <- seq ( -7, 7, length = 100 )

df <- data.frame ( mu = rep ( mu, each = length ( x ) ),
                   sigma = rep ( sigma, each = length ( x ) ),
                   x = rep ( x, length ( mu ) )
                 )
df$density <- with ( df, dnorm ( x, mean = mu, sd = sigma ) )

p <- ggplot ( df, aes ( x = x, y = density, linetype = paste ( mu, sigma ) ) )
p + geom_line() +
    ylab ( expression ( p[mu*','*sigma] (x) ) ) +
    scale_linetype ( name = expression ( mu*','~sigma ),
                     labels = c ( "-.5, 3", "-2, 1", "0, .5", "0, 2", "2, .3"  )
                   ) +
    theme_bw()
```

We wanted to display the Normal density for 5 combinations of (μ, σ), so we began by specifying those values. Next we specified 100 values of x between -7 and 7 at which to evaluate and plot the five densities. Then we put all those `mu`'s, `sigma`'s and `x`'s into a dataframe. Each `mu`, and each `sigma` has to be repeated 100 times, for the 100 values of `x`. If you don't understand the code, it might be helpful for you to print out the first and last several rows of `df`. Once we have the dataframe with `mu`, `sigma`, and `x`, we used the `dnorm` command to create a new column called `density`. You have to tell `dnorm` at which values of `x`, `mu`, and `sigma` you want to calculate the density. We also introduced a new trick, the `with` command. Instead of writing `dnorm ( x, mean = df$mu, sd = df$sigma )` we wrote `with ( df, dnorm ( x, mean = mu, sd = sigma ) )`, which tells R to use the columns `x`, `mu`, and `sigma` within the dataframe `df`.

We also used a new trick in the `aes` part of the `ggplot` command. We said `linetype = paste ( mu, sigma )`, which tells `ggplot` to paste together the values of `mu` and `sigma` to determine which linetype to use. It uses a different linetype for each combination of `(mu, sigma)`.

Figure 1.17 provides an illustration. It shows the density of Galton's parents' heights from Example 1.4 estimated with 1-inch bins, just as in Figure 1.6, along with an approximating Normal density. The Normal density is a reasonably good approximation and may be adequate for some purposes. However, the density estimated from the data has a dip around 67–68 inches, suggesting that we might better describe the data as two populations, presumably one for mothers and one for fathers, which could each be approximated with its own Normal density. We shall see later, in Example 2.5, how to compare the single-Normal approximation to the two-Normal approximation.

1.7 Cumulative Distribution Functions

In Statistics we often want to talk about the probability that a random variable lies in a particular range. For example,

- The probability that eight or more employees at the Slater school (Example 1.2) develop cancer: $\Pr[X \geq 8]$.
- The probability that a given quadrat (Example 1.3) has fewer than 4 seedlings: $\Pr[X < 4]$.
- The probability that a parent's height (Example 1.4) is less than 62 or greater than 72 inches: $\Pr[X < 62 \text{ or } X > 72]$.

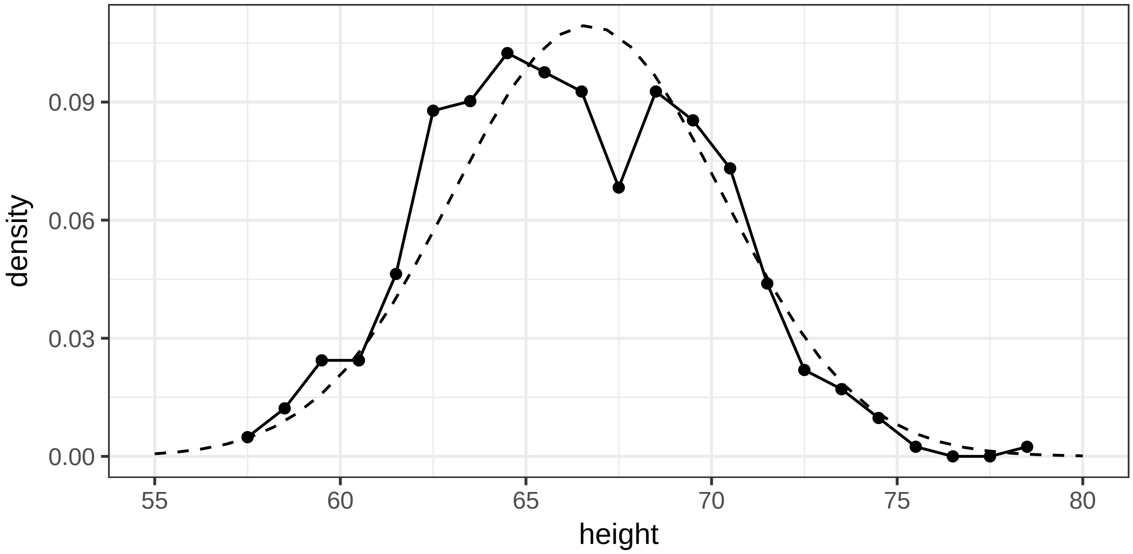


Figure 1.17: Galton's parents' heights with Normal approximation. Solid line: density estimated from 1-inch bins. Dashed line: The $N_{66.66,3.645}$ density.

- The probability that a given day's S&P 500 return (Example 1.5) is between -0.01 and 0.01: $\Pr[X \in (-0.01, 0.01)]$.

- The probability that a neuron (Example 1.6) fires in the next tenth of a second: $\Pr[X < 0.10]$.

- The probability that a lung cancer patient (Example 1.7) survives more than 1 year: $\Pr[X > 365]$.

The fundamental quantity for calculating such probabilities is the *cumulative distribution function*, or cdf. For any random variable X, the definition is

$$F(x) \equiv \Pr[X \leq x] \tag{1.11}$$

where X is the random variable, x is a specific number, and F is typical statistics notation for the cdf. The probabilities in the preceding list can all be written in terms of F:

$$\begin{gathered} 1 - F(7) \\ F(3) \\ F(61) + (1 - F(72)) \\ F(0.01) - F(-0.01) \\ F(0.1) \\ 1 - F(365) \end{gathered} \tag{1.12}$$

The probability that a random variable X lies in a specified interval, or set of intervals, can always be written as addition or subtraction of F evaluated at the endpoints of the intervals.

The notation in (1.12) is imprecise because F is different on each line: on the first line it refers to probabilities in the Slater school in Example 1.2; on the second line it refers to probabilities of seedlings in Example 1.3; and so on. When it's necessary to be more specific — i.e., when it's not obvious from the context — we will use subscripts on F.

R has built-in functions for computing the cdf of many families of distributions. Here's how they work for the distributions we've learned so far.

Binomial distribution In Example 1.2, the probability that 8 or more employees develop cancer depends on the cancer rate. Assuming the rate at Slater is the same as the national rate, we would write

$$\begin{aligned} &\mathrm{Pr}_{p=.03}[\text{8 or more employees develop cancer}] \\ =&1 - \mathrm{Pr}_{p=.03}[\text{7 or fewer employees develop cancer}] \\ =&1 - F_{p=.03}(7). \end{aligned}$$

In R we would enter `1 - pbinom ( 7, 145, .03 )`. R's convention is that `p` stands for *probability* and `binom` says which distribution we want. So `pbinom` computes the cdf for the Binomial distribution. The `( 7, 145, .03 )` says to compute $F(7)$ under the conditions $N = 145$ and $p = 0.03$.

Poisson distribution In Example 1.3, the probability that a given quadrat has fewer than 4 seedlings depends on λ, the arrival rate of new seedlings. If we want to compare, say, $\lambda = .5$ to $\lambda = 1$ to $\lambda = 2$, then we would want

$$\begin{gathered} \mathrm{Pr}_{\lambda=.5}[X \leq 3] = F_{\lambda=.5}(3) \\ \mathrm{Pr}_{\lambda=1}[X \leq 3] = F_{\lambda=1}(3) \\ \mathrm{Pr}_{\lambda=2}[X \leq 3] = F_{\lambda=2}(3) \end{gathered}$$

In R we would enter `ppois(3,.5)`, `ppois(3,1)`, and `ppois(3,2)` and get the answers 0.9982484, 0.9810118, and 0.8571235.

Here are two R tricks that can make your life easier.

1. You can enter multiple commands on one line if you separate them with a semicolon. That is, you can enter `ppois ( 3, .5 ); ppois ( 3, 1 ); ppois ( 3, 2 )` on one line.
2. You can often enter vectors instead of individual numbers into R commands. In this example you can enter `ppois ( 3, c ( .5, 1, 2 ) )`.

Exponential distribution In Example 1.6 we might be interested in the probability the neuron fires in the next tenth of a second. That probability depends on the firing rate λ. We would write $\Pr_\lambda[X < 0.1] = F_\lambda(0.1)$. We saw Exp(10.84) approximates the firing density well, so in R we would write `pexp ( 0.1, 10.84 )` and learn there is about a 2/3 chance the neuron fires in the next tenth of a second.

In Example 1.7 we might be interested in the probability a particular patient survives at least a year. We saw the Exp(1/121.6277) density models the data reasonably well, so we might want $\Pr_{\lambda=1/121.6277}[X > 364] = 1 - F_{1/121.6277}(364)$. In R we would write `1 - pexp ( 364, 1/121.6277 )` and learn there is about a 5% chance of surviving at least a year.

Normal distribution In Figure 1.17 we saw the $\mathrm{N}(66.66, 3.645)$ density was a reasonably good model for the heights of parents in Example 1.4. For the probability that a parent's height is less than 62 or greater than 72 inches we would write `pnorm ( 62, 66.66, 3.645 ) + 1 - pnorm ( 72, 66.66, 3.645 )`.

When working with densities, probability can be visualized as area under the density curve. Figure 1.18 illustrates in the context of Example 1.4. The shaded area is $F(62) + (1 - F(72)) = \Pr[X \leq 62 \text{ or } X > 72]$.

When working with probabilities that are not densities, the same visualization still often works, at least to a good approximation, as illustrated in Figure 1.19. Each rectangle outlined in dots and dashes has width 1, so the area of the rectangle is equal to its height. Therefore,

$$\begin{aligned}
F_{p=.03}(10) &\equiv \Pr_{p=.03}[X \leq 10] \\
&= \text{height of bar } 1 + \cdots + \text{height of bar } 11 \\
&= \text{area of rectangle } 1 + \cdots + \text{area of rectangle } 11 \\
&= \text{area under dashed curve}
\end{aligned}$$

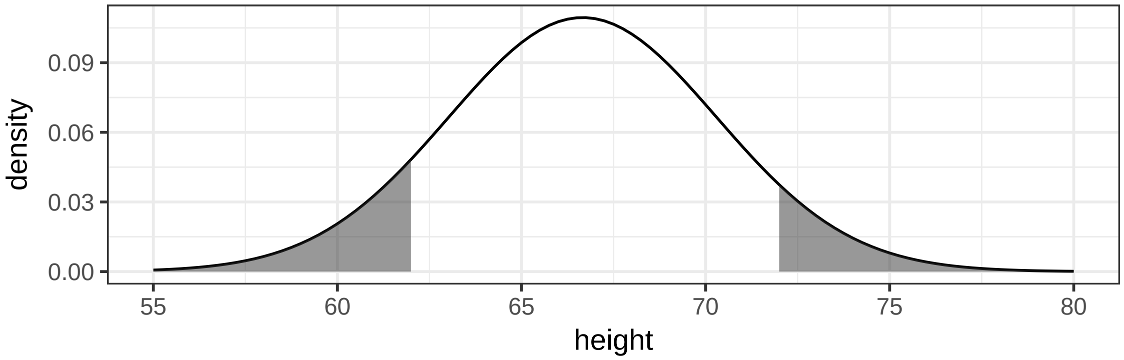


Figure 1.18: Normal approximation to Galton's parents' heights, same as dashed line in Figure 1.17. $\Pr[X \leq 62 \text{ or } X > 72]$ = area of shaded region.

When working with densities, probability is area under the curve, so in Figure 1.18 $\Pr[X = 62]$ is the area of the line segment above the point $x = 62$, the boundary between the shaded and unshaded regions. That area is 0, so $\Pr[X = 62] = 0$. In fact, for any height x, $\Pr[X = x \text{ exactly}] = 0$, according to the density. Thus, it is also true that

$$F(62) \equiv \Pr[X \leq 62] = \Pr[X < 62]$$

and it doesn't matter whether we write "$\leq$" or "$<$". Similarly, it doesn't matter whether we write "$\Pr[X > 72]$" or "$\Pr[X \geq 72]$". In practice, heights, or any other measurement, are not recorded with infinite precision, so $\Pr[\text{recorded height} = x \text{ exactly}] \neq 0$. When working with probabilities that are not densities, as in Figure 1.19, $\Pr[\text{exactly } x \text{ employees develop cancer}] \neq 0$, so it makes a little bit of difference whether we write $\Pr[X < 10]$ or $\Pr[X \leq 10]$.

1.8 Joint, Marginal, and Conditional Distributions

Statisticians often have to deal with the probabilities of several events, quantities, or random variables simultaneously. For example, we may classify voters in a city according to political party affiliation and support for a school bond referendum. Let A and S be a voter's affiliation and support, respectively.

$$A = \begin{cases} D & \text{if Democrat} \\ R & \text{if Republican} \end{cases} \qquad \text{and} \qquad S = \begin{cases} Y & \text{if in favor} \\ N & \text{if opposed} \end{cases}$$

Suppose a polling organization finds that 80% of Democrats and 35% of Republicans favor the bond referendum. The 80% and 35% are called *conditional* probabilities

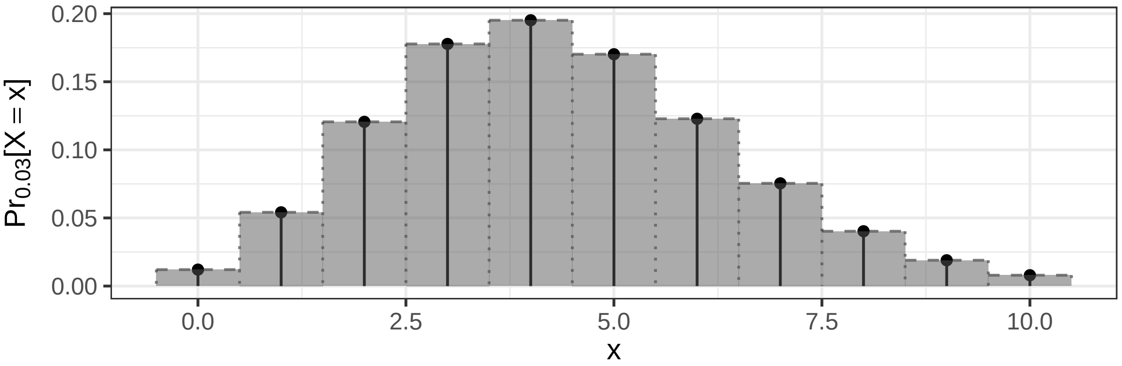


Figure 1.19: $\Pr_{p=.03}[X = x]$ in the Slater School (Example 1.2)

because they are conditional on party affiliation. The notation for conditional probabilities uses a vertical bar, "$\,|\,$". Specifically,

$$\begin{aligned} \Pr[S = Y \mid A = D] &= 0.80; & \Pr[S = N \mid A = D] &= 0.20; \\ \Pr[S = Y \mid A = R] &= 0.35; & \Pr[S = N \mid A = R] &= 0.65. \end{aligned}$$

The vertical bar is read "given." We say "the conditional probability that $S = N$ given $A = D$ is 0.20", etc.

Suppose further that 60% of voters in the city are Democrats. Then 80% of 60% = 48% of the voters are Democrats who favor the referendum. The 48% is called a *joint* probability because it is the probability of $(A = D, S = Y)$ jointly. The notation is $\Pr[A = D, S = Y] = .48$. Likewise, $\Pr[A = D, S = N] = .12$; $\Pr[A = R, S = Y] = .14$; and $\Pr[A = R, S = N] = 0.26$. Table 1.1 summarizes the probabilities. The quantities .60, .40, .62, and .38 are called *marginal* probabilities. The name comes from their being written in the margins of the table. Marginal probabilities are probabilities for one variable alone, the ordinary probabilities we've been talking about all along.

Joint probabilities can be written as the product of marginal and conditional probabilities. In our example,

$$\Pr[A = D, S = Y] = \Pr[A = D] \times \Pr[S = Y \mid A = D].$$

The general rule is that for random variables X and Y, and specific values x and y,

$$\Pr[X = x, Y = y] = \Pr[X = x] \times \Pr[Y = y \mid X = x].$$

	For	Against	Total
Democrat	48%	12%	60%
Republican	14%	26%	40%
Total	62%	38%	100%

Table 1.1: Party Affiliation and Referendum Support

The event $A = D$ can be partitioned into the two smaller events $(A = D, S = Y)$ and $(A = D, S = N)$. So

$$\Pr[A = D] = \Pr[A = D, S = Y] + \Pr[A = D, S = N] = .48 + .12 = .60.$$

(The numbers come from Table 1.1.) The event $A = R$ can be partitioned similarly. Too, the event $S = Y$ can be partitioned into $(A = D, S = Y)$ and $(A = R, S = Y)$, so

$$\Pr[S = Y] = \Pr[A = D, S = Y] + \Pr[A = R, S = Y] = .48 + .14 = .62.$$

These calculations illustrate another general rule: *To get a marginal probability for one variable, add the joint probabilities for all values of the other variable.* To illustrate further, let's make our political polling example more complicated by supposing there are four major political parties: Democrat (D), Republican (R), Libertarian (L), and Green (G). Then,

$$\begin{aligned}\Pr[S = Y] = \Pr[S = Y, A = D] &+ \Pr[S = Y, A = R] \\ &+ \Pr[S = Y, A = L] + \Pr[S = Y, A = G]\end{aligned}$$

Sometimes we know joint probabilities and need to find marginals and conditionals; sometimes it's the other way around. And sometimes we know probabilities for X and for $Y \mid X$ and need to find probabilities for Y or $X \mid Y$. The following story is an example of the latter. It is a common problem in drug testing, disease screening, polygraph testing, and many other fields.

The participants in an athletic competition are tested for steroid use. The test is 90% accurate in the following sense: for athletes who use steroids, the test has a 90% chance of returning a positive result; for non-users, the test has a 90% chance of returning a negative result. Suppose that only 30% of athletes use steroids. An athlete is randomly selected and her test returns a positive result. What is the probability she is a steroid user?

This is a story of two random variables, steroid use U and test result T. Let

$$U = \begin{cases} 1 & \text{if the athlete uses steroids} \\ 0 & \text{if not} \end{cases} \quad \text{and} \quad T = \begin{cases} 1 & \text{if the test is positive} \\ 0 & \text{if not} \end{cases}$$

	$T = 0$	$T = 1$	Total
$U = 0$	.63	.07	.70
$U = 1$	.03	.27	.30
Total	.66	.34	1.00

Table 1.2: Steroid Use and Test Results

We're given $\Pr[U = 1] = 0.3$; $\Pr[T = 1 \,|\, U = 1] = 0.9$; and $\Pr[T = 1 \,|\, U = 0] = 0.1$. We want $\Pr[U = 1 \,|\, T = 1]$. The calculation are

$$\begin{aligned}
\Pr[U = 1 \,|\, T = 1] &= \frac{\Pr[U = 1, T = 1]}{\Pr[T = 1]} \\
&= \frac{\Pr[U = 1, T = 1]}{\Pr[U = 1, T = 1] + \Pr[U = 0, T = 1]} \\
&= \frac{\Pr[U = 1] \times \Pr[T = 1 \,|\, U = 1]}{\Pr[U = 1] \times \Pr[T = 1 \,|\, U = 1] + \Pr[U = 0] \times \Pr[T = 1 \,|\, U = 0]} \\
&= \frac{0.3 \times 0.9}{0.3 \times 0.9 + 0.7 \times 0.1} = \frac{.27}{.27 + .07} = \frac{.27}{.34} \approx .79.
\end{aligned}$$

The calculation works by writing the thing we want in terms of the things we know. The result is that even though the test is 90% accurate, an athlete who tests positive has only about an 80% chance of being a steroid user. If that doesn't seem intuitively reasonable, think of a large number of athletes, say 100. About 30 will be steroid users of whom about 27 will test positive. About 70 will be non-users of whom about 7 will test positive. So there will be about 34 athletes who test positive, of whom about 27, or about 80%, will be users.

Table 1.2 is another representation of the steroid problem. It is important to become familiar with the concepts and notation of marginal, conditional, and joint distributions, and not rely too heavily on tables, because in more complicated problems there may not be a convenient table.

Example 2.4 is a further illustration of joint, conditional, and marginal distributions.

Example 1.8 (Seedlings, continued)

Example 1.3 introduced an observational study at the Coweeta Ecological Research Station to learn about the rate of seedling production and survival. For a particular quadrat in a particular year, let N be the number of new seedlings that emerge and suppose $N \sim \mathsf{Poi}(\lambda)$ describes the data well for some $\lambda > 0$. Each seedling either dies over

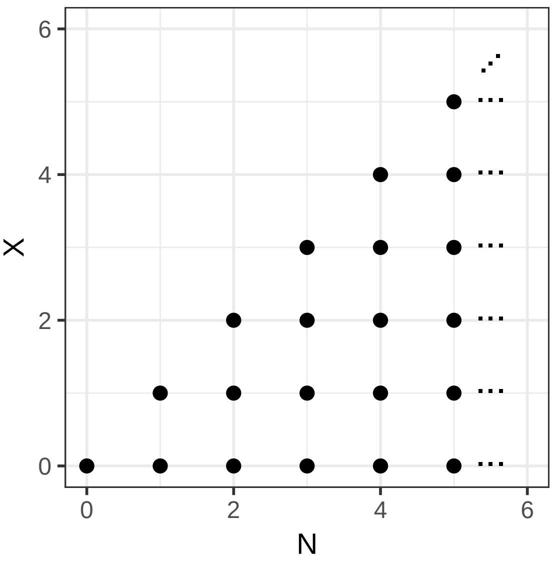


Figure 1.20: Permissible values of N and X in the seedlings example. Permissible values continue up and to the right.

the winter or survives to become an established seedling the next year. Let p be the probability of survival and X the number of seedlings that survive. Suppose that the survival of any one seedling is not affected by the survival of any other seedling. Then $X \sim \text{Bin}(N, p)$ is a reasonable model for X, if we know N. That's the conditional distribution of $X \mid N$. More formally, we would write $X \mid N \sim \text{Bin}(N, p)$. Since we know the formula (Equation (1.5)) for Binomial probabilities we could also write

$$\Pr_p[X = x \mid N = n] = \binom{n}{x} p^x (1-p)^{n-x}$$

Figure 1.20 shows the possible values of the pair (N, X). Of course $X \leq N$.

What are the joint, marginal, and conditional probabilities? They depend on the emergence rate λ and the survival probability p, so let's suppose those have been specified. Our model says $N \sim \text{Poi}(\lambda)$. That's a marginal distribution. The probabilities are (Equation (1.8))

$$\Pr_\lambda[N = n] = \frac{\lambda^n e^{-\lambda}}{n!}.$$

You should be able to plug in a value of λ and compute these probabilities with R's `dpois` command.

Our model also says $X \mid N \sim \text{Bin}(N, p)$. That's $\Pr_p[X = x \mid N = n] = \binom{n}{x} p^x (1 - p)^{n-x}$. Since our model specifies marginal probabilities for N and conditional probabilities for $X \mid N$, we can calculate joint probabilites:

$$\begin{aligned}\Pr_{p,\lambda}[N = n, X = x] &= \Pr_\lambda[N = n] \times \Pr_p[X = x \mid N = n] \\ &= \frac{\lambda^n e^{-\lambda}}{n!} \times \binom{n}{x} p^x (1-p)^{n-x}.\end{aligned}$$

You should be able to compute these probabilities with R's `dpois` and `dbinom` commands by plugging in a value of λ and a value of p.

The preceding formula is for values (n, x) where $x \le n$. For a value $x > n$, $\Pr[N = n, X = x] = 0$.

Another good example is the dataset `iris`, which comes with R and contains the lengths and widths of sepals and petals of 150 iris plants. The first and last several lines of `iris` are

```
> rbind ( head(iris), tail(iris) )
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1            5.1         3.5          1.4         0.2    setosa
2            4.9         3.0          1.4         0.2    setosa
3            4.7         3.2          1.3         0.2    setosa
4            4.6         3.1          1.5         0.2    setosa
5            5.0         3.6          1.4         0.2    setosa
6            5.4         3.9          1.7         0.4    setosa
145          6.7         3.3          5.7         2.5 virginica
146          6.7         3.0          5.2         2.3 virginica
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

As usual, `?iris` gives information about the data.

Figure 1.21 shows the distribution of petal length. There seem to be two groups of irises: about a third of the plants have petals about 1–2 centimeters in length, while the other two-thirds have petals about 3–7 centimeters long. Figure 1.22 helps explain why there are seemingly two groups. In fact, there are three groups, one for each species. *setosa* petals are about 1–2 cm; *versicolor* petals are about 3–5 cm; and *virginica* petals are about 4.5–7 cm. There is a little bit of overlap between *versicolor* and *virginica*, but none with *setosa*.

How did we know that distinguishing irises by species would reveal something interesting? We didn't. We tried plotting the data several ways to see what we could see. Of course, species in an obvious candidate for a variable that *might* matter, but we didn't know in advance that it *would* matter. Part of good statistical practice is looking at the data in various ways to see what emerges. Good statisticians don't follow a script; they use intelligence guided by experience. In the iris data, intelligence and experience both suggested that species might matter, so we tried it.

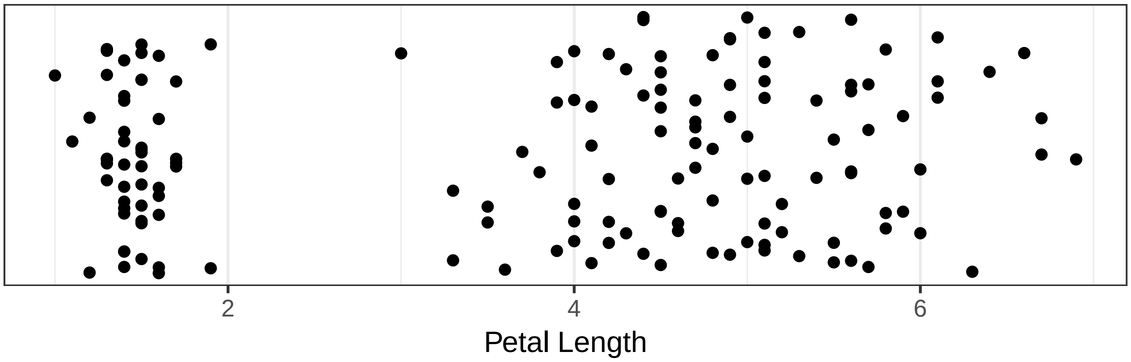


Figure 1.21: Lengths of petals of 150 iris plants

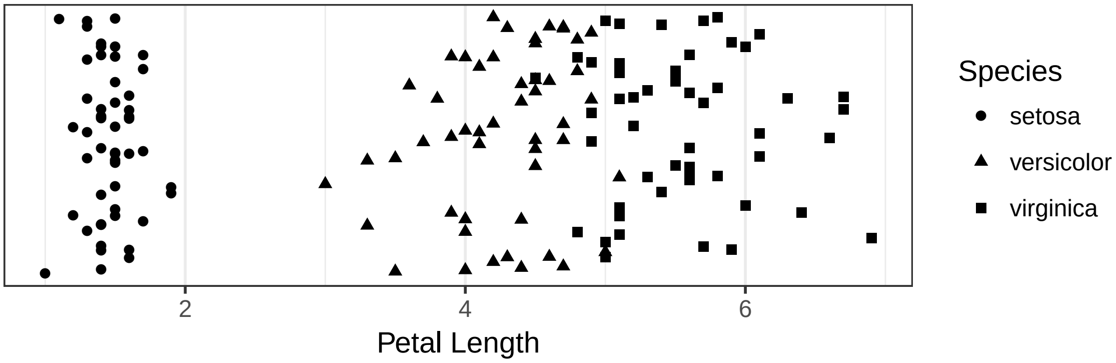


Figure 1.22: Lengths of petals of 150 iris plants, by species

We can describe the situation with two random variables. One, `Species`, has three possible values. The other, `Petal.Length`, can be any positive number. Figure 1.23 shows the petal lengths separated by species, with approximating Normal densities. The solid curve is the approximate density for all plants combined. It is the same in each facet. It is not a good approximation to the points shown in Figure 1.21. The dashed curves are approximate densities for each species separately. They do seem to be good approximations to the points in Figure 1.22, or to the rugs in Figure 1.23, which display each species on its own.

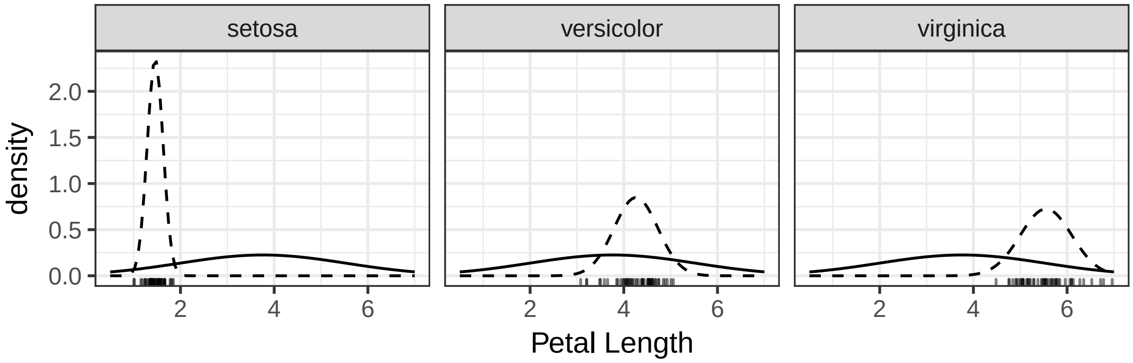


Figure 1.23: Petal length of iris plants with rug plots and approximating Normal densities. Solid curve: Normal density for all plants combined. Dashed curves: Normal densities for each species separately.

Figures 1.15 and 1.23 illustrate an important idea in statistics. In Figure 1.15 the density of survival time seems to be about the same for each group:

$$p(\text{survival time} \mid \text{treatment} = \text{standard}) = p(\text{survival time} \mid \text{treatment} = \text{new}).$$

But in Figure 1.23 the density of petal length seems to be different for each group:

$$\begin{aligned} &p(\text{petal length} \mid \text{species} = \text{setosa}) \\ \neq &p(\text{petal length} \mid \text{species} = \text{versicolor}) \\ \neq &p(\text{petal length} \mid \text{species} = \text{virginica}). \end{aligned}$$

In mathematics and statistics, p is a common notation for probability density.

More generally, in a problem with two random variables X and Y, we can ask whether the distribution of X depends on the value of Y. If it does not, then we say X and Y are *independent*. If it does, then X and Y are *dependent*. Figure 1.15 suggests that treatment and survival time are independent. Figure 1.23 suggests that species and petal length are dependent. In Table 1.1, party affiliation and referendum support are dependent because $\Pr[S = Y \mid A = D] \neq \Pr[S = Y \mid A = R]$. In Table 1.2, U and T are dependent because $\Pr[U = 1 \mid T = 1] \neq \Pr[U = 1 \mid T = 0]$.

If the distribution of X depends on the value of Y, then also the distribution of Y depends on the value of X. So when checking whether X and Y are dependent, it doesn't matter whether we compare $p(X \mid Y = y)$ for different values of y or $p(Y \mid X = x)$ for different values of x.

When X and Y are independent, then the marginal distribution of X is the same as the conditional distribution of $X \mid Y$. In symbols, $p(x) = p(x \mid y)$ for densities or $\Pr[X = x] = \Pr[X = x \mid Y = y]$ for probabilities, for all values of x and y. Also, $p(y) = p(y \mid x)$ and $\Pr[Y = y] = \Pr[Y = y \mid X = x]$.

When X and Y are independent, then knowing X gives no information about Y and knowing Y gives no information about X. We used that idea in Example 1.1 to calculate the probability of winning a game of craps on the come-out roll. For a fair die, $\Pr[1] = \cdots = \Pr[6] = 1/6$. For a pair of dice, (D_1, D_2), and for any particular values (d_1, d_2),

$$\begin{aligned}
\Pr[D_1 = d_1, D_2 = d_2] &= \Pr[D_1 = d_1] \times \Pr[D_2 = d_2 \mid D_1 = d_1] \\
&= \Pr[D_1 = d_1] \times \Pr[D_2 = d_2] \\
&= (1/6) \times (1/6) \\
&= 1/36.
\end{aligned}$$

The same idea applies to Figure 1.15, which suggests that knowing which treatment was applied gives no information about how long the patient is likely to survive and that the new treatment is no better and no worse than the standard treatment.

But when X and Y are dependent, then knowing X does give information about Y. E.g., in the `iris` data, knowing the species does give information about the likely petal length. *setosa* petals are likely to be shorter than *versicolor* petals, which in turn are likely to be shorter than *virginica* petals. Similarly, if we know an iris's petal is shorter than about 2 cm, then we can be fairly sure the species is *setosa*.

Do not confuse independent with mutually exclusive. Let X denote the outcome of a die roll and let

$$A = \begin{cases} 1 & \text{if } X \in \{1, 2, 3\} \\ 0 & \text{if } X \in \{4, 5, 6\}. \end{cases}$$

A is called an *indicator* variable because it indicates the occurrence of a particular event.

There is a special notation for indicator variables:

$$A = \mathbf{1}_{\{1,2,3\}}(X).$$

$\mathbf{1}_{\{1,2,3\}}$ is an *indicator function.* $\mathbf{1}_{\{1,2,3\}}(X)$ is either 1 or 0 according to whether X is in the subscript.

Let $B = \mathbf{1}_{\{4,5,6\}}(X)$, $C = \mathbf{1}_{\{1,3,5\}}(X)$, $D = \mathbf{1}_{\{2,4,6\}}(X)$ and $E = \mathbf{1}_{\{1,2,3,4\}}(X)$. A and B are dependent because $\Pr[A] = .5$ but $\Pr[A \mid B] = 0$. D and E are independent because $\Pr[D] = \Pr[D \mid E] = .5$. You can also check that $\Pr[E] = \Pr[E \mid D] = 2/3$. Do not confuse dependence with causality. B and D are dependent, but neither causes the other.

Example 1.9 uses joint, conditional, and marginal probabilities to calculate the probability of winning at craps.

Example 1.9 (Craps, continued)
We'll continue Example 1.1 by calculating the probability of winning at craps. Example 1.1 explains the rules.

$$\begin{aligned}&\Pr[\text{winning at craps}] = \\ &\qquad \Pr[\text{winning on the come-out roll}] + \Pr[\text{winning after the come-out roll}].\end{aligned}$$

Example 1.1 calculated the first term, so here we'll concentrate on the second.

$$\begin{aligned}
&\Pr[\text{winning after the come-out roll}] \\
&= \quad \Pr[\text{come-out} = 4, \text{win}] + \Pr[\text{come-out} = 5, \text{win}] + \Pr[\text{come-out} = 6, \text{win}] \\
&\quad + \Pr[\text{come-out} = 8, \text{win}] + \Pr[\text{come-out} = 9, \text{win}] + \Pr[\text{come-out} = 10, \text{win}] \\
&= \quad \Pr[\text{come-out} = 4] \times \Pr[\text{win} \mid \text{come-out} = 4] \\
&\quad + \Pr[\text{come-out} = 5] \times \Pr[\text{win} \mid \text{come-out} = 5] \\
&\quad + \Pr[\text{come-out} = 6] \times \Pr[\text{win} \mid \text{come-out} = 6] \\
&\quad + \Pr[\text{come-out} = 8] \times \Pr[\text{win} \mid \text{come-out} = 8] \\
&\quad + \Pr[\text{come-out} = 9] \times \Pr[\text{win} \mid \text{come-out} = 9] \\
&\quad + \Pr[\text{come-out} = 10] \times \Pr[\text{win} \mid \text{come-out} = 10]
\end{aligned}$$

Since the terms for 4, 5, 6, 8, 9, and 10 all have the same form, we examine just the term for 4 more carefully.

$$\Pr[\text{come-out} = 4] = \Pr[(1,3)] + \Pr[(2,2)] + \Pr[(3,1)] = \frac{3}{36} = \frac{1}{12}.$$

Let $p = \Pr[\text{win} \,|\, \text{come-out} = 4]$ To calculate p, it helps to consider the next roll after the come-out. Call it X. We partition the event "win", into three smaller events to get

$$\begin{aligned}
p = \;& \Pr[\text{win}, X = 4 \,|\, \text{come-out} = 4] \\
& + \Pr[\text{win}, X = 7 \,|\, \text{come-out} = 4] \\
& + \Pr[\text{win}, X = \text{other} \,|\, \text{come-out} = 4] \\
= \;& \Pr[X = 4 \,|\, \text{come-out} = 4] \times \Pr[\text{win} \,|\, X = 4, \text{come-out} = 4] \\
& + \Pr[X = 7 \,|\, \text{come-out} = 4] \times \Pr[\text{win} \,|\, X = 7, \text{come-out} = 4] \\
& + \Pr[X = \text{other} \,|\, \text{come-out} = 4] \times \Pr[\text{win} \,|\, X = \text{other}, \text{come-out} = 4] \quad \text{(Why?)} \\
= \;& \Pr[X = 4] \times \Pr[\text{win} \,|\, X = 4, \text{come-out} = 4] \\
& + \Pr[X = 7] \times \Pr[\text{win} \,|\, X = 7, \text{come-out} = 4] \\
& + \Pr[X = \text{other}] \times \Pr[\text{win} \,|\, X = \text{other}, \text{come-out} = 4] \quad \text{(Why?)} \\
= \;& \Pr[X = 4] \times 1 \\
& + \Pr[X = 7] \times 0 \\
& + \Pr[X = \text{other}] \times p \qquad \text{(Why?)} \\
= \;& \frac{3}{36} + \frac{27}{36} \times p \qquad \text{(Why?)}
\end{aligned}$$

Solving the previous equation yields $p = 1/3$. Here's some intuition behind the $1/3$. There are 9 rolls that determine whether you win or lose: $(1,3)$, $(2,2)$, $(3,1)$, $(1,6)$, $(2,5)$, $(3,4)$, $(4,3)$, $(5,2)$, and $(6,1)$. Any other rolls leave you in the same situation you were already in. Of the 9, 3 are winners and 6 are losers. So the probability of winning is 3/9 and the probability of losing is 6/9.

Similar calculations yield

$$\begin{aligned}
\Pr[\text{win} \,|\, \text{come-out} = 4] &= \frac{1}{3} \\
\Pr[\text{win} \,|\, \text{come-out} = 5] &= \frac{2}{5} \\
\Pr[\text{win} \,|\, \text{come-out} = 6] &= \frac{5}{11} \\
\Pr[\text{win} \,|\, \text{come-out} = 8] &= \frac{5}{11} \\
\Pr[\text{win} \,|\, \text{come-out} = 9] &= \frac{2}{5} \\
\Pr[\text{win} \,|\, \text{come-out} = 10] &= \frac{1}{3}
\end{aligned}$$

Putting everything together gives

$$
\begin{aligned}
&\Pr[\text{winning at craps}]\\
&\quad= \Pr[\text{winning on the come-out roll}] + \Pr[\text{winning after the come-out roll}]\\
&\quad= 2/9 + \Pr[\text{winning after the come-out roll}]\\
&\quad= 2/9 + \Pr[\text{come-out} = 4, \text{win}] + \Pr[\text{come-out} = 5, \text{win}] + \Pr[\text{come-out} = 6, \text{win}]\\
&\qquad+ \Pr[\text{come-out} = 8, \text{win}] + \Pr[\text{come-out} = 9, \text{win}] + \Pr[\text{come-out} = 10, \text{win}]\\
&\quad= 2/9 + \frac{3}{36}\cdot\frac{1}{3} + \frac{4}{36}\cdot\frac{2}{5} + \frac{5}{36}\cdot\frac{5}{11} + \frac{5}{36}\cdot\frac{5}{11} + \frac{4}{36}\cdot\frac{2}{5} + \frac{3}{36}\cdot\frac{1}{3} \approx 0.493.
\end{aligned}
$$

Craps is a very fair game; the house has only a slight edge.

1.9 Simulation

Computer simulations can be very helpful in calculating probabilities. Section 1.9 illustrates some of the possibilities. We begin with a simulation to estimate $\Pr[7 \text{ or } 11]$, the probability of winning on the come-out roll in Example 1.1. We should get $2/9$, just as we calculated. The point is not to estimate the probability, but to learn how to do simulations.

Example 1.10 (Craps, continued)
I wrote the following R script to estimate the probability of winning on the come-out roll.

```
> sim.size <- 10
> D1 <- sample ( 1:6, sim.size, replace = TRUE )
> D2 <- sample ( 1:6, sim.size, replace = TRUE )
> tot <- D1 + D2
> win <- ifelse ( tot == 7 | tot == 11, TRUE, FALSE )
> est.prob <- sum ( win ) / sim.size
> est.prob
[1] 0.7
>
```

First I set the simulation size to be 10. That's the number of come-out rolls I'm going to simulate. In practice, you would use a larger simulation, but for present purposes, 10 is a good number because you can run the simulation step-by-step and see the results.

Then I simulated the first die, `D1`, with the `sample` command. The command says to take a sample of size `sim.size` from the numbers 1 through 6. Each time you sample a number, you replace it (`replace = TRUE`) so it's still there, available to be sampled again in the next simulation. If you don't quite understand the commands, try them and examine `D1`. (Just enter `D1` in the console.) Keep repeating the commands until you understand them.

Then I sampled `D2` the same way as `D1`. Next, I made a new variable `tot`. If you don't see what `tot` is, then run the commands up to this point and examine `tot`. Repeat the commands as many times as you need in order to understand them. Then I used the command `ifelse` to make a new variable `win`. `ifelse` works like this. You write `ifelse(a,b,c)`. The result of `ifelse` is `b` whenever `a` is true and `c` whenever `a` is false. I stored the result in a new variable `win`. My `a` was `tot == 7 | tot == 11`. The "|" means "or", so `win` becomes `TRUE` whenever `tot` is 7 or 11. The vertical bar here is R's symbol for "or", even though the vertical bar elsewhere is the mathematical symbol for "given." R uses the double equals sign, "==", for comparing two quantities and reserves the single equals sign "=" (not used here) for assignment. As usual, try to understand the code. If need be, run it a few times and examine the result.

The number of times we won on the come-out roll is `sum ( win )`, so the estimate of $\Pr[\text{win on come-out roll}]$ is $\frac{\texttt{sum (win)}}{\texttt{sim.size}}$. According to Example 1.1, $\Pr[\text{win on come-out roll}] = 2/9$. The simulation above estimated $\Pr[\text{win on come-out roll}]$ as 0.7. The difference is due to the small simulation size. Try running the simulation yourself with various sample sizes. See how close you come to the right answer with different sized simulations.

Example 2.4 was concerned with the number of new seedlings that emerge and survive over the winter to become established. Example 1.11 shows how a simulation can estimate the relevant probabilities.

Example 1.11 (Seedlings, continued)
In Example 2.4 we said that the marginal distribution of N, the number of new seedlings, can be described as Poi(λ) and the conditional distribution of $X \mid N$, the number that survive over the winter given that N emerge, can be described as Bin(N, p). Now we want the marginal distribution of X; i.e., $\Pr[X = 0]$, $\Pr[X = 1]$, $\Pr[X = 2]$, and so on. An R script might look something like this.

```
sim.size <- 10000
lambda <- 1 # change this as needed
p <- .5 # change this as needed
N <- rpois ( sim.size, lambda )
X <- rbinom ( sim.size, N, p )
table ( X )
X
```

```
   0    1    2    3    4    5
6023 3048  779  132   13    5
table ( X ) / sim.size
X
     0      1      2      3      4      5
0.6023 0.3048 0.0779 0.0132 0.0013 0.0005
```

First we set the simulation size and choose values for λ and p. There may be particular values of λ and p we want to use, or we may want to run the simulation many times, for many different choices of λ and p. Then we simulate N with the `rpois` command. We have to tell `rpois` how many samples to generate and what value of λ to use. Once we have samples of N, we use `rbinom` to generate samples of X. We have to tell `rbinom` how many samples to generate and what values of N and p to use for each of them. Finally, the `table` command tells us how many times $X = 0$, $X = 1$, and so on. My simulation estimated $\Pr[X = 0] \approx 60\%$, $\Pr[X = 1] \approx 30\%$, and $\Pr[X > 1] \approx 10\%$. Try it yourself. Vary the simulation size. Vary λ and p.

The marginal distribution of X can also be calculated mathematically.

$$\begin{aligned}
\Pr[X = x] &= \sum_{n=x}^{\infty} \Pr[N = n, X = x] \\
&= \sum_{n=x}^{\infty} \Pr[N = n] \times \Pr[X = x \mid N = n] \\
&= \sum_{n=x}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{n=x}^{\infty} \frac{e^{-\lambda(1-p)}(\lambda(1-p))^{n-x}}{(n-x)!} \frac{e^{-\lambda p}(\lambda p)^x}{x!} \\
&= \frac{e^{-\lambda p}(\lambda p)^x}{x!} \sum_{z=0}^{\infty} \frac{e^{-\lambda(1-p)}(\lambda(1-p))^z}{z!} \\
&= \frac{e^{-\lambda p}(\lambda p)^x}{x!}
\end{aligned}$$

The next to last equality follows from setting $z = n - p$, the number of seedlings that emerge, but die. The last equality follows since $\sum_{z=0}^{\infty} \cdots = 1$ because it is the sum of probabilities from the Poi$(\lambda(1-p))$ distribution, and the sum of all the probabilities from any distribution is 1. The final result is recognized as a probability from the Poisson distribution with rate λp. So $X \sim$ Poi(λp).

The intuition is that seedlings emerge at rate λ. A fraction p of them survive. So the rate at which surviving seedlings emerge is λp.

Example 1.9 calculated Pr[winning at craps] ≈ 0.493 by a long and tedious sequence of steps. Example 1.12 estimates the same probability by simulation.

Example 1.12 (Craps, continued)
We will simulate a large number of craps games and count how many we win. We begin by using the following code to make our own R command called `MakePoint`.

```
MakePoint <- function ( point ) {
  made <- NA  # made will eventually be TRUE or FALSE
  while ( is.na ( made ) ) {
    roll <- sum ( sample ( 1:6, 2, replace = TRUE ) )
    if ( roll == point ) made <- TRUE
    else if ( roll == 7 ) made <- FALSE
  }
  return ( made )
}
```

`MakePoint <- function ( point ) {...}` creates an R command called `MakePoint` that accepts a single number, `point`, as input, then runs all the commands between the braces. First it creates a new variable called `made`. Initially, `made` is `NA`, which stands for Not Available, but eventually `made` will become either `TRUE` or `FALSE`. Then `while (...) {...}` creates a loop. If what's inside the parentheses is `TRUE`, it runs the commands inside the braces. Then it goes back to the top of the loop and checks whether what's in the parentheses is still `TRUE`. If it is, it runs the commands in the braces again, and so on.

In our example, the loop checks whether `made` is Not Available. If it is, R simulates a roll. If the roll equals the point, `made` becomes `TRUE`. If the roll equals 7, `made` becomes `FALSE`. If neither of those happen, `made` remains `NA` and R goes back to the beginning of the loop and simulates another roll. Eventually `made` becomes either `TRUE` or `FALSE`.

You can create the `MakePoint` command and try it a few times. Enter things like `MakePoint(4)` or `MakePoint(8)` or even `MakePoint(-8)`. Try to understand what `MakePoint` does.

After we create the `MakePoint` command we come to the main part of the simulation, which is given below.

```
n.games <- 10  # number of simulated games
win <- rep ( NA, n.games )  # a vector that will eventually be TRUE for the games
                            # we win and FALSE for the games we lose
for ( i in 1:n.games ) {
  ComeOut <- sum ( sample ( 1:6, 2, replace = TRUE ) )
  if ( ComeOut == 7 || ComeOut == 11 )
    win[i] <- TRUE
  else if ( ComeOut == 2 || ComeOut == 3 || ComeOut == 12 )
    win[i] <- FALSE
  else
    win[i] <- MakePoint ( ComeOut )
}
```

n.games says how many games we're going to simulate. win is a vector of length n.games. Each element of win starts as NA but eventually becomes either TRUE or FALSE. Then we begin a "for" loop. First the loop sets i equal to 1 and determines the value of win[1]. Then it sets i equal to 2 and determines the value of win[2]. The loop keeps incrementing i until it determines the value of win[n.games]. Once all values in win have been determined, we can estimate the probability of winning a craps game by entering either sum(win)/n.games or mean(win). Try it. See how many games you must simulate in order to estimate the probability accurately.

1.10 Exercises

1. **Simulating Dice Rolls**

   (a) Simulate 6000 fair dice rolls. Count the number of 1's, 2's, ..., 6's.

   (b) You expect about 1000 of each number. How close was your result to what you expected?

   (c) About how often would you expect to get more than 1030 1's? Run an R simulation to estimate the answer.

2. **The Game of Risk** In the board game *Risk* players place their armies in different countries and try eventually to control the whole world by capturing countries one at a time from other players. To capture a country, a player must attack it from an adjacent country. If player A has $A \geq 2$ armies in country A, she may attack adjacent country D. Attacks are made with from 1 to 3 armies. Since at least 1 army must be left behind in the attacking country, A may choose to attack with a minimum of 1 and a maximum of $\min(3, A-1)$ armies. If player D has $D \geq 1$ armies in country D, he may defend himself against attack using a minimum of 1 and a maximum of $\min(2, D)$ armies. It is almost always best to attack and defend with the maximum permissible number of armies.

   When player A attacks with a armies she rolls a dice. When player D defends with d armies he rolls d dice. A's highest die is compared to D's highest. If both players use at least two dice, then A's second highest is also compared to D's second highest. For each comparison, if A's die is higher than D's then A wins and D removes one army from the board; otherwise D wins and A removes one army from the board. When there are two comparisons, a total of two armies are removed from the board.

   (a) If A attacks with one army (she has two armies in country A, so may attack with just one) and D defends with one army (he has only one army in country D) what is the probability that A will win?

   (b) Suppose that Player 1 has two armies each in countries C_1, C_2, C_3 and C_4, that Player 2 has one army each in countries B_1, B_2, B_3 and B_4, and that country C_i attacks country B_i. What is the chance that Player 1 will be successful in at least one of the four attacks?

   (c) If A attacks with 2 armies and D defends with 1, what is the probability A will win?

(d) Write a simulation to estimate the probability in the previous part. Make sure your answers agree.

3. When spun, an unbiased spinner points to some number X in the interval $[0, 1]$.

   (a) What density $p(x)$ is likely to describe X accurately? Draw a graph of it.

   (b) $Y = 2X$. On what interval does Y live? What density $p(y)$ is likely to describe Y accurately? Draw a graph of it.

4. X is a random variable in the interval $(-1, 1)$. The pdf is $p(x) = kx$ for some constant k.

   (a) Find k.

   (b) Plot the pdf.

   (c) Let $Y = 3X$. Find the pdf of Y and plot it.

   (d) Let $Z = X/2$. Find the pdf of Z and plot it.

5. A teacher randomly selects a student from a Sta 103 class. Let X be the number of math courses the student has completed. Let $Y = 1$ if the student is female and $Y = 0$ if the student is male. Fifty percent of the class is female. Among the women, thirty percent have completed one math class, forty percent have completed two math classes and thirty percent have completed three. Among the men, thirty percent have completed one math class, fifty percent have completed two math classes and twenty percent have completed three.
   True or **False**: X and Y are independent. Explain.

6. Sue is studying the Bin(25,.4) distribution. In R she types

```
y <- rbinom(50, 25, .4)
m1 <- mean(y)
m2 <- sum(y) / 25
```

   (a) Is `y` a number, a vector or a matrix?

   (b) What is the approximate value of `m1`?

   (c) What is the approximate value of `m2`?

7. Isaac is in 5th grade. Each sentence he writes for homework has a 90% chance of being grammatically correct. The correctness of one sentence does not affect the correctness of any other sentence. He recently wrote a 10 sentence paragraph for a writing assignment.

   (a) What is the probability Isaac's first sentence is grammatically correct?

   (b) Let the random variable X be the number of correct sentences in Isaac's paragraph. What is the distribution of X?

   (c) Find $\Pr[X = 10]$.

   (d) Find $\Pr[X = 9]$.

   (e) Find the probability that no more than two sentences are grammatically incorrect.

8. As part of his math homework Isaac had to roll two dice and record the results. Let X1 be the result of the first die and X2 be the result of the second. What is the probability that X1=1 given that X1 + X2 = 5?

9. Teams A and B play each other in the World Series of baseball. Team A has a 60% chance of winning each game. What is the chance that B wins the series? (The winner of the series is the first team to win 4 games.)

10. A basketball player shoots ten free throws in a game. She has a 70% chance of making each shot. If she misses the shot, her team has a 30% chance of getting the rebound.

    (a) Let M be the number of shots she makes. What is the distribution of M?

    (b) What is the chance that she makes somewhere between 5 and 9 shots, inclusive?

    (c) Let R be the number of rebounds her team gets from her free throws. What is the conditional distribution of $R \mid M$?

    (d) What is the chance that $R \geq 1$?

11. A doctor suspects a patient has the rare medical condition DS, or disstaticularia, the inability to learn statistics. DS occurs in .01% of the population, or one person in 10,000. The doctor orders a diagnostic test. The test is quite accurate. Among people who have DS the test yields a positive result 99% of the time. Among people who do not have DS the test yields a positive result only 5% of the time.

For the patient in question, the test result is positive. Calculate the probability the patient has DS.

12. Country A suspects country B of having hidden chemical weapons. Based on secret information from their intelligence agency they calculate Pr[B has weapons] = .8. But then country B agrees to inspections, so A sends inspectors. If there are no weapons then of course the inspectors won't find any. But if there are weapons then they will be well hidden, with only a 20% chance of being found. I.e.,

$$\text{Pr}[\text{finding weapons}|\text{weapons exist}] = .2. \tag{1.13}$$

No weapons are found. Find the probability that B has weapons. I.e., find

$$\text{Pr}[\text{B has weapons}|\text{no weapons are found}].$$

13. There are two coins. One is fair; the other is two-headed. You randomly choose a coin and toss it.

    (a) What is the probability the coin lands Heads?

    (b) What is the probability the coin is two-headed given that it landed Heads?

    (c) What is the probability the coin is two-headed given that it landed Tails? Give a formal proof, not intuition.

    (d) You are about to toss the coin a second time. What is the probability that the second toss lands Heads given that the first toss landed Heads?

14. There are two coins. For coin A, $\text{Pr}[H] = 1/4$; for coin B, $\text{Pr}[H] = 2/3$. You randomly choose a coin and toss it.

    (a) What is the probability the coin lands Heads?

    (b) What is the probability the coin is A given that it landed Heads? What is the probability the coin is A given that it landed Tails?

    (c) You are about to toss the coin a second time. What is the probability the second toss lands Heads given that the first toss landed Heads?

15. At Dupont College (apologies to Tom Wolfe) Math SAT scores among math majors are distributed $\text{N}(700, 50)$ while Math SAT scores among non-math majors are distributed $\text{N}(600, 50)$. 5% of the students are math majors. A randomly chosen student has a math SAT score of 720. Find the probability that the student is a math major.

16. Ecologists are studying salamanders in a forest. There are two types of forest. Type A is conducive to salamanders while type B is not. They are studying one forest but don't know which type it is. Types A and B are equally likely.

    During the study, they randomly sample quadrats. (A quadrat is a square-meter plot.) In each quadrat they count the number of salamanders. Some quadrats have poor salamander habitat. In those quadrats the number of salamanders is 0. Other quadrats have good salamander habitat. In those quadrats the number of salamanders is either 0, 1, 2, or 3, with probabilities 0.1, 0.3, 0.4, and 0.2, respectively. (Yes, there might be no salamanders in a quadrat with good habitat.) In a type A forest, the probability that a quadrat is good is 0.8 and the probability that it is poor is 0.2. In a type B forest the probability that a quadrat is good is 0.3 and the probability that it is poor is 0.7.

    (a) On average, what is the probability that a quadrat is good?

    (b) On average, what is the probability that a quadrat has 0 salamanders, 1 salamander, 2 salamanders, 3 salamanders?

    (c) The ecologists sample the first quadrat. It has 0 salamanders. What is the probability the quadrat is good?

    (d) Given that the quadrat had 0 salamanders, what is the probability the forest is type A?

    (e) Now the ecologists prepare to sample the second quadrat. Given the results from the first quadrat, what is the probability the second quadrat is good?

    (f) Given the results from the first quadrat, what is the probability they find no salamanders in the second quadrat?

17. In 1973 UC Berkeley investigated its graduate admissions rates for potential sex bias. The built-in data set `UCBAdmissions` gives the acceptance and rejection data from the six largest graduate departments on which the study was based. Typing `?UCBAdmissions` or `help(UCBAdmissions)` tells more about the data. Typing `UCBAdmissions` displays the data. It is in the form of a 2 by 2 by 6 table in which the first dimension is admission status, either `Admitted` or `Rejected`; the second dimension is Gender, either `Male` or `Female`; and the third dimension is department, either `A`, `B`, `C`, `D`, `E`, or `F`. The entries in the table give the number of people in each cell. Let A, G, and D be the admission status, gender, and department of a randomly selected student.

(a) Find Pr[A = Admitted] and Pr[A = Rejected].

(b) Find Pr[G = Female] and Pr[G = Male].

(c) Find Pr[D = A],Pr[D = B], ..., and Pr[D = F].

(d) Find Pr[A = Admitted | G = Female] and Pr[A = Admitted | G = Male]. Are A and G dependent or independent? Who is more likely to be admitted: a randomly selected female or randomly selected male applicant?

(e) Find Pr[A = Admitted | G = Female, D = A]. That is, find the probability of admission for a randomly selected female who applied to department A. Also find Pr[A = Admitted | G = Male, D = A]. Within department A, Are A and G dependent or independent? Who is more likely to be admitted: a randomly selected female or randomly selected male applicant from department A?

(f) Repeat part (e) but for departments B, C, D, E, and F.

(g) Do your answers to parts (e) and (f) agree with your answers to part (d)? Explain.

18. The Great Randi is a professed psychic and claims to know the outcome of coin flips. This problem concerns a sequence of 20 coin flips that Randi will try to guess (or not guess, if her claim is correct).

(a) Suppose

$$\text{Pr[Randi is psychic]} = .01$$

expresses your skepticism of Randi's claim.

i. Before any guesses have been observed, find
Pr[first guess is correct] and Pr[first guess is incorrect].

ii. After observing 10 consecutive correct guesses, find the updated
Pr[Randi is psychic].

iii. After observing 10 consecutive correct guesses, find
Pr[next guess is correct] and Pr[next guess is incorrect].

iv. After observing 20 consecutive correct guesses, find
Pr[next guess is correct] and Pr[next guess is incorrect].

(b) Suppose that Randi doesn't claim to guess coin tosses perfectly, only that she can guess them at better than 50%. 100 trials are conducted. Randi gets 60 correct. How strongly do the data support Randi's claim? What if 70 were correct? How strongly would that support her claim?

(c) The Great Sandi, a statistician, writes the following R code to calculate a probability for Randi.

```
y <- rbinom ( 500, 100, .5)
sum ( y == 60 ) / 500
```

What is Sandi trying to calculate? Write a formula (Don't evaluate it.) for the quantity Sandi is trying to calculate.

19. For various reasons, researchers often want to know the number of people who have participated in embarassing activities such as illegal drug use, cheating on tests, robbing banks, etc. An opinion poll which asks these questions directly is likely to elicit many untruthful answers. To get around the problem, researchers have devised the method of randomized response. The following scenario illustrates the method.

A pollster identifies a respondent and gives the following instructions. "Toss a coin, but don't show it to me. If it lands Heads, answer question (a). If it lands tails, answer question (b). Just answer 'yes' or 'no'. Do not tell me which question you are answering.

Question (a): Does your telephone number end in an even digit?

Question (b): Have you ever used cocaine?"

Because the respondent can answer truthfully without revealing his or her cocaine use, the incentive to lie is removed. Researchers hope respondents will tell the truth.

You may assume respondents are truthful and that telephone numbers are equally likely to be odd or even. Let p be the probability that a randomly selected person has used cocaine.

(a) What is the probability that a randomly selected person answers "yes"?

(b) Suppose we survey 100 people. Let X be the number who answer "yes". What is the distribution of X?

(c) What is the probability a respondent uses cocaine given she says "yes"?

(d) What is the probability a respondent uses cocaine given she says "no"?

20. In a 1991 article (See Utts, 1991 and discussants.) Jessica Utts reviews some of the history of probability and statistics in ESP research. This question concerns

a particular series of *autoganzfeld* experiments in which a sender looking at a picture tries to convey that picture telepathically to a receiver. Utts explains:

> "...'autoganzfeld' experiments require four participants. The first is the Receiver (R), who attempts to identify the target material being observed by the Sender (S). The Experimenter (E) prepares R for the task, elicits the response from R and supervises R's judging of the response against the four potential targets. (Judging is double blind; E does not know which is the correct target.) The fourth participant is the lab assistant (LA) whose only task is to instruct the computer to randomly select the target. No one involved in the experiment knows the identity of the target.
>
> "Both R and S are sequestered in sound-isolated, electrically shielded rooms. R is prepared as in earlier ganzfeld studies, with white noise and a field of red light. In a nonadjacent room, S watches the target material on a television and can hear R's target description ('mentation') as it is being given. The mentation is also tape recorded.
>
> "The judging process takes place immediately after the 30-minute sending period. On a TV monitor in the isolated room, R views the four choices from the target pack that contains the actual target. R is asked to rate each one according to how closely it matches the ganzfeld mentation. The ratings are converted to ranks and, if the correct target is ranked first, a direct hit is scored. The entire process is automatically recorded by the computer. The computer then displays the correct choice to R as feedback."

In the series of autoganzfeld experiments analyzed by Utts, there were a total of 355 trials. Let X be the number of direct hits.

(a) What are the possible values of X?

(b) Assuming there is no ESP, and no cheating, what is the distribution of X?

(c) Plot the distribution in part (b).

21. Suppose extensive testing has revealed that people in Group A have IQ's that are well described by a $\text{N}(100, 10)$ distribution while the IQ's of people in Group B have a $\text{N}(105, 10)$ distribution. Write some R code to answer the question *What is the probability that a randomly chosen individual from Group A has a higher IQ than a randomly chosen individual from Group B?*

22. The so-called *Monty Hall* or *Let's Make a Deal* problem has caused much consternation over the years. It is named for an old television program. A contestant is presented with three doors. Behind one door is a fabulous prize; behind the other two doors are virtually worthless prizes. The contestant chooses a door. The host of the show, Monty Hall, then opens one of the remaining two doors, revealing one of the worthless prizes. Because Monty is the host, he knows which doors conceal the worthless prizes and always reveals one of them, but never the door chosen by the contestant. Then the contestant is offered the choice of keeping what is behind her original door or trading for what is behind the remaining unopened door. What should she do?

    There are two popular answers.

    - There are two unopened doors, they are equally likely to conceal the fabulous prize, so it doesn't matter which one she chooses.
    - She had a $1/3$ probability of choosing the right door initially, a $2/3$ chance of getting the prize if she trades, so she should trade.

    (a) Create a simulation in R to discover which answer is correct.

    (b) Show using formal arguments of conditional probability which answer is correct.

    Make sure your answers to (a) and (b) agree!

23. Suppose you want to test whether the random number generator in R generates each of the digits $0, 1, \ldots, 9$ with probability 0.1. How could you do it? You may consider first testing whether R generates 0 with the right frequency, then repeating the analysis for each digit.

Chapter 2

Inference

This chapter takes up the heart of statistics: making inferences from data. In our opinion, the most important part of inference is looking at data and displaying it for others to see. We agree with Lindley (1993), who says, "In the earlier stages ...consideration of models can be done informally by inspection of the data. Only when the issues become more delicate is there a need for a formal system ...". Often, the early stages are all that's needed, so we begin our discussion of statistical inference by illustrating several ways of informally inspecting and displaying data, usually by plots. There is a rich literature on visualizing data and we won't cover it all here. After visualization, we will present more formal ways of analyzing data.

2.1 Data Visualization

Example 2.1 (Hot Dogs)
In June of 1986, Consumer Reports published a study of hot dogs. The data are available at DASL, THE *Data and Story Library*, a collection of data sets for free use by statistics students. DASL says the hot dog data are

> "Results of a laboratory analysis of calories and sodium content of major hot dog brands. Researchers for Consumer Reports analyzed three types of hot dog: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat)."

You can download the hot dog data from DASL. The first few lines look like this:

```
> head ( hotdogs )
  Type Calories Sodium
```

```
1 Beef       186    495
2 Beef       181    477
3 Beef       176    425
4 Beef       149    322
5 Beef       184    482
6 Beef       190    587
>
```

The data come in a `.txt` file, for which you should use the `read.table` command like this:

```
  hotdogs <- read.table ( file name here,  header = TRUE )
```

`header = TRUE` tells R the first line of the file contains the names of the columns.

This example shows several ways to display the calorie content of the beef hot dogs.

Figure 2.1A is a histogram. From the histogram one might form the impression there are two major varieties of beef hot dogs, one with about 130–160 calories or so, another with about 180 calories or so, and a rare outlier with fewer calories. Figure 2.1B is another histogram of the same data but with a different bin width. It gives a different impression, namely that calorie content is approximately evenly distributed from about 130 to about 190 with a small number of lower calorie hot dogs. Figure 2.1C gives much the same impression as 2.1B. It was made with the same bin width as 2.1A, but with cut points starting at 105 instead of 110. These histograms illustrate that one's impression can be influenced by both bin width and cut points. Because one's impression can be so easily influenced by what ought to be irrelevant details, we should consider other types of display.

Figures 2.1D, 2.1E, and 2.1F are density estimates. Density estimates have a control parameter, usually called *bandwidth*, similar to histograms' bin width. A larger bandwidth makes the density look smoother — it's like using larger bin width in a histogram. The three bandwidths in Figure 2.1 give different impressions. 2.1D looks like a roughly constant density in the range 130–190 calories, with a small tail below 130; 2.1E gives the impression of two main populations of hot dogs, with a third, small population around 110 calories; 2.1F could be showing there are three or even four separate populations of hot dogs. Which of these bandwidths to choose is not just a statistics question and there is not just one right answer. The decision is not only the statistician's, but may require help from the scientist with whom the statistician is working. You and your collaborator may choose a bandwidth that depends on what else you know about hot dogs and what aspect of the data you want to emphasize.

Good density estimation is not as simple as we made it seem in Example 1.4 where we estimated probability density by dividing parents' heights into bins. More sophisticated density estimates don't use sharp cutoffs like bin boundaries but are designed to produce smoothly varying density estimates.

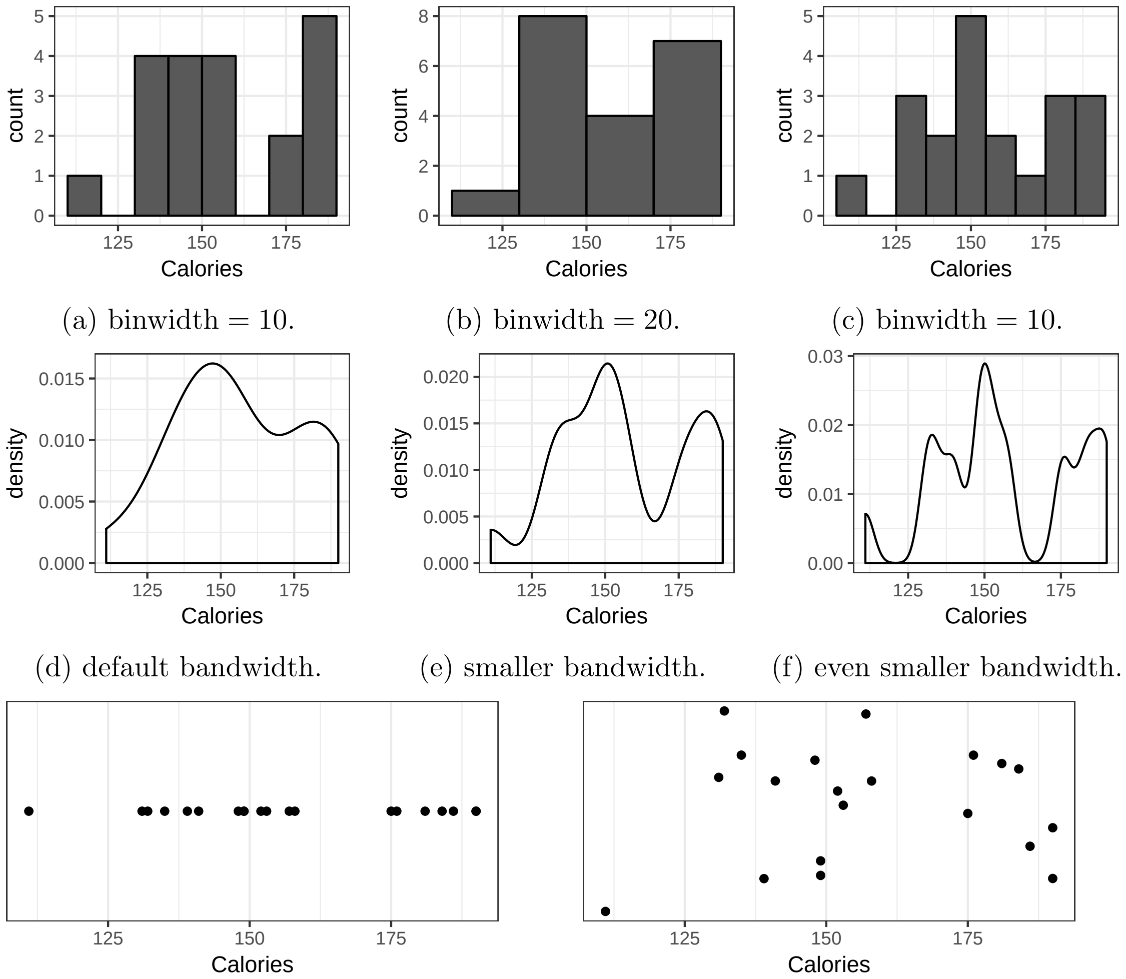
Figures 2.1G and 2.1H are strip charts. The x-axis is calorie count, the y-axis is meaningless. 2.1H has random jitter added in the y direction because that sometimes helps the reader see the data better, especially when the points would otherwise overlap. Both strip charts reveal a gap between about 115 and 130 calories and a second gap between about 160 and 175 calories. Whether that gap represents something interesting about hot dogs or is due just to the vagaries of random sampling, we don't know. Part of the statistician's job is to discover the gap. Then we can report it to our collaborator, who can help decide whether it's interesting enough to explore further.

Figure 2.1 was made with the following R commands.

```
p <- ggplot ( subset ( hotdogs, Type == "Beef" ), aes ( x = Calories ) )
p + geom_histogram ( binwidth = 10, center = 115, color = "black" ) +
    scale_y_continuous ( minor_breaks = NULL ) +
    theme_bw()
p + geom_histogram ( binwidth = 20, center = 120, color = "black" ) +
    scale_y_continuous ( minor_breaks = NULL ) +
    theme_bw()
p + geom_histogram ( binwidth = 10, center = 120, color = "black" ) +
    scale_y_continuous ( minor_breaks = NULL ) +
    theme_bw()
p + geom_density () +
    theme_bw()
p + geom_density ( adjust = .5 ) +     # half the default bandwidth
    theme_bw()
p + geom_density ( adjust = .25 ) +   # quarter the default bandwidth
    theme_bw()
p + geom_point ( mapping = aes ( y = 0 ) ) +
    scale_y_continuous ( name = NULL, breaks = NULL ) +
    theme_bw()
p + geom_jitter ( mapping = aes ( y = 0 ), width = 0 ) +  # no horizontal jitter
    scale_y_continuous ( name = NULL, breaks = NULL ) +
    theme_bw()
```

These commands show some of the thought behind `ggplot`. A single plot is created with the `p <- ggplot (...)` command, then displayed multiple ways with the many `p + ...` commands.

Example 2.2 (Hot Dogs, continued)
Sometimes it is necessary to compare the distributions of several populations. Example 2.2 illustrates by comparing the calorie counts of the three types of hot dogs in Example 2.1. Figures 2.2A and 2.2B both use strip charts. In 2.2A, the hot dog type is on the y-axis. In 2.2B, each type has its own facet. The populations of Beef and Meat



(a) binwidth = 10.

(b) binwidth = 20.

(c) binwidth = 10.

(d) default bandwidth.

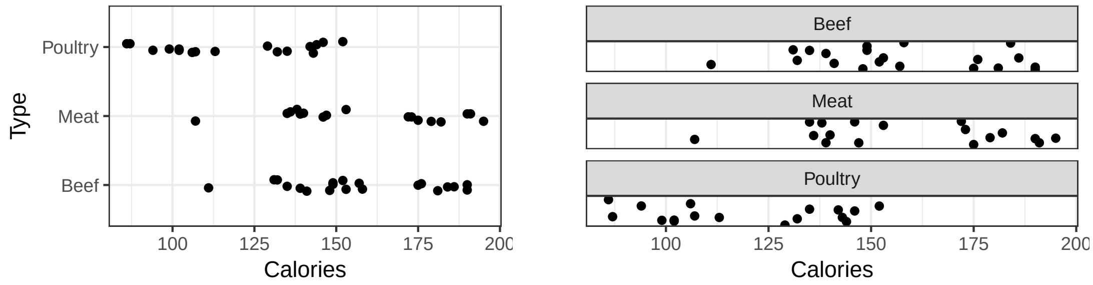
(e) smaller bandwidth.

(f) even smaller bandwidth.

(g) strip chart

(h) strip chart with vertical jitter

Figure 2.1: Multiple ways to display the calorie content of beef hotdogs



(a) Type on the y-axis

(b) Each type in its own facet

Figure 2.2: Calorie content of Beef, Meat, and Poultry hot dogs

hot dogs appear to be about the same, but the Poultry hot dogs seem to have fewer calories. No formal or quantitative analysis is needed.

Whichever display you think is more effective in this example, the answer may change for other data sets. The point is not to prescribe how to display data, but to show some of the possibilities. When you analyze data, use your judgement about which display is best for your problem and your audience.

Figure 2.2 was produced with the following commands

```
p <- ggplot ( hotdogs, aes ( x = Calories ) )
p + geom_jitter ( mapping = aes ( y = Type ), width = 0, height = .1 ) +
    theme_bw()
p + geom_jitter ( mapping = aes ( y = 0 ), width = 0, height = .1 ) +
    facet_wrap ( ~ Type, ncol = 1 ) +
    scale_y_continuous ( name = NULL, breaks = NULL ) +
    theme_bw()
```

It sometimes happens that statisticians examine the relationship between two random variables X and Y, but the relationship depends on a third random variable Z. A plot of Y vs. X doesn't suffice to reveal the dependence on Z. Example 2.3 illustrates the idea with some oceanographic data. X is latitude, Y is ocean temperature, and Z is longitude. In this example, the relationship between X and Y is crucial because it tells us about deep ocean currents. We find that the relationship between X and Y varies slightly with Z. We didn't know in advance that would be the case; we discovered it by plotting data.

Example 2.3 (Ocean temperatures)
Physical oceanographers study physical properties such as temperature, salinity, pres-

sure, oxygen concentration, and potential vorticity of the world's oceans. Data about the oceans' surface can be collected by satellites' bouncing signals off the surface, but satellites cannot collect data about deep ocean water. Until the 1970s, the main source of data about deep water came from ships that lowered instruments to various depths to record properties of ocean water. Starting around the 1970s oceanographers began to use neutrally buoyant floats. A brief description and history of the floats can be found on the web at `www.soc.soton.ac.uk/JRD/HYDRO/shb/float.history.html`. Figure 2.3 shows locations, called *hydrographic stations*, off the coast of Africa and Europe where ship-based measurements were taken between about 1902 and 1998 at a depth of 1000 meters. The outline of the continents is apparent on the right-hand side of the figure due to the lack of measurements over land.

Deep ocean currents cannot be seen but can be inferred from physical properties, as illustrated by Figure 2.4, which plots temperature vs. latitude for a range of longitudes, using the same hydrographic stations as Figure 2.3. In each facet, temperature peaks around 35°N latitude. Data like these allow oceanographers to deduce the presence of a large outpouring of relatively warm water called the *Mediterranean tongue* from the Mediterranean Sea into the Atlantic ocean. The Mediterranean tongue is centered at about 1000 meters depth and 35°N latitude, flows from east to west, and is warmer than the surrounding Atlantic waters into which it flows. The relationship between latitude and temperature is most strongly peaked for longitudes around -10 to -15, i.e., near the coast, and gradually becomes more mildly peaked as longitude moves toward -40, in the middle of the ocean. The gradual change in the peak's sharpness tells us something about how far into the Atlantic the Mediterranean Tongue persists. No formal inference is needed.

Part of a statistician's job is to look at data in many ways, to see whether there are interesting relationships in the data and whether those relationships depend on other factors. Sometimes the other factors are substantive, like longitude in Example 2.3. Other times they are extraneous, like band width, bin width, and bin boundaries in Example 2.1. Below we present two more examples in which extraneous considerations are important. In one case, it's the shape of the plotting region; in the other it's the scale in which the data are plotted. As usual, there's no rule that tells us how to plot the data. Use intelligence guided by experience, and try different plots, even if you don't know in advance whether they will help. Use your judgement to find a plot that will reveal interesting features of the data to readers who haven't spent as much time with the data as you have.

One data set, from Weisberg (2013), gives "average soil temperature in degrees C at 20 cm depth in Mitchell, Nebraska, for 17 years beginning January 1976."

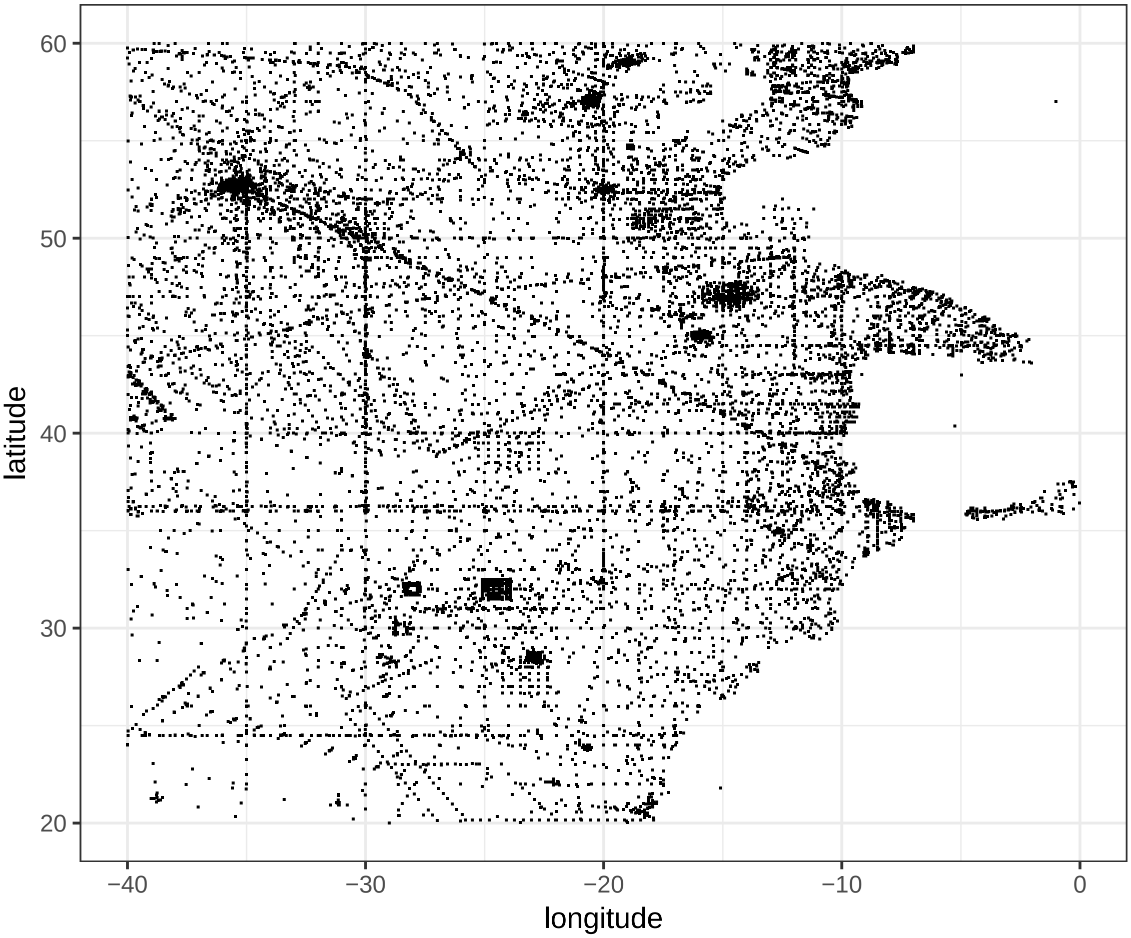


Figure 2.3: Hydrographic stations near the coast of Africa and Europe, 1902–1998

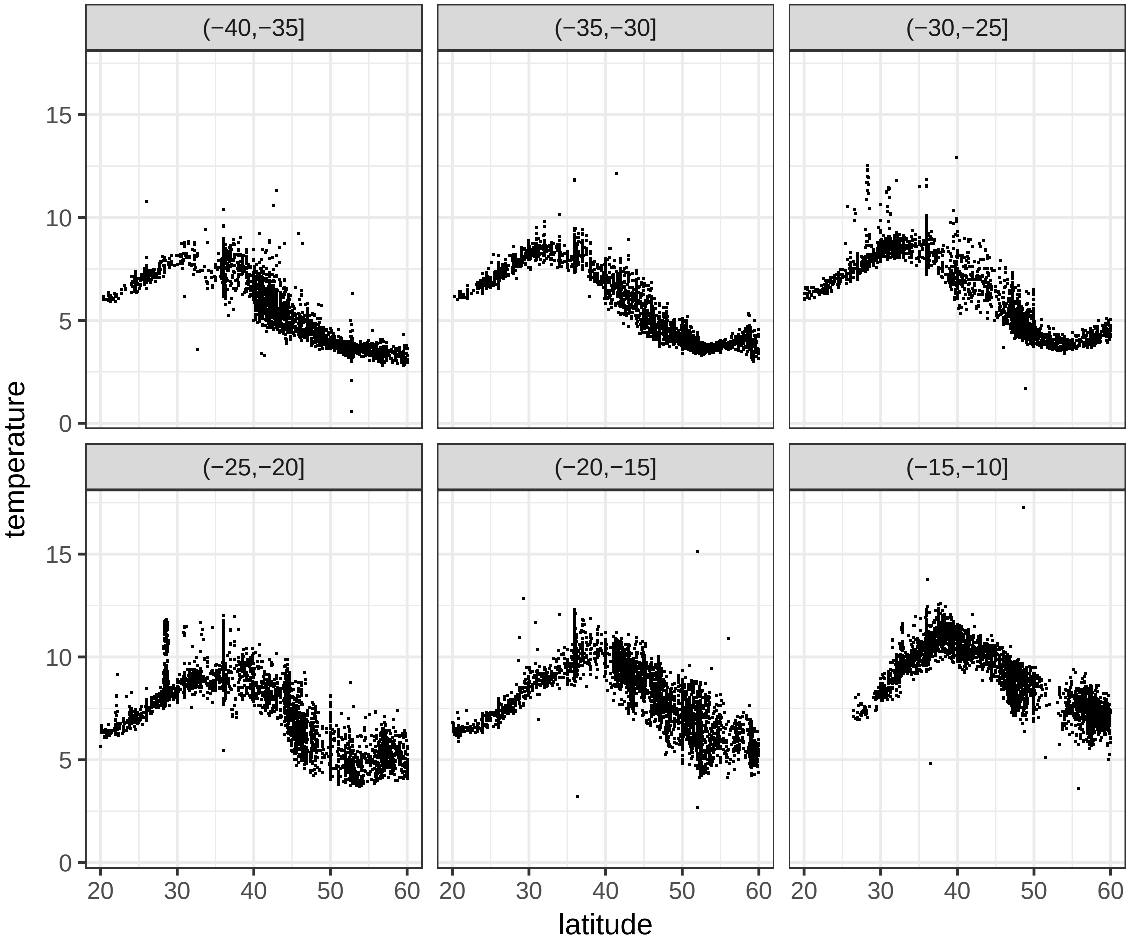


Figure 2.4: Temperature vs. latitude with facets for longitude

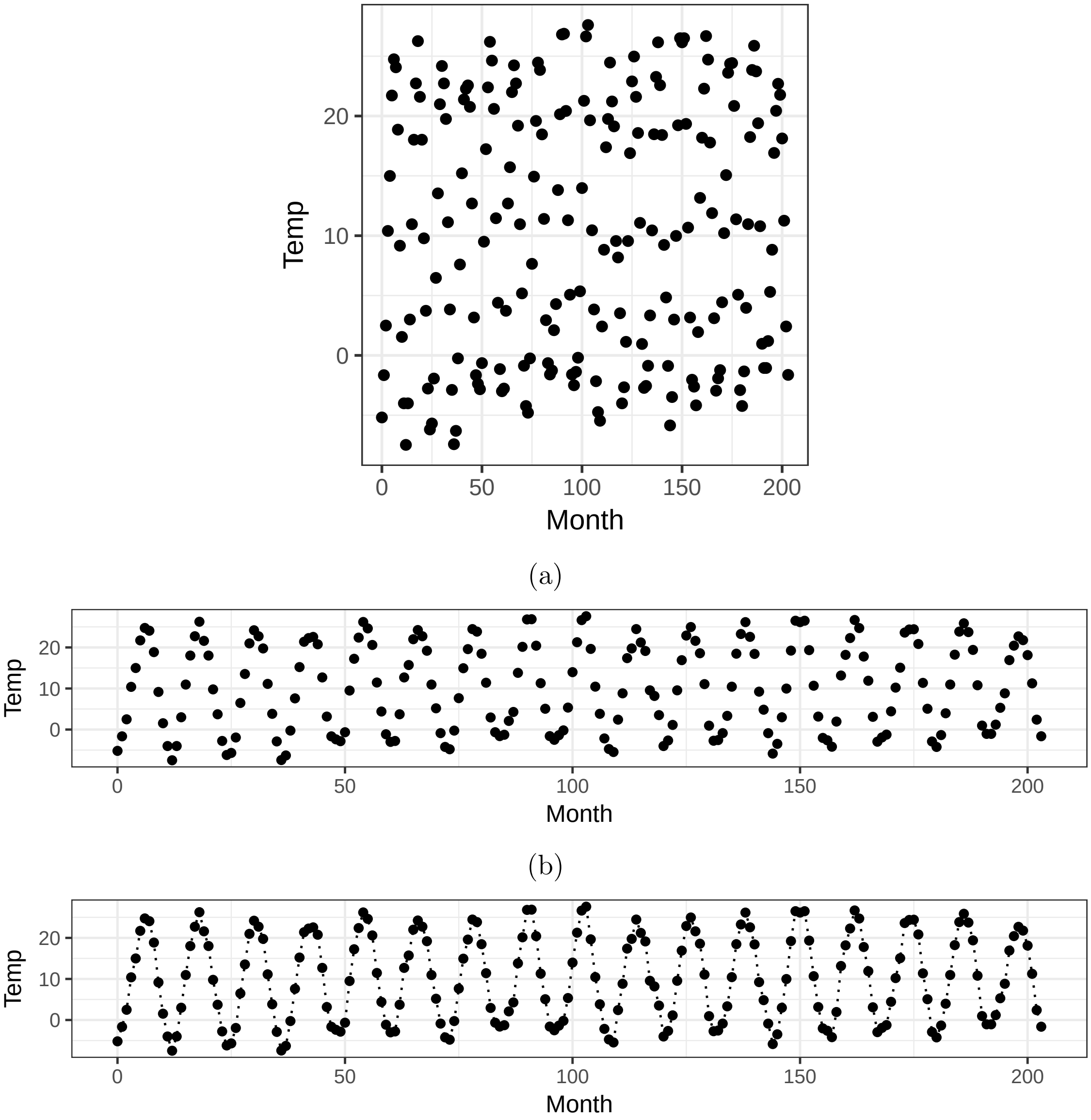
Figure 2.5 shows three ways of plotting the data. In 2.5A the relationship between Temperature and Month is obscure, at best. 2.5B plots the same data, but changes the aspect ratio. The annual cycle begins to emerge. Once we've seen 2.5B, it might occur to us to add a line, as in 2.5C, which makes the annual cycle even clearer. We didn't follow rules to make these plots. Experience told us there might be a relationship between Temperature and Time. When we didn't see it in 2.5A, we played with the plot until we saw something interesting in 2.5B. Then experience guided us again, to say that adding a line might make a clearer plot.

The next data set comes from Allison and Cicchetti (1976) via Weisberg (2013), and is the body weight in kilograms and brain weight in grams of 62 species of mammals. 2.6A is a simple plot of brain weight vs. body weight. Most points are clustered near the origin and no clear pattern is apparent. Experience tells us that when most points are clustered near the origin, and there are a few distant points, that plotting on a logarithmic scale sometimes clarifies the picture. Accordingly, 2.6B is the same data, but with both x and y on logarithmic scales. An interesting relationship is apparent: log(brain size) appears to scale roughly linearly with log(body size), over a wide variety of mammalian species.

The brain and body weight data also comes with the name of each species. Can you figure out how to get R to tell you the names of the two large species? The species with moderate body weight but brain weight over 1000 g? Which point is *homo sapiens*?

2.2 Quantitative Inference

There are several ways to make quantitative statistical inference, but the main idea is to assess statistical models according to how well they describe data. We saw one instance of quantitative inference in Example 1.2, the Slater School. There were 145 employees, some of whom developed invasive cancer. We adopted the model that all employees have the same chance of developing invasive cancer and that employees develop invasive cancer independently of each other. Those assumptions lead naturally to the model $X \sim \text{Bin}(145, p)$ where X is the number who develop invasive cancer and p is an important but unknown number. According to a national database, $p \approx 0.03$, for people whose characteristics are typical of Slater employees. The data turned out to be $X = 8$, and $8/145 \approx 0.055$ is the value of p that best describes the data. We want to compare $p = 0.03$ and $p = 0.055$. If $p = 0.055$ describes the data much better than $p = 0.03$, then there is a strong suggestion that something at Slater School, possibly its proximity to high-tension power lines, causes

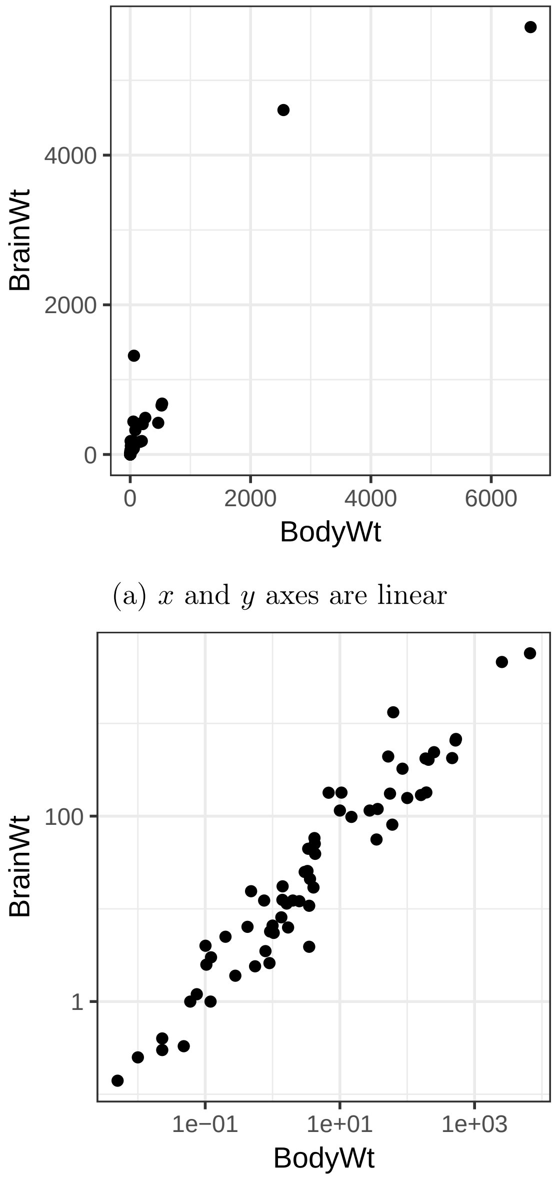


(a)

(b)

(c)

Figure 2.5: Soil temperature at Mitchell, NE



(a) x and y axes are linear

(b) x and y axes are logarithmic

Figure 2.6: Brain wt (g) vs. body wt (kg) for 62 species of mammals

cancer. But if $p = 0.055$ describes the data only slightly better than $p = 0.03$, then there is only a slight suggestion that something is harmful at Slater School.

We saw in Example 1.2 $\frac{\Pr_{p=.055}[X=8]}{\Pr_{p=.03}[X=8]} \approx 3.6$. In other words, $p = 0.055$ describes the data about 3.6 times better than $p = 0.03$. Is that a lot better or just a little better? To answer that question, it helps to think about a comparable situation where we have some intuition. Think about tossing a coin twice and let Y be the number of heads. A good model is $Y \sim \text{Bin}(2, p)$. If the coin is fair, then $p = .5$. If the coin is two-headed, then $p = 1$. Suppose it turns out that $Y = 2$. Then to compare the two-headed theory to the fair theory we would use the ratio $\frac{\Pr_{p=1}[Y=2]}{\Pr_{p=.5}[Y=2]} = \frac{1}{.25} = 4$. In other words, the two-headed theory describes the data 4 times better than the fair-coin theory. But 2 heads in 2 tosses is only weak evidence in favor of the two-headed theory. The ratio in Example 1.2 is less than 4, so the $p = .055$ model describes the data only a very little bit better than the $p = .03$ model. There is very little evidence that the rate of invasive cancer at Slater School differs from the national average rate.

If the rate of invasive cancer at Slater School is not 0.03, it also may not be exactly $8/145 \approx 0.055$. Figure 2.7 shows how well other values of p describe the data $X = 8$. So far we've compared $p = 0.055$ to $p = 0.03$ by forming the ratio

$$\frac{\text{height of curve at } p = .055}{\text{height of curve at } p = .03} \approx 3.6,$$

but we could compare any two values p_1 and p_2 by forming the ratio

$$\frac{\text{height of curve at } p_1}{\text{height of curve at } p_2},$$

which tells us how much better $X = 8$ is described by p_1 than by p_2.

If the data had turned out differently, say $X = 9$, then the y-axis of the curve would be $\Pr_p[X = 9]$. More generally, if the data turned out to be $X = x$ for some specific value x, the y-axis would be $\Pr_p[X = x]$. Only the x that actually occurs is relevant, and the curve tells us how well each value of p describes $X = x$.

$X \sim \text{Bin}(145, p)$ may not be a perfect model for cancer at the Slater School. One unrealistic assumption is that all employees have the same chance p of getting cancer. In fact, they probably have different chances, which may depend on their age, how long they've worked at Slater, their genetics, their pre-Slater background, and other factors. A more thorough analysis would make a more elaborate model to account for such factors. So, to be more precise, instead of saying, "there is very little evidence . . . ," we should say that $X = 8$ provides very little evidence . . . when analyzed with the $\text{Bin}(145, p)$ model.

Figure 2.7 illustrates two big ideas in quantitative inference.

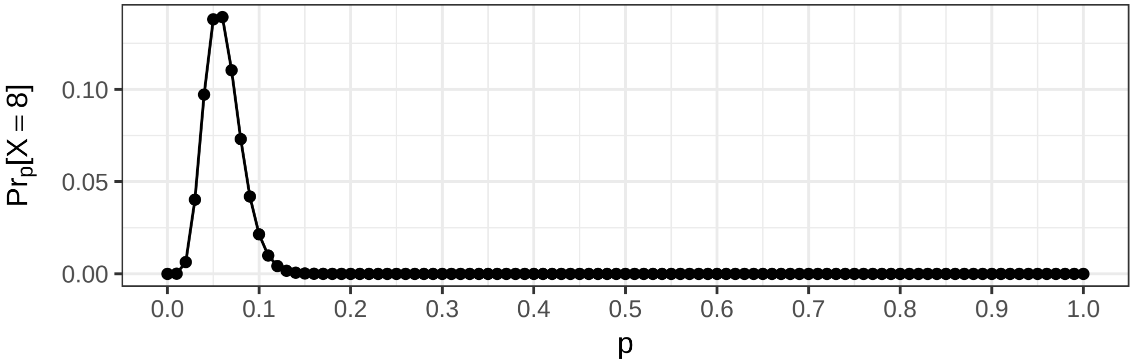


Figure 2.7: $\Pr_p[X = 8]$ as a function of p

1. We modelled the random variable X with a family of probability distributions. For the Slater school we used the family of Binomial distributions. There is one Binomial distribution for each value of p. p is called a *parameter*. In other problems we might use the Poisson family with parameter λ, or the exponential family with parameter λ, or the Normal family, which has a two-dimensional parameter (μ, σ). Each statistics problem has its own parametric family. Part of statistical analysis is to specify a parametric family that seems well suited to the situation at hand and is likely to describe the data well.

2. The parameter is unknown, and another part of statistical analysis is to compare how well different values of the parameter describe the data. Figure 2.7 shows that values of p from about 0.02 to about 0.12 describe the data not too much worse than the best, $p \approx 0.055$. In numbers, $\frac{\Pr_{p=.055}[X=8]}{\Pr_{p=.02}[X=8]} \approx 22$ and $\frac{\Pr_{p=.055}[X=8]}{\Pr_{p=.12}[X=8]} \approx 34$. Figure 1.3 compared how well different values of λ described $X = 3$ for the number of seedlings to emerge in one quadrat, and showed "any value of λ from about 0.5 to about 9 describes the data not too much worse than $\lambda = 3$." In numbers, $\frac{\Pr_{\lambda=3}[X=3]}{\Pr_{\lambda=.5}[X=3]} \approx 18$ and $\frac{\Pr_{\lambda=3}[X=3]}{\Pr_{\lambda=9}[X=3]} \approx 15$.

   It is common in statistics to use the Greek letter θ (theta) for the parameter. θ could stand for p in the Binomial family, λ in the Poisson or exponential families, (μ, σ) in the Normal family, or other parameters in other parametric families. To compare different values of θ we compare how well they describe the data $X = x$, where X is a random variable and x is the value that actually occurred. In symbols, we look at $\Pr_\theta[X = x]$ as a function of θ. That function of θ is called the *likelihood* function. Figures 1.3 and 2.7 are plots of likelihood functions.

Whichever parametric family we use, and whatever the data turns out to be, there is some value of θ at which the likelihood function achieves its maximum. That value is usually called the *maximum likelihood estimate*, or m.l.e., and denoted $\hat{\theta}$. In symbols,

$$\hat{\theta} \equiv \text{argmax}_\theta \, p_\theta(X = x). \tag{2.1}$$

The m.l.e. is the value of θ that says which member of the parametric family best describes the data, at least if "best describes" means having the largest value of $p_\theta(X = x)$.

It's easy to find the m.l.e. for the parametric models we've seen so far.

Binomial Distribution mle.Binomial If $X \sim \text{Bin}(n, p)$, where n is known and p is the unknown parameter, then $\hat{p} = x/n$.

Poisson Distribution mle.Poisson If $X_1, \dots, X_n$ are independent and all distributed $\text{Poi}(\lambda)$, then $\hat{\lambda} = \frac{x_1 + \cdots + x_n}{n}$.

Exponential Distribution mle.Exponential If $X_1, \dots, X_n$ are independent and all distributed $\text{Exp}(\lambda)$, then $\hat{\lambda} = \frac{n}{x_1 + \cdots + x_n}$.

Normal Distribution If $X_1, \dots, X_n$ are independent and all distributed $\text{N}(\mu, \sigma)$, then $\hat{\mu} = \frac{x_1 + \cdots + x_n}{n}$ and $\hat{\sigma} = \sqrt{\frac{(x_1 - \hat{\mu})^2 + \cdots + (x_n - \hat{\mu})^2}{n}}$.

It seems obvious that more data ought to be better than less. Example 2.4 shows how that works quantitatively.

Example 2.4 (Seedlings, continued)

Example 1.3 reported on a single quadrat in a single year in an observational study to learn about the rate at which new seedlings emerge in the forest. If N is the number of seedlings to emerge, we adopted the model $N \sim \text{Poi}(\lambda)$, where λ is the typical emergence rate. The point of the study is to learn what values of λ will describe the data well, which in turn will tell us something about how quickly the forest can extend its range. The full study collected data from 60 quadrats over multiple years.

You can download the data from the book's website and read it with

```
seedlings <- read.table ( paste ( datadir, "seedlings.txt", sep="/" ),
                          header = TRUE,
                          na.strings = "."
                        )
```

The column named `Block` identifies the quadrat, from 1 to 60. Columns named something like `Xyy.t` are the number of seedlings that emerged in year 19yy. Columns named something like `Xyy.1` are the number of those seedlings that survived over the winter.

Example 2.4 reports on the data from 1993. Let N_1, N_2, …, N_{60} be the numbers of seedlings to emerge in the 60 quadrats. If we think the emergence rate is similar in all quadrats then we might adopt the model $N_i \sim$ Poi(λ). We write λ and not λ_i, i.e., we use the same λ for all quadrats, to express the idea that all quadrats' emergence rates are sufficiently similar that we can reasonably model them as being the same. A good analysis should check whether that idea agrees with the data, but we will proceed for now without checking.

The data will turn out to be $N_1 = n_1$, $N_2 = n_2$, …, $N_{60} = n_{60}$, for some specific values $n_1, n_2, \ldots, n_{60}$. The joint probability will be

$$\Pr_\lambda[N_1 = n_1, N_2 = n_2, \ldots, N_{60} = n_{60}] \tag{2.2}$$

When n_1, …, n_{60} are observed we plug them into (2.2) to get the likelihood function for λ. Our model combined with a little math shows

$$\begin{aligned}
&\Pr_\lambda[N_1 = n_1, N_2 = n_2, \ldots, N_{60} = n_{60}] \\
&\quad = \Pr_\lambda[N_1 = n_1] \times \Pr_\lambda[N_2 = n_2 \,|\, N_1 = n_1] \times \Pr_\lambda[N_3 = n_3 \,|\, N_1 = n_1, N_2 = n_2] \\
&\qquad\quad \times \cdots \times \Pr_\lambda[N_{60} = n_{60} \,|\, N_1 = n_1, N_2 = n_2, \ldots, N_{59} = n_{59}] \qquad \text{Why?} \\
&\qquad\quad = \Pr_\lambda[N_1 = n_1] \times \Pr_\lambda[N_2 = n_2] \times \cdots \times \Pr_\lambda[N_{60} = n_{60}] \qquad \text{Why?} \\
&\qquad\quad = \left(\frac{e^{-\lambda}\lambda^{n_1}}{n_1!}\right) \times \left(\frac{e^{-\lambda}\lambda^{n_2}}{n_2!}\right) \times \cdots \times \left(\frac{e^{-\lambda}\lambda^{n_{60}}}{n_{60}!}\right) \qquad \text{Why?} \\
&\qquad\quad = \left(\prod_{i=1}^{60} \frac{1}{n_i!}\right) e^{-60\lambda} \lambda^{\sum_{i=1}^{60} n_i} \qquad \text{Why?}
\end{aligned}$$

```
> seedlings$X93.t
 [1] 1 1 0 0 2 1 4 1 2 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 1 6 1 4 1 4 3 0 2 6 2 3 1 0 2
[40] 0 4 3 0 0 1 1 3 1 2 3 5 1 2 6 2 2 1 1 0 1
>
```

It turned out there was a total of 90 new seedlings in 1993, i.e., $\sum_{i=1}^{60} n_i = 90$, so the likelihood function is $\Pr_\lambda[N_1 = n_1, N_2 = n_2, \ldots, N_{60} = n_{60}] = \left(\prod_{i=1}^{60} \frac{1}{n_i!}\right) e^{-60\lambda}\lambda^{90}$.

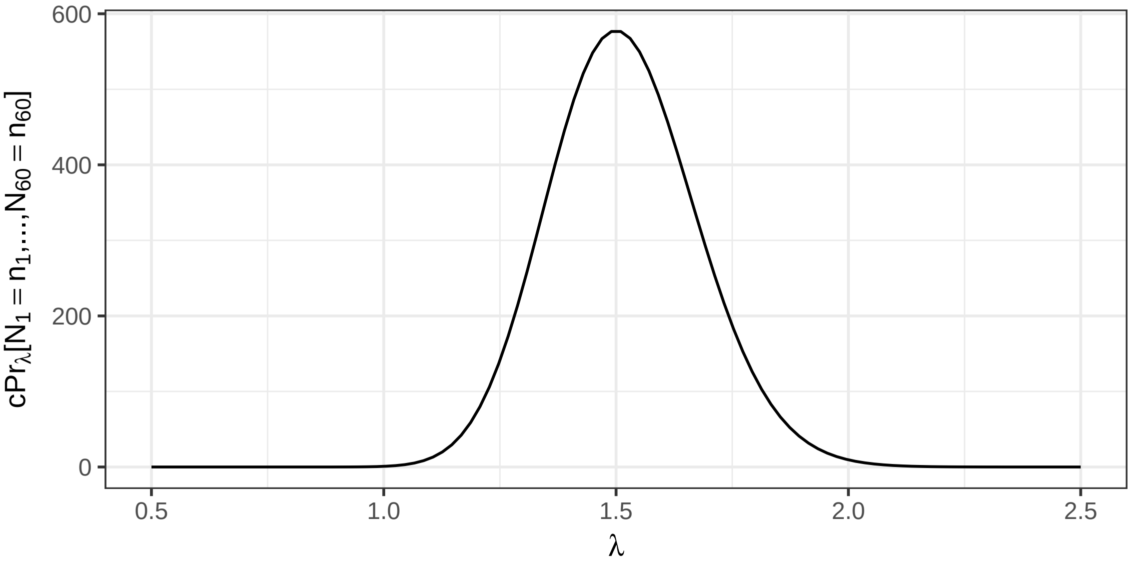


Figure 2.8: The likelihood function $\Pr_\lambda[N_1 = n_1, \ldots, N_{60} = n_{60}]$

To compare two values of λ, say λ_1 and λ_2, we use the ratio

$$\frac{\Pr_{\lambda_1}[N_1 = n_1, N_2 = n_2, \ldots, N_{60} = n_{60}]}{\Pr_{\lambda_2}[N_1 = n_1, N_2 = n_2, \ldots, N_{60} = n_{60}]} = \frac{e^{-60\lambda_1}\lambda_1^{90}}{e^{-60\lambda_2}\lambda_2^{90}}$$

and the term with the factorials drops out, so when we plot the likelihood function it doesn't matter whether we include the term with the factorials. Figure 2.8 plots the likelihood function $\Pr_\lambda[N_1 = n_1, \ldots, N_{60} = n_{60}]$ as a function of λ and shows that values of $\lambda \approx 1.5$ describe the data best — which agrees with the formula for the m.l.e. on page 79 — but any value of λ between about 1 and 2 describe the data not much worse than the best.

Contrast Figure 1.3, which showed the likelihood function for one quadrat's data with Figure 2.8, which shows the likelihood function for 60 quadrats' data. The x-axis in Figure 2.8 has been reduced because the likelihood function has become much sharper. That's the typical effect of having more data: it reduces the range of reasonable descriptions of the data.

For any likelihood function for a parameter θ, we're interested in ratios like $\frac{\text{likelihood of }\theta_1}{\text{likelihood of }\theta_2}$ so multiplicative terms not involving θ drop out and can be ignored. We sometimes indicate the presence of irrelevant multiplicative constants by the letter "c", for "constant," as in the y-axis of Figure 2.8.

In the examples we've discussed so far there has been a single parametric family of distributions — Binomial for the Slater School and Poisson for seedling emergence — and our analysis has found which values of the parameter provide reasonably good descriptions of the data. The next example shows a different type of statistics problem: determining whether a population can be well-described by a single member of a parametric family, or whether the population ought to be divided into two subpopulations, each with its own set of parameters.

Example 2.5 (Galton's Heights, continued)
Figure 1.17 raised the question of whether the heights of the parents in Galton's sample can be well-described by a single Normal density, or whether they ought to be described by one Normal distribution for the mothers and another for the fathers. Most of us have good intuition about the matter, so the point of Example 2.5 is not to learn the answer, but to learn the statistical analysis and see how it accords with our intuition.

Let's begin with the single-Normal approximation. A Normal distribution has parameters μ and σ, so the likelihood is a function of (μ, σ). There are 410 parents and we treat their heights as random variables $X_1, \ldots, X_{410}$. We model the X_i's as independent of each other, so the likelihood function is

$$
\begin{aligned}
&p_{(\mu,\sigma)}(x_1, \ldots, x_{410}) \\
&\quad = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2} \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma}\right)^2} \times \cdots \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_{410}-\mu}{\sigma}\right)^2} \\
&\quad = \left(\frac{1}{2\pi}\right)^{205} \left(\frac{1}{\sigma}\right)^{410} e^{-\frac{1}{2\sigma^2}\left[(x_1-\mu)^2+\cdots+(x_{410}-\mu)^2\right]} \\
&\quad = c\left(\frac{1}{\sigma}\right)^{410} e^{-\frac{1}{2\sigma^2}\left[(x_1-\mu)^2+\cdots+(x_{410}-\mu)^2\right]} \qquad (2.3)
\end{aligned}
$$

where we used the letter "c" to indicate a multiplicative constant that does not involve μ or σ.

Figure 2.9 displays the likelihood function in (2.3), rescaled so its maximum is 1. Regions of high likelihood are white; regions of low likelihood are black. The maximum occurs at $(\hat{\mu} = 66.66, \hat{\sigma} = 3.645)$, which can be calculated with the formula on page 79. To describe the data nearly as well as the maximum requires μ to be between about 66 and 67 and σ to be between about 3.4 and 3.9 or so.

Figure 2.10 displays the density of parents' heights in Galton's data, estimated from 1-inch bins, just as in Figure 1.17, along with 9 approximating Normal densities with μ being one of $(66, 66.66, 67)$ and σ being one of $(3.4, 3.645, 3.9)$. These densities span the range of good approximating Normal densities.

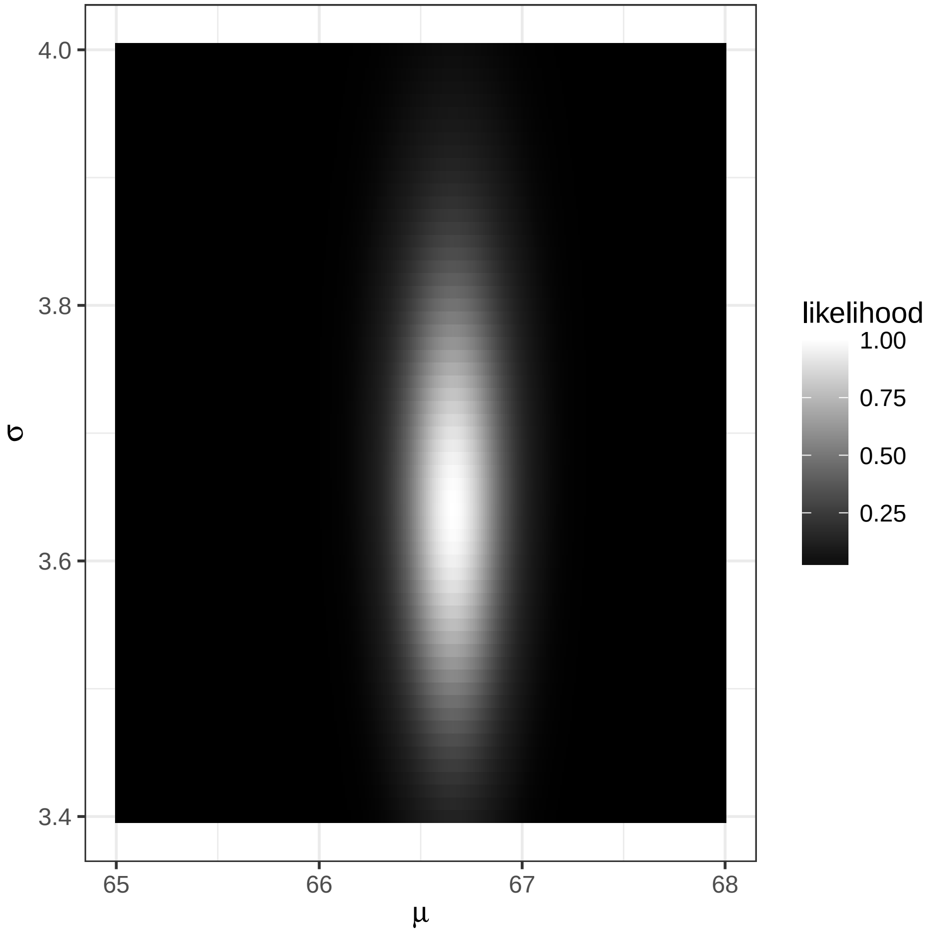


Figure 2.9: The likelihood function (2.3), rescaled so its maximum is 1.

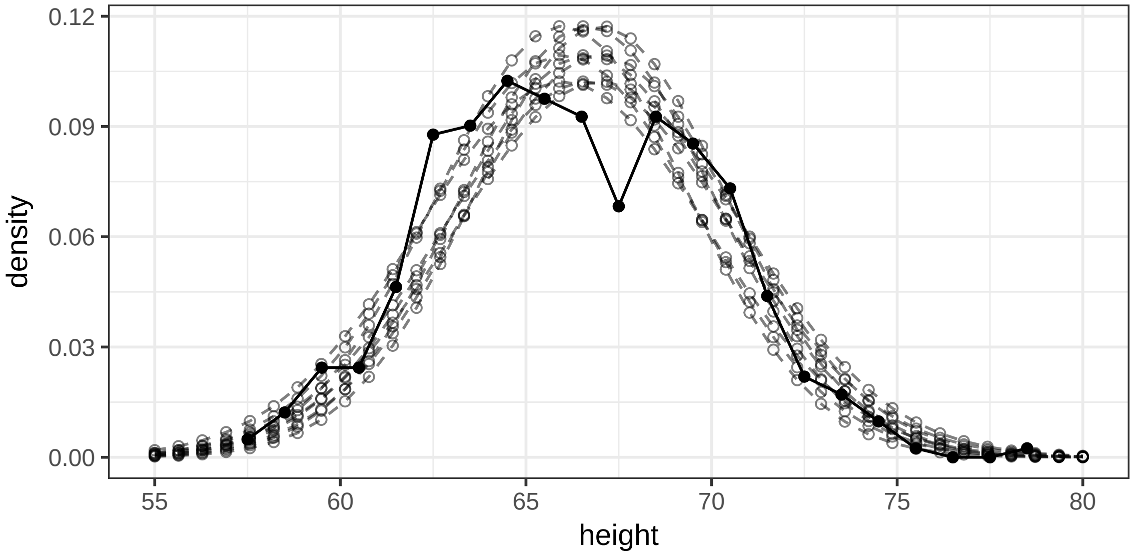


Figure 2.10: The density of parents' heights in Galton's data (black) with several approximating Normal densities (gray). $\mu \in \{66, 66.66, 67\}$; $\sigma \in \{3.4, 3.645, 3.9\}$.

Now let's consider models in which mothers' heights have a Normal density with parameters $(\mu_{\mathsf{M}}, \sigma_{\mathsf{M}})$ and fathers' heights have a Normal density with parameters $(\mu_{\mathsf{F}}, \sigma_{\mathsf{F}})$. In our dataframe, the first 205 heights belong to mothers and the last 205 to fathers, so the likelihood function is

$$\begin{aligned}
p_{(\mu_{\mathsf{M}},\sigma_{\mathsf{M}},\mu_{\mathsf{F}},\sigma_{\mathsf{F}})}(x_1,\ldots,x_{410}) &= \\
&\frac{1}{\sqrt{2\pi}\sigma_{\mathsf{M}}} e^{-\frac{1}{2}\left(\frac{x_1-\mu_{\mathsf{M}}}{\sigma_{\mathsf{M}}}\right)^2} \times \cdots \times \frac{1}{\sqrt{2\pi}\sigma_{\mathsf{M}}} e^{-\frac{1}{2}\left(\frac{x_{205}-\mu_{\mathsf{M}}}{\sigma_{\mathsf{M}}}\right)^2} \\
&\times \frac{1}{\sqrt{2\pi}\sigma_{\mathsf{F}}} e^{-\frac{1}{2}\left(\frac{x_{206}-\mu_{\mathsf{F}}}{\sigma_{\mathsf{F}}}\right)^2} \times \cdots \times \frac{1}{\sqrt{2\pi}\sigma_{\mathsf{F}}} e^{-\frac{1}{2}\left(\frac{x_{410}-\mu_{\mathsf{F}}}{\sigma_{\mathsf{F}}}\right)^2} = \\
c\left(\frac{1}{\sigma_{\mathsf{M}}}\right)^{205} & e^{-\frac{1}{2\sigma_{\mathsf{M}}^2}\left[(x_1-\mu_{\mathsf{M}})^2+\cdots+(x_{205}-\mu_{\mathsf{M}})^2\right]} \left(\frac{1}{\sigma_{\mathsf{F}}}\right)^{205} e^{-\frac{1}{2\sigma_{\mathsf{F}}^2}\left[(x_{206}-\mu_{\mathsf{F}})^2+\cdots+(x_{410}-\mu_{\mathsf{F}})^2\right]} \\
&= cp_{(\mu_{\mathsf{M}},\sigma_{\mathsf{M}})}(x_1,\ldots,x_{205}) \times p_{(\mu_{\mathsf{F}},\sigma_{\mathsf{F}})}(x_{206},\ldots,x_{410}) \quad (2.4)
\end{aligned}$$

Because we modelled the heights as independent of each other, in particular the mothers' heights are independent of the fathers' heights, the likelihood function (2.4) breaks into a product of two likelihood functions, one for $(\mu_{\mathsf{M}}, \sigma_{\mathsf{M}})$ using the mothers' data and one for $(\mu_{\mathsf{F}}, \sigma_{\mathsf{F}})$ using the fathers' data. Figure 2.11 displays the two likelihood

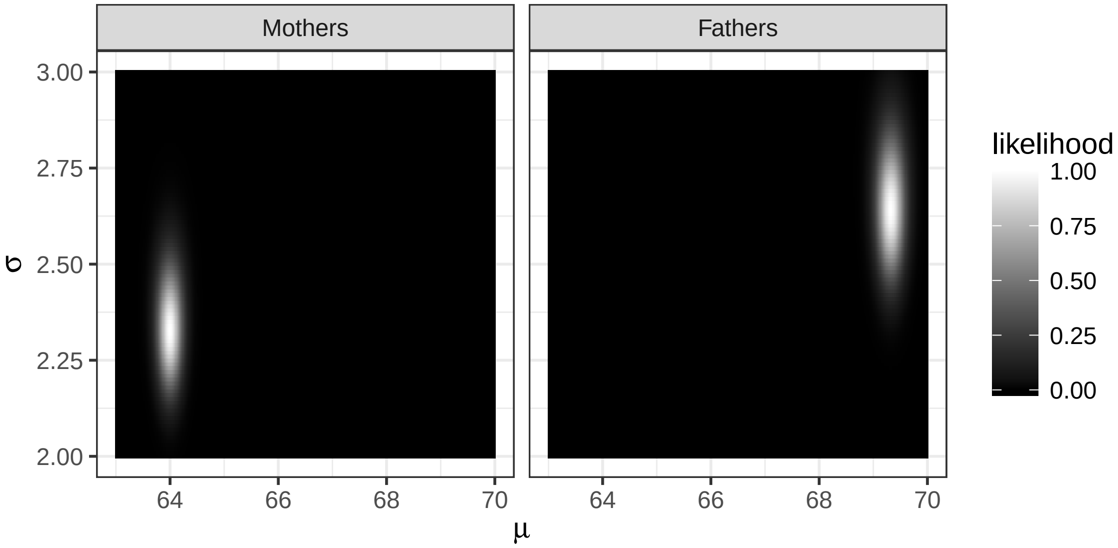


Figure 2.11: Likelihood functions from (2.4). Each likelihood function has been rescaled so its maximum is 1.

functions. Mothers' heights are best described by a Normal density with mean $\mu_\mathsf{M} \approx 64$ and $\sigma_\mathsf{M} \approx 2.3$ while fathers' heights are best described with mean $\mu_\mathsf{F} \approx 69$ and $\sigma_\mathsf{F} \approx 2.6$. (The m.l.e.'s can be calculated with the formula on page 79.) Fathers are taller and more variable in height than mothers. Compare Figure 2.11 to Figure 2.9. We had to expand the x-axis and shift the y-axis to show values of μ_M, σ_M, μ_F, and σ_F with high likelihood. Do you see why?

Using $\hat{\mu}_\mathsf{M}$, $\hat{\sigma}_\mathsf{M}$, $\hat{\mu}_\mathsf{F}$, and $\hat{\sigma}_\mathsf{F}$ for the two-Normal model and $\hat{\mu}$ and $\hat{\sigma}$ for the one-Normal model, we can say how much better the two-Normal model can describe the data by the ratio

$$\frac{p_{\hat{\mu}_\mathsf{M},\hat{\sigma}_\mathsf{M},\hat{\mu}_\mathsf{F},\hat{\sigma}_\mathsf{F}}(x_1,\ldots,x_{410})}{p_{\hat{\mu},\hat{\sigma}}(x_1,\ldots,x_{410})} \approx 2.7 \times 10^{68}.$$

The two-Normal model is a vast improvement over the one-Normal model. Both models are displayed in Figure 2.12.

Example 2.6 (FACE)
The amount of carbon dioxide, or CO_2, in the Earth's atmosphere has been steadily increasing over the last century or so, as illustrated in Exercise 4 of Chapter 2. CO_2 is a greenhouse gas that traps heat in the atmosphere instead of letting it radiate out,

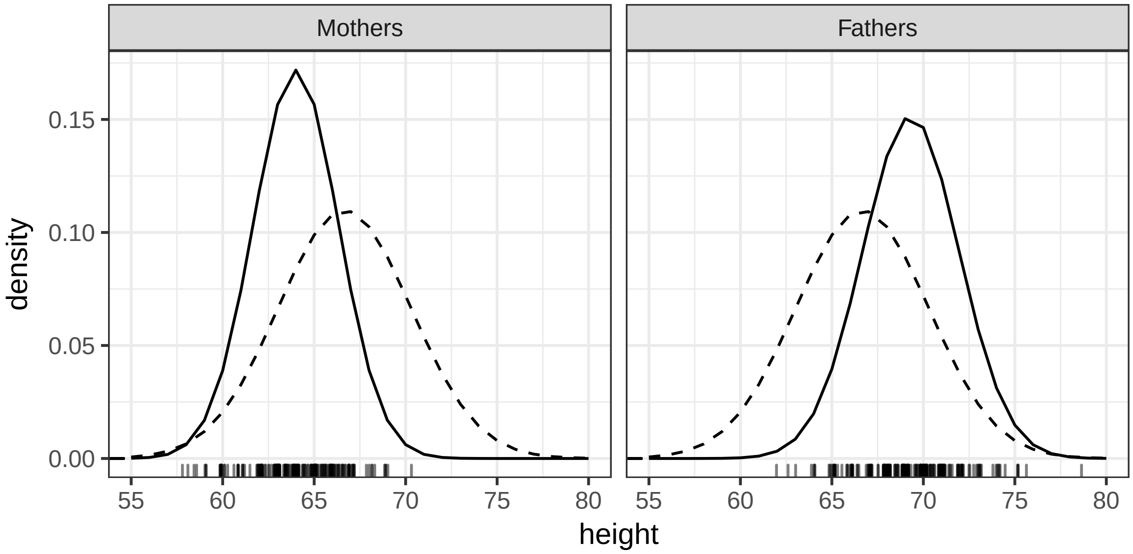


Figure 2.12: Heights of Mothers and Fathers in Galton's data. The solid curves are the best Normal approximations to Mothers' and Fathers' heights separately. The dashed curve is the best Normal approximation to all parents' heights combined, and is the same in each facet. The rug shows the actual data points, jittered horizontally to avoid overplotting.

so an increase in atmospheric CO_2 will eventually result in an increase in the Earth's temperature. But what is harder to predict is the effect on the Earth's plants. Carbon is a nutrient needed by plants, so it's possible that an increase in CO_2 will cause an increase in plant growth which in turn will partly absorb some of the extra carbon.

To learn about plant growth under elevated CO_2, ecologists began by conducting experiments in greenhouses. In a greenhouse, two sets of plants could be grown under conditions that are identical except for the amount of CO_2 in the atmosphere. But the controlled environment of a greenhouse is quite unlike the uncontrolled natural environment, so, to gain verisimilitude, experiments soon moved to open-top chambers. An open-top chamber is a space, typically a few meters in diameter, enclosed by a solid, usually plastic, wall and open at the top. CO_2 can be added to the air inside the chamber. Because the chamber is mostly enclosed, not much CO_2 will escape, and more can be added as needed. Plants grown in chambers with excess CO_2, can be compared to plants grown in chambers with normal CO_2. But as with greenhouses, open-top chambers are not completely natural and ecologists wanted to conduct experiments under even more natural conditions.

To that end, in the late 1980's the Office of Biological and Environmental Research in the U.S. Department of Energy (DOE) began supporting research using a technology called FACE, or *Free Air CO_2 Enrichment*, developed at the Brookhaven National Laboratory. As the lab's webpage `WWW.FACE.BNL.GOV/FACE1.HTM` explains

> "FACE provides a technology by which the microclimate around growing plants may be modified to simulate climate change conditions. Typically CO2-enriched air is released from a circle of vertical pipes into plots up to 30m in diameter, and as tall as 20 m.
>
> "Fast feedback control and pre-dilution of CO2 provide stable, elevated [CO2] simulating climate change conditions.
>
> "No containment is required with FACE equipment and there is no significant change in natural air-flow. Large FACE plots reduce effects of plot edge and capture fully-functioning, integrated ecosystem-scale processes. FACE Field data represent plant and ecosystems responses to concentrations of atmospheric CO2 expected in the mid-twenty-first century."

See the website for pictures and more information. In a FACE experiment, CO_2 is released into some treatment plots. The level of CO_2 inside the plot is continually monitored so more CO_2 can be released as needed to keep the amount of CO_2 in the atmosphere at some pre-specified level, typically the level expected in the mid-21st century. Other plots are reserved as control plots. Plant growth in the treatment plots is compared to that in the control plots.

Because a FACE site is not enclosed, CO_2 continually drifts out of the site and needs to be replenished. Keeping enough CO_2 in the air is the major expense in FACE experiments.

The first several FACE sites were in Arizona (sorghum, wheat, cotton), Switzerland (rye grass, clover) and California (native chaparral). All of these contained low-growing plants. By the early 1990's, ecologists wanted to conduct a FACE experiment in a forest, and such an experiment was proposed by investigators at Duke University, to take place in Duke Forest. Before the experiment could be funded the investigators had to convince DOE it would be worthwhile. In particular, they had to demonstrate that the experiment would have a good chance of uncovering whatever growth differences would exist between treatment and control. The demonstration was carried out by computer simulation. Below, we present a modified version of the simulation's R script — modified to use the statistical material in this book. The actual R script was different.

The main steps in the simulation are

1. Load the necessary packages. `ggplot2` is for plotting; `reshape2` is for the `melt` command.

2. The plan is for the experiment to be carried out in six rings. Three will be treated with extra CO_2; three will be control. We set that number at the beginning of the script so it can be easily changed, if needed.

3. Biologists will measure the size of each tree, each year. We set the number of trees per ring at 20, as an approximation to the conditions in this particular forest, for the size of the proposed FACE rings. It can be easily changed to simulate what would happen under different conditions.

4. The total number of trees is the number of trees per ring times the number of rings per treatment times the number of treatments.

5. From past experience we know the typical growth rate of trees in this forest under natural conditions. It's about 2% per year.

6. The treated trees will likely grow faster than the control trees. In the script below we set the expected growth rate of treated trees to be about 4% per year. Several considerations go into setting this number, including the following.

   (a) We don't know exactly what the new growth rate will be, but we set it at the beginning of the simulation in a way that can be easily changed, so we can run the simulation multiple times under a variety of conditions.

   (b) It is both less important and more difficult to detect the treatment effect if it is small. That is, if the new growth rate is only 2.001%, we will have only a small chance of detecting it, but we won't care very much if we don't detect it. We set the new growth rate in the script to a value that may be important to detect. It is set early in the script so it can be easily changed, to see what would happen under other conditions.

7. We know from past experience that different trees grow at different rates each year. We set `sigma` based on that experience to describe the typical standard deviation of the trees' growth rates.

8. We set how many years' growth we will simulate.

9. We will simulate `n.years` growth for each tree and store the results in a dataframe called `sims`. The first column of `sims` will have the tree number; the second column will record whether it's a control or a treatment tree; and the third column will have the size of the tree in year 0. Because we're interested in relative growth, the initial size of each tree is irrelevant, so we set them all to 1 for convenience.

10. `rate.ave` contains the typical growth rate of each tree, and depends on whether it's a control or treatment tree.

11. `rate` will eventually contain an individualized growth rate for each tree in each year. It's initially set to `NA`.

12. For each simulated year, do two things.

    (a) Simulate an individual growth rate for that tree by generating a random number from the Normal distribution. See whether you can understand the `mean = ...` and `sd = ...` parts of the `rnorm` command.

    (b) Simulate the size of each tree by multiplying its previous size by its growth rate.

13. Create the names of the dataframe's columns that contain the simulated sizes.

14. We now have a dataframe containing the simulated sizes. Each year is one column of the dataframe. To make plotting easier, I want to convert to a dataframe in which all the sizes are in one column. That's what `sims.m <- melt (...)` does.

15. If you examine `sims.m` at this point in the simulation, you'll see it has a column called `year` with values `year0`, `year1`, ..., `year10`. The `as.numeric ( sub (...) )` command converts them to the numbers 0 through 10.

16. Lastly, we draw a plot to show the simulation results.

```
library ( ggplot2 )
library ( reshape2 )

n.trees <- 20 # trees per ring
n.rings <- 3 # rings per treatment
trees.tot <- n.trees * n.rings * 2 # total number of trees in the experiment

rate.contr <- 1.02 # expected growth rate of control trees
rate.treat <- 1.04 # possible growth rate of treated trees
sigma.rate <- 0.1 # standard deviation of growth rate

n.years <- 10

sims <- data.frame ( tree = 1:trees.tot,
                     treat = rep ( c ( "contr", "treat" ),
                                   each = n.trees * n.rings
                                 ),
                     year0 = 1 # size of each tree in year 0
                   )

rate.ave <- ifelse ( sims$treat == "contr", rate.contr, rate.treat )
rate <- matrix ( NA, nrow = nrow(sims), ncol = n.years )
for ( i in 1:n.years )
  rate[,i] <- rnorm ( trees.tot, mean = rate.ave, sd = sigma.rate * ( rate.ave - 1 ) )
  sims[,i+3] <- sims[,i+2]*rate[,i]

names(sims)[3 + 1:n.years] <- paste ( "year", 1:n.years, sep="" )

sims.m <- melt ( sims, id.vars = c ( "tree", "treat" ),
                       measure.vars = 3:(3+n.years),
                       variable.name = "year",
                       value.name = "size"
               )
sims.m$year <- as.numeric ( sub ( "year", "", sims.m$year ) )


pdf ( "face_sim.pdf", height = 3, width = 6 )
p <- ggplot ( sims.m, aes ( x = year, y = size, shape = treat ) )
p + geom_jitter ( height = 0, width = .1 ) +
    scale_x_continuous ( breaks = 0:10 ) +
    scale_y_continuous ( name = "size of tree" ) +
    scale_shape ( name = "treatment" ) +
    theme_bw()
dev.off()
```

Figure 2.13 plots the results of one run of the R script. The question is, from simulated data such as this, can we tell whether the treated trees are growing faster than the control trees? To put it another way,

1. Does the Normal distribution give a reasonable approximation to the tree sizes?

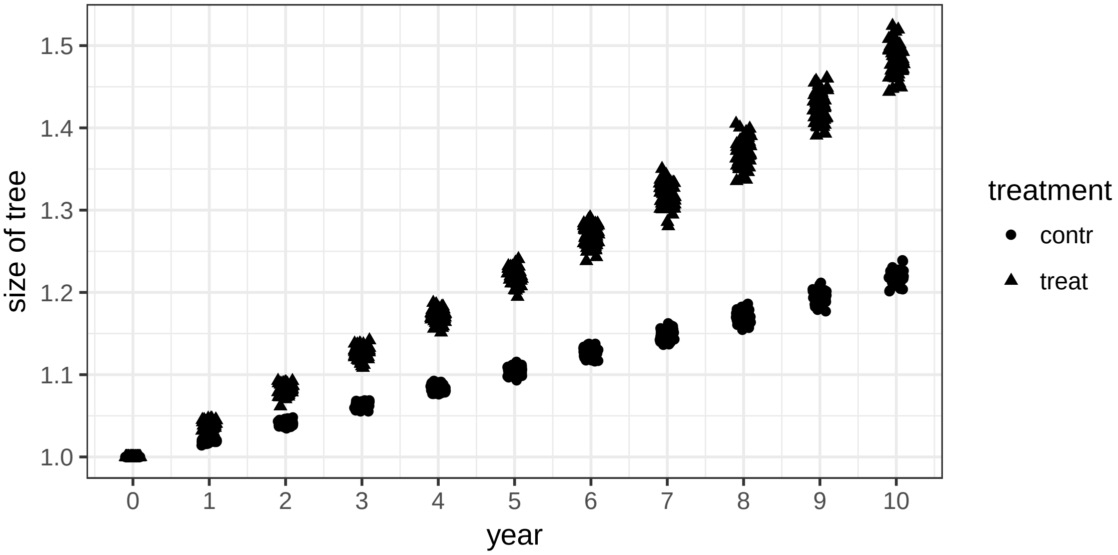


Figure 2.13: Simulated tree sizes in the FACE experiment

2. If it does, then how much better are sizes described by two Normals, with parameters $(\mu_{\text{contr}}, \sigma_{\text{contr}})$ and $(\mu_{\text{treat}}, \sigma_{\text{treat}})$, than by a single Normal with a single (μ, σ) that applies to both sets of trees combined?

If the data are much better described by two Normals, then the experiment is a success: we can detect that the treated trees are growing faster. Figure 2.13 suggests the experiment will be a success, at least under the conditions in this simulation. The DOE did decide to fund the proposal for a FACE experiment in Duke Forest, at least partly because of the demonstration that such an experiment would have a reasonably large chance of success.

2.3 Exercises

1. Figure 2.1g has 18 distinct points but Figure 2.1h has 20. Why?

2. R comes with many built-in data sets. You can type `data()` to see what they are. This example uses the `ToothGrowth` data set. You can type `?ToothGrowth` for an explanation of the data. Our goal is to plot the data in various ways to see how the dose of vitamin C and the delivery method affect the length of odontoblasts.

> The value of `dose` is either 0.5, 1.0, or 2.0, so R treats `dose` as a continuous variable. However, in this exercise we want to treat `dose` as categorical. You can do that by replacing `dose` with `as.factor(dose)` in your R commands. The value of `supp` is either `OJ` or `VC` so R already knows `supp` is not a number. You don't have to do anything special to get R to treat it as categorical.

   (a) Make a histogram of `len`. Make a dot plot of `len`. Make a density plot of `len`. What do you conclude? Which method of display do you prefer? Does there appear to be one population or several?

   (b) Make a plot with `len` on the x-axis and `supp` on the y-axis. Jitter the points vertically. Does delivery method seem to affect length?

   (c) Repeat the previous plot, but use different shapes for different doses. Does delivery method seem to affect length?

   (d) Make a plot with `len` on the x-axis and `dose` on the y-axis. Jitter the points vertically. Does dose seem to affect length?

   (e) Repeat the previous plot, but use different shapes for different delivery methods. Does dose seem to affect length? Does the amount by which dose affects length seem to be same for `OJ` and `VC`?

3. Another data set that comes with R is `faithful`, which R describes as "Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA." As usual, `?faithful` gives more complete information. Old Faithful is an attraction for tourists who want to witness its eruptions. To help them, park rangers post predictions of the next eruption time. This exercise investigates whether data collected by the rangers can help them predict the next eruption time more accurately.

   (a) Plot `waiting` vs. `eruptions`. What do you learn from this plot? Is the duration of one eruption useful in predicting the time until the next eruption.

(b) Can the duration of one eruption help predict the duration of the next? Make a plot with the duration of eruption i on the x-axis and the duration of eruption $i + 1$ on the y-axis. You can make a new variable with a command such as
`faithful$next.erup <- c ( faithful$eruptions[-1], NA )`.
Plot `next.erup` vs. `eruptions`. What do you learn?

(c) Can the previous waiting time help predict the next waiting time? Plot `next.wait` vs. `waiting`. What do you learn?

4. Climate scientists are in wide agreement that human activity affects the earth's climate. It is easy to find plots of global temperature vs. time to illustrate the climate trend. Climate scientists also agree that the way in which human activity affects climate is by releasing heat-trapping gases into the atmosphere. Carbon dioxide, CO_2, is one of those gases. R comes with the data set `co2`, which gives the atmospheric concentration of CO_2 at Mauna Loa, measured monthly from 1959 through 1997. `co2` is in an R format called *time series*. Because `ggplot` is designed to plot data frames, we should convert `co2` to a data frame with a command such as

```
co2.df <- data.frame ( month = 1:length(co2),
                       co2 = as.numeric ( co2 )
                     )
```

Plot `co2` vs. `month`. What do you learn?

5. Refer to the hotdog data in Example 2.1.

(a) Plot the sodium content vs. the calorie content of hotdogs.

(b) Plot the sodium content vs. the calorie content of hotdogs, but use different plotting symbols for beef, meat, and poultry.

(c) Plot the sodium content vs. the calorie content of hotdogs, but use different facets for beef, meat, and poultry.

(d) Summarize your findings.

6. Is poverty related to academic performance in school? The file `schpov.csv` at this text's website contains relevant data from the Durham, NC school system in 2001. The first few lines are

```
  pfl eog type
1  66  65    e
2  32  73    m
3  65  65    e
```

Each school in the Durham public school system is represented by one line in the file. The variable `pfl` stands for *percent free lunch* and records the percentage of the school's student population that qualifies for a free lunch program. It is an indicator of poverty. The variable `eog` stands for *end of grade*. It is the school's average score on end of grade tests and is an indicator of academic success. Finally, `type` indicates the type of school — `e`, `m`, or `h` for elementary, middle, or high school, respectively. You are to investigate whether `pfl` is predictive of `eog`.

(a) Read the data into `R` and plot it in a sensible way. Use different plot symbols for the three types of schools.

(b) Does there appear to be a relationship between `pfl` and `eog`? Is the relationship the same for the three types of schools?

(c) During the 2000-2001 school year Duke University, in Durham, NC, sponsored a tutoring program in one of the elementary schools. Many Duke students served as tutors. From looking at the plot, and assuming the program was successful, can you figure out which school it was?

7. This exercise relies on Example 1.2 about the Slater school where there were 8 cancers among 145 teachers. Figure 2.7 shows the likelihood function.

   (a) If X is the number of invasive cancers at Slater and we adopt the model $X \sim \text{Bin}(145, p)$, find the m.l.e. $\hat{p}$.

   (b) Suppose there had been 9 cancers among the 145 teachers. How would that affect the likelihood function? Make a plot similar to Figure 2.7, but pretending there had been 9 cancers among 145 teachers. Compare to Figure 2.7. What is the result? Does it make sense? Try other numbers if it helps you see what is going on.

   (c) If there had been 9 cancers among the 145 teachers, what would $\hat{p}$ be?

   (d) How many cancers would there have to be among the Slater employees, i.e. what would x have to be, in order for $\frac{\text{Pr}_{\hat{p}}[x \text{ cancers}]}{\text{Pr}_{.03}[x \text{ cancers}]} \geq 100$?

(e) Suppose the same incidence rate had been found among more teachers. How would that affect the likelihood function? Make a plot similar to Figure 2.7, but pretending there had been 80 cancers among 1450 teachers. Compare to Figure 2.7. What is the result? Does it make sense? Try other numbers if it helps you see what is going on.

8. This exercise continues Example 1.5 on the S&P 500. There we said "the density falls slightly more steeply to the left of 0 than to the right, which means there are more days with positive than with negative returns." This exercise will examine the situation quantitatively. We'll begin by looking only at the days where the S&P 500 return can be calculated and is not equal to 0.

   (a) How many days are represented in the S&P 500 dataframe? For how many of them is the return zero? Not Available? Available and non-zero? We're interested in returns on the available non-zero days. Use the letter N to denote that number of days.

   (b) Make a new vector `sp.small` containing the available non-zero returns. Make sure the length of `sp.small` is equal to N.

   (c) Let X be the number of entries in `sp.small` that are positive. We don't know X in advance, but we'll treat daily returns as independent of each other, so we'll treat X as a $\text{Bin}(N, p)$ random variable and try to discover whether $p = .5$ describes the data well or whether there is a value $p \neq .5$ that describes the data much better. Find the actual number of positive returns x.

   (d) Find the m.l.e. $\hat{p}$.

   (e) Find $\Pr_{p=.5}[X = x]$ and $\Pr_{p=\hat{p}}[X = x]$.

   (f) How much better does $p = \hat{p}$ describe the data than $p = 0.5$?

   Let $p_{\text{pos}} = \Pr[\text{return} > .01 \,|\, \text{return} > 0]$ and $p_{\text{neg}} = \Pr[\text{return} < -.01 \,|\, \text{return} < 0]$. p_{pos} and p_{pos} are the probabilities of big swings in the positive or negative directions, respectively. (I used 0.01 as an arbitrary but convenient threshhold.) The observation that "the density falls slightly more steeply to the left of 0 than to the right" suggests $p_{\text{pos}} > p_{\text{neg}}$.

   (g) How much better can the data be described with $p_{\text{pos}} > p_{\text{neg}}$ than with $p_{\text{pos}} = p_{\text{neg}}$?

9. This exercise continues Exercise 19 in Chapter 1 about randomized response. Let p be the fraction of the population that uses cocaine. We conduct an

experiment to learn about p by giving 100 people the randomized response question. Let X be the number of people who answer "yes".

(a) What is the likelihood function for p?

(b) Plot the likelihood function for p if $X = 20$, if $X = 50$, and if $X = 90$.

(c) Now suppose we sample 1000 people. Plot the likelihood function for p if $X = 200$, if $X = 500$, and if $X = 900$.

(d) About how much does the larger sample size help you narrow down the range of reasonable values of p?

10. Refer to Example 2.5. Verify the calculations of $\hat{\mu}$, $\hat{\sigma}$, $\hat{\mu}_{\mathrm{M}}$, $\hat{\sigma}_{\mathrm{M}}$, $\hat{\mu}_{\mathrm{F}}$, and $\hat{\sigma}_{\mathrm{F}}$.

11. Our primary means of quantitative inference is likelihood ratios. But how big a likelihood ratio does it take before we say one model describes the data much better than another? Page 77 referred to a reference experiment in which we tossed a coin, got a sequence of Heads, and compared a two-headed description of the coin to a one-head and one-tail description. Page 77 talked about tossing the coin twice and getting two Heads. What would the likelihood ratio be if we tossed the coin 10 times and got 10 Heads? What would the likelihood ratio be if we tossed the coin 10 times and got 9 Heads?

12. This exercise continues Example 1.7 on the VA lung cancer trial. In that example we compared the standard and new treatments visually. In this exercise we'll use exponential distributions to compare them quantitatively.

(a) Combining the patients on standard treatment with those on the new treatment, verify that the m.l.e. is $\hat{\lambda} \approx 1/121.6277$.

(b) Treating the standard and new treatments separately, find the mle's $\hat{\lambda}_{\text{standard}}$ and $\hat{\lambda}_{\text{new}}$.

(c) Find $p_{\hat{\lambda}}(x_1, \ldots, x_{137})$ and $p_{\hat{\lambda}_{\text{standard}}, \hat{\lambda}_{\text{new}}}(x_1, \ldots, x_{137})$.

(d) How much better can the data be described by different parameters for the two groups separately than by one common parameter for the two groups combined? Does that agree with the conclusion in Example 1.7?

13. This exercise continues Example 2.4. Make figures similar to Figure 2.8 but for 1991, 1992, 1994, 1995, 1996, and 1997. Use the columns `Xyy.t`, not `Xyy.1` Compare them to Figure 2.8 and to each other. Does it appear that different years have different emergence rates?

14. The book *Data* by Andrews and Herzberg contains lots of data sets that have been used for various purposes in statistics. One famous data set records the annual number of deaths by horsekicks in the Prussian Army from 1875-1894 for each of 14 corps. Download the data from `STATLIB` at `HTTP://LIB.STAT.CMU.EDU/DATASETS/ANDREWS/T04.1`. (It is Table 4.1 in the book.) Let Y_{ij} be the number of deaths in year i, corps j, for $i = 1875, \dots, 1894$ and $j = 1, \dots, 14$. The Y_{ij}s are in columns 5–18 of the table.

    (a) What are the first four columns of the table?

    (b) What is the last column of the table?

    (c) What is a good model for the data?

    (d) Suppose you model the data as i.i.d. Poi(λ). (Yes, that's a good answer to the previous question.)

    i. Plot the likelihood function for λ.

    ii. Find $\hat{\lambda}$.

    iii. What can you say about the rate of death by horsekick in the Prussian calvary at the end of the 19th century?

    (e) Is there any evidence that different corps had different death rates? How would you investigate that possibility?

15. John is a runner and frequently runs from his home to his office. He wants to measure the distance, so once a week he will drive to work and measure the distance on his car's odometer. He also drives a lot each week in addition to his commute. Unfortunately, John's odometer records distance only to the nearest mile and changes abruptly from one digit to the next. When the odometer displays a digit x, it is not possible to infer how close it is to becoming $x + 1$. John will drive the route ten times and record digits $X_1, \dots, X_{10}$. Let D be the exact distance from home to office.

    (a) Why doesn't the odometer record the same distance each day? I.e., why doesn't $X_1 = \cdots = X_{10}$?

    (b) What are the possible values of each X_i?

    (c) What is a good model for $X_1, \dots, X_{10}$? Make reasonable assumptions as needed.

    (d) Suppose the data turn out to be $x_1 = 3$, $x_2 = 3$, $x_3 = 3$, $x_4 = 4$, $x_5 = 3$, $x_6 = 4$, $x_7 = 3$, $x_8 = 3$, $x_9 = 4$, and $x_{10} = 3$. Find the m.l.e. of D.

16. Continue Exercise 20 from Chapter 1. The autoganzfeld trials resulted in $X = 122$ direct hits.

    (a) What is the parameter in this problem?

    (b) Plot the likelihood function.

    (c) What do you conclude?

17. This exercise continues Example 2.6 about simulations to discover whether running a FACE experiment is likely to uncover differences in the growth of treated and untreated trees. Figure 2.13 assessed the simulation results graphically. Here, we're going to assess them quantitatively.

    (a) Run the simulation.

    (b) For year 2, how much better can the simulated data be described by two Normals — one for treatment, one for control — than by one Normal for all trees combined?

    (c) The simulation was run under the assumption that the average growth rate of treated trees would be about twice the average growth rate of control trees. Rerun the simulation under the assumption that the average growth rate of treated trees would be about 1.5 times the average growth rate of control trees. How much better is the two-Normal model better than the one-Normal model?

    (d) The simulation was run under the assumption that the standard deviation of tree-to-tree growth rates is about 10% of the typical growth rate. Rerun the simulation under the assumption that the standard deviation is about 50% of the typical growth rate. How much better is the two-Normal model better than the one-Normal model?

    (e) With the simulation of the previous part, how many years of simulated data would be needed so the advantage of the two-Normal model over the one-Normal model is about the same as the advantage after two years of data in the original simulation?

18. This exercise continues Example 2.6 about the FACE experiment. To describe the size of trees, ecologists sometimes use Diameter at Breast Height, or DBH. The dominant tree species in that part of the Duke Forest is *pinus taeda*, or loblolly pine. DBH was recorded every year for each loblolly pine tree in the FACE experiment. The dataset in the file `pinecones.txt`, which can be read

with the `read.table` command, contains data on the 644 loblolly pines in the FACE experiment. The columns are

ring FACE was run in six rings. Rings 1, 5, and 6 were control; 2, 3, and 4 were treatment.

ID Each tree has a unique identifier.

xcoor, ycoor The x- and y-coordinates of the tree.

spec Species. For this file spec is always *pita*, for *pinus taeda*.

dbh Diameter at Breast Height. The size of the tree at the beginning of the experiment.

X1998, X1999, X2000 The number of pine cones on the tree in 1998, 1999, and 2000.

The data show there were about 100 trees per ring, whereas the simulation in Example 2.6 was done with 20 trees per ring. You could think about how that affects the validity of the simulation.

One potential effect of elevated CO_2 is for trees to reach sexual maturity and hence be able to reproduce earlier than otherwise. If they do mature earlier, ecologists would like to know whether that's due only to their increased size, or whether trees will reach maturity not just at younger ages, but also at smaller sizes. Sexually mature trees can produce pine cones but immature trees cannot. So to investigate sexual maturity, a graduate student counted the number of pine cones on each tree. For each tree let X be its DBH and Y be either 1 or 0 according to whether the tree has pine cones.

(a) Add a new column to the dataframe that indicates whether a tree is treated or control. The new column should have one value for trees in the control rings and another value for trees in the treated ring.

(b) Plot number of pinecones vs. DBH in a way that shows whether treated trees produce more cones than control trees.

(c) Add a new column to the dataframe that indicates the year in which a tree first produces pinecones. We'll call that the tree's maturity year.

(d) Plot maturity year vs. DBH in a way that shows whether treated trees mature earlier than control trees.

Chapter 3

Models

Statistics uses probability to model real-world phenomena. For example, earlier in this book we used a binomial model to describe cancers at the Slater School, a Poisson model to describe seedling emergence in a forest, and a Normal model to describe the heights of parents in Galton's study of inheritance. The probability models are simplified descriptions of the real world. The binomial model for cancer is simplified because it doesn't account for employees' age, sex, length of employment at Slater, genetic predisposition, and other factors that influence the development of cancer. The Poisson model for seedlings is simplified because it doesn't account for different forest conditions in different locations, for wind direction, and for various other factors. The Normal model for height is simplified because it doesn't account for genetics, nutrition, health, and various other factors. Nevertheless, the models are useful. When we use a Normal model for heights, we're not saying that parents' heights are determined solely by random forces; but we are saying they look enough like they're determined by random forces for our present purpose. For another purpose, the Normal model might be insufficient. As Box (1979) says,

> "Now it would be very remarkable if any system existing in the real world could be *exactly* represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law PV = RT relating pressure P, volume V and temperature T of an 'ideal' gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation . . . .
>
> "For such a model there is no need to ask the question 'Is the model true?' If 'truth' is to be the 'whole truth' the answer must be 'No'. The only question of interest is 'Is the model illuminating and useful?' "

Kaplan (2011) also says some useful things about models in statistics.

> "*A model is a representation for a particular purpose.* ....
>
> "There are three main uses for statistical models. They are closely related, but distinct enough to be worth enumerating.
>
> **Description.** Sometimes you want to describe the range or typical values of a quantity. For example, what's a 'normal' white blood cell count? Sometimes you want to describe the relationship between things. Example: What's the relationship between the price of gasoline and consumption by automobiles?
>
> **Classification or prediction.** You often have information about some observable traits, qualities, or attributes of a system you observe and want to draw conclusions about other things that you can't directly observe. For instance, you know a patient's white blood-cell count and other laboratory measurements and want to diagnose the patient's illness.
>
> **Anticipating the consequences of interventions.** Here, you intend to do something: you are not merely an observer, but an active participant in the system. For example, people involved in setting or debating public policy have to deal with questions like these: To what extent will increasing the tax on gasoline reduce consumption? To what extent will paying teachers more increase student performance?"

This chapter will review some of the models we've already seen and introduce a few new ones. It will also touch on model checking.

3.1 Model Review

We've seen and used the following probability models.

Binomial The Binomial distribution is used to model seemingly random events that result in either success or failure (1 or 0), that all have the same probability of success, and that are independent of each other. We modelled the occurrence of cancer at the Slater school with a Binomial distribution in Example 1.2.

Poisson The Poisson distribution is used to model events that arise seemingly at random in a block of space or time, when the arrival rate is roughly constant over the entire block. We modelled the emergence of tree seedlings with a Poisson distribution in Example 1.3.

Normal The Normal distribution is used to model data that appear to have a symmetric unimodal density. We modelled parents' heights with a Normal distribution in Example 1.4.

Exponential The Exponential distribution is used to model the time between one event and the next when the arrival rate of events is roughly constant. We modelled the firing of a neuron with an Exponential distribution in Example 1.6.

A probability model is often a collection of probability distributions. For example, we modelled the number of cancers at the Slater school with the Bin$(145, p)$ model. There is one probability distribution for each value of $p \in [0, 1]$. We say the model is a collection of distributions indexed by the parameter p. We study the data to learn about p. More specifically, we learn which values of p describe the data well and which values describe the data poorly.

Similarly, we modelled the emergence of seedlings with a Poi(λ) model. There is one probability distribution for each value of $\lambda \in (0, \infty)$ We say the model is a collection of distributions indexed by the parameter λ. We study the data to learn about λ. More specifically, we learn which values of λ describe the data well and which ones don't.

3.2 Regression Models

A common problem in statistics is to study how the distribution of one quantity depends on another. For example, how does a child's height depend on its parents' heights? How does the firing rate of a rat's neuron depend on what the rat is tasting? In statistics, such problems are called *regression* problems. It is common to use the symbol Y for the thing whose distribution is being studied and X for the things that may affect the distribution of Y. In our examples, Y would be the child's height or the neuron's firing rate while X would be the parent's height or the tastant. (If we accounted for both parents' height then X would be a vector (X_1, X_2) containing both heights.) We use notation like

$$Y \sim X \tag{3.1}$$

to mean that the distribution of Y may depend on the value of X. (3.1) does not mean that Y is completely determined by X, only that its distribution varies with X. For instance, in Examples 2.1 and 2.2, Y is the calorie content of a hot dog and X is its type: Poultry, Beef, or Meat. When X = Poultry then Y has a distribution centered around 125, but when X = Beef or X = Meat then Y has a distribution centered around 160.

3.2.1 Graphs as Regression Models

Displaying data graphically is often a good way of describing a regression model. Figure 2.12 suggests

$$\text{Height} \sim \text{Sex} \tag{3.2}$$

for parents in Galton's data. Figure 2.4 suggests that the distribution of ocean temperatures depends on both latitude and longitude. We would write

$$\text{temperature} \sim \text{latitude} + \text{longitude}. \tag{3.3}$$

Equations (3.1), (3.2), and (3.3) are examples of a modelling language that statisticians use. Even though we earlier used "$\sim$" to mean "is distributed as," in the modelling language we use "$\sim$" to mean "has a distribution that depends on". Similarly, in (3.3) "+" does not mean addition. It means that the distribution of temperature depends on both latitude and longitude. It happens sometimes in mathematics, statistics, and life that we use the same word or symbol to mean different things in different contexts. The reader should be able to figure out the meaning from the context.

Figure 3.1 shows child's height vs. parent's height in Galton's data with separate panels for fathers and mothers. It suggests models like

$$\text{child's height} \sim \text{father's height}$$

and

$$\text{child's height} \sim \text{mother's height}$$

We should also consider the model

$$\text{child's height} \sim \text{father's height} + \text{mother's height}$$

even though we haven't yet made a figure showing how a child's height depends on both parents' heights together.

3.2.2 Linear Regression Models

In each panel of Figure 3.1 the points are scattered around a line. To keep things simple we'll concentrate on the relationship between a child's height and its father's height for now and save the mothers for the Exercises. Let y be a child's height and x be its father's height. For each value of x, the points in Figure 3.1 are scattered

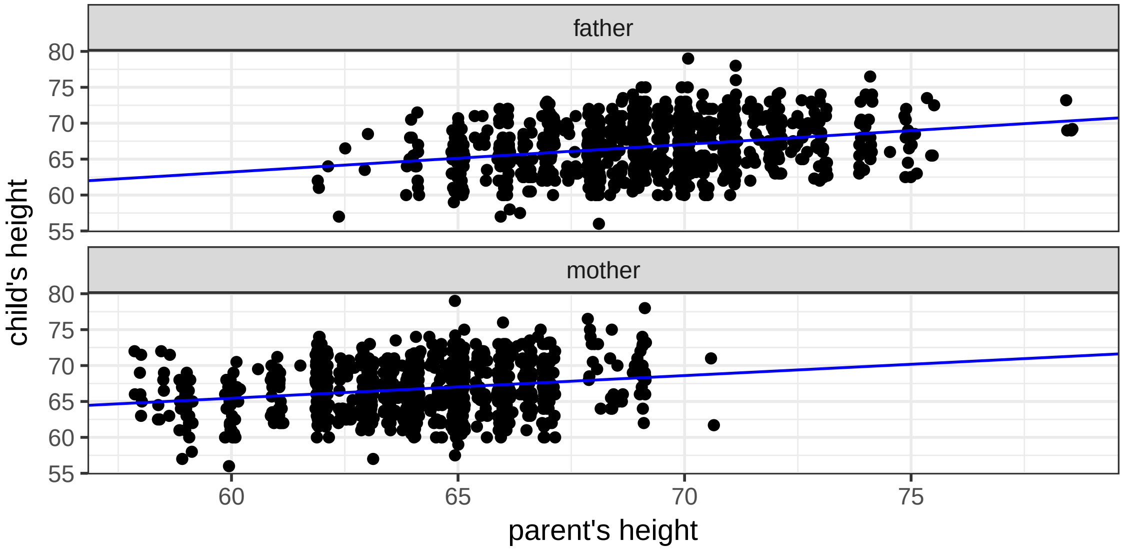


Figure 3.1: Child's height vs. parent's height in Galton's data. Separate panels for mothers and fathers. Points are jittered horizontally to avoid overplotting. The blue lines illustrate the nearly linear trend. Each child appears twice: once in the fathers' panel and once in the mothers'.

roughly symmetrically above and below the line and denser near the line than farther away. Those observations suggest a Normal distribution, so we consider a model like

$$y \sim \mathrm{N}(\mu(x), \sigma(x)) \tag{3.4}$$

where $\mu(x)$ and $\sigma(x)$ indicate that the mean and standard deviation of the Normal distribution may depend on the father's height x. In fact, Figure 3.1 has about the same amount of scatter around the line for all values of x so we can simplify (3.4) to

$$y \sim \mathrm{N}(\mu(x), \sigma) \tag{3.5}$$

in which the standard deviation does not depend on x. Exercise 9 examines this simplification more carefully. The linear trend in Figure 3.1 suggests that $\mu(x)$ is roughly a linear function of x. I.e.,

$$\mu(x) = \beta_0 + \beta_1 x \tag{3.6}$$

where β_0 is the intercept and β_1 is the slope. Together, (β_0, β_1) are called the *coefficients* and are unknown parameters. The data tell us which values of the parameters are more or less plausible. The line in the fathers' panel in Figure 3.1 has intercept $\beta_0 = 40.14$ and slope $\beta_1 = 0.385$. You'll find the intercept and slope of the mothers' line in Exercise 7.

Combining (3.5) and (3.6) gives

$$y \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma). \tag{3.7}$$

We could rewrite (3.7) as

$$y = \beta_0 + \beta_1 x + \mathit{residual}. \tag{3.8}$$

The *residual* is there because a child's height is not exactly a linear function of its father's height; i.e. the points in Figure 3.1 do not fall exactly on the line. In (3.8), β_0 and β_1 are the same for every child but x, y, and *residual* vary from child to child. More thorough notation would replace (3.8) with $y_i = \beta_0 + \beta_1 x_i + r_i$ where the subscript i indicates the i'th child. The Normal distribution in (3.7) means the r_i's have a Normal distribution.

Model (3.7) is not yet a complete model for all the childrens' heights because it does not say how the 934 heights relate to each other. We're going to model them as independent of each other so

$$\begin{aligned} p_{\beta_0,\beta_1,\sigma}(y_1, \dots, y_{934}) &= p_{\beta_0,\beta_1,\sigma}(y_1) \times \cdots \times p_{\beta_0,\beta_1,\sigma}(y_{934}) \qquad \text{(Why?)} \\ &= \prod_{i=1}^{934} p_{\beta_0,\beta_1,\sigma}(y_i) \qquad \text{(math notation)} \end{aligned} \tag{3.9}$$

Modelling the y_i's as independent isn't perfectly accurate. For one thing, children from the same family are likely to have similar heights in ways (3.9) doesn't account for: they have the same mother; they have similar nutrition; and so on. Nevertheless, (3.9) is a good enough model for our purposes.

Equations (3.7) and (3.9) are a complete model for how a child's height depends on its father's height. They have three parameters: β_0, β_1, and σ. We want to compare how well different values of $(\beta_0, \beta_1, \sigma)$ describe the data. We're especially interested in β_1 because if a positive value of β_1 describes the data much better than $\beta_1 = 0$ that's evidence that height can be inherited. To compare values we'll need to evaluate the likelihood function $\ell(\beta_0, \beta_1, \sigma) \equiv p_{\beta_0,\beta_1,\sigma}(y_1, \ldots, y_{934}) = \prod_{i=1}^{934} p_{\beta_0,\beta_1,\sigma}(y_i)$. We'll also want to find the m.l.e. $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}) \equiv \operatorname{argmax}_{\beta_0,\beta_1,\sigma} \ell(\beta_0, \beta_1, \sigma)$.

Even if a positive value of β_1 describes the data much better than $\beta_1 = 0$, that's not necessarily evidence for heritibility of height because there are other possible reasons why β_1 might be positive. For example, height might be related to nutrition, nutrition might be related to wealth, and parents and children might be rich or poor together. Or maybe tall fathers tend to marry tall mothers and the child's height is inherited through its mother.

One aspect of good scientific thinking is being alert to alternate explanations of observed phenomena and not settling for just the first or easiest explanation we find. Once we are aware of alternate explanations we can either examine the data in alternate ways or conduct experiments to try to rule out some of the alternatives.

Equation (3.7) together with the assumption of independence in (3.9) are called a Normal linear model. R has a command lm for working with linear models. We write

```
> lm1 <- lm ( childHeight ~ father, data = GaltonFamilies )
```

to tell R to summarize a linear model of childHeight as a function of father in the dataframe GaltonFamilies and to call the summary lm1. R also has commands for examining aspects of the summary. One of the most important is the coefficients command.

```
> coefficients ( lm1 )
(Intercept)      father
  40.139295    0.384505
```

gives the m.l.e.'s $\hat{\beta}_0 \approx 40.14$ and $\hat{\beta}_1 \approx 0.385$.

Figure 3.1 was produced with the following R commands.

```
library ( ggplot2 )
library ( HistData )
library ( reshape2 )
...
Galton.m <- melt ( GaltonFamilies,
                   id.vars = "childHeight",
                   measure.vars = c ( "father", "mother" ),
                   variable.name = "parent",
                   value.name = "parentHeight"
                 )

lm1 <- lm ( childHeight ˜ father, data = GaltonFamilies )
lm2 <- lm ( childHeight ˜ mother, data = GaltonFamilies )

galtlines <- data.frame ( intercept = c ( coefficients(lm1)[1], coefficients(lm2)[1] ),
                          slope = c ( coefficients(lm1)[2], coefficients(lm2)[2] ),
                          parent = c ( "father", "mother" )
                        )
pdf ( "GaltonChild_v_Parent.pdf", height = 3, width = 6 )
p <- ggplot ( Galton.m, aes ( x = parentHeight, y = childHeight ) )
p + geom_jitter ( width = .15, height = 0 ) +
    xlab ( "parent’s height" ) +
    ylab ( "child’s height" ) +
    facet_wrap ( ˜ parent, ncol = 1 ) +
    geom_abline ( data = galtlines,
                  mapping = aes ( intercept = intercept, slope = slope ),
                  color = "blue"
                ) +
    theme_bw()
dev.off()
```

If you don’t remember what `melt` does either compare GaltonFamilies to Galton.m or look up the command. The `geom_abline` command plots a line; you have to tell it the slope and intercept. That’s how we got the lines in Figure 3.1. We’re using `facet_wrap` and we want different lines in different facets so we create a dataframe `galtlines` that contains one column of intercepts, one column of slopes, and one column that says to which facets they apply. We take the coefficients from `lm1` and `lm2` to use the m.l.e.’s $(\hat\beta_0, \hat\beta_1)$ and we use the same names for facets as in `Galton.m`.

Finding the m.l.e. $\hat\sigma$ is a bit harder and uses the formula on page 79 for the m.l.e. of the Normal distribution. Equation (3.7) can be rewritten as

$$y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma). \tag{3.10}$$

for $i = 1, \dots, 934$. (3.10) represents 934 y_i’s, each with a different Normal distribution. (3.10) can be rewritten as

$$(y_i - (\beta_0 + \beta_1 x_i)) \sim \mathrm{N}(0, \sigma) \quad \text{(Why?)} \tag{3.11}$$

which still represents 934 quantities, but they all have the same $\mathrm{N}(0, \sigma)$ distribution.

To find $\hat{\sigma}$ we plug $(\hat{\beta}_0, \hat{\beta}_1)$ into (3.11) and use the formula on page 79:

$$\begin{aligned}\hat{\sigma} &= \sqrt{\frac{(y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1))^2 + \cdots + (y_{934} - (\hat{\beta}_0 + \hat{\beta}_1 x_{934}))^2}{934}} \\ &= \sqrt{\frac{\sum_{i=1}^{934}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{934}}\end{aligned}$$

The quantities $(\hat{\beta}_0 + \hat{\beta}_1 x_i)$ are called *fitted values*; the quantities $(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$ are called *residuals*. There are 934 fitted values and 934 residuals. You can easily extract them from `lm1` as shown in the following R snippet.

```
> head ( fitted ( lm1 ) )
       1        2        3        4        5        6
70.32294 70.32294 70.32294 70.32294 69.16942 69.16942

> head ( residuals ( lm1 ) )
        1         2         3         4         5         6
 2.877060 -1.122940 -1.322940 -1.322940  4.330575  3.330575
```

We can use the residuals to calculate $\hat{\sigma}$:

```
> sighat <- sqrt ( mean ( residuals(lm1)^2 ) )
> sighat
[1] 3.448416
```

We find $\hat{\sigma} \approx 3.45$.

R also has a `summary` command to extract information from `lm1`:

```
> summary ( lm1 )

Call:
lm(formula = childHeight ~ father, data = GaltonFamilies)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2856  -2.7374  -0.2275   2.6763  11.9454

Coefficients:
```

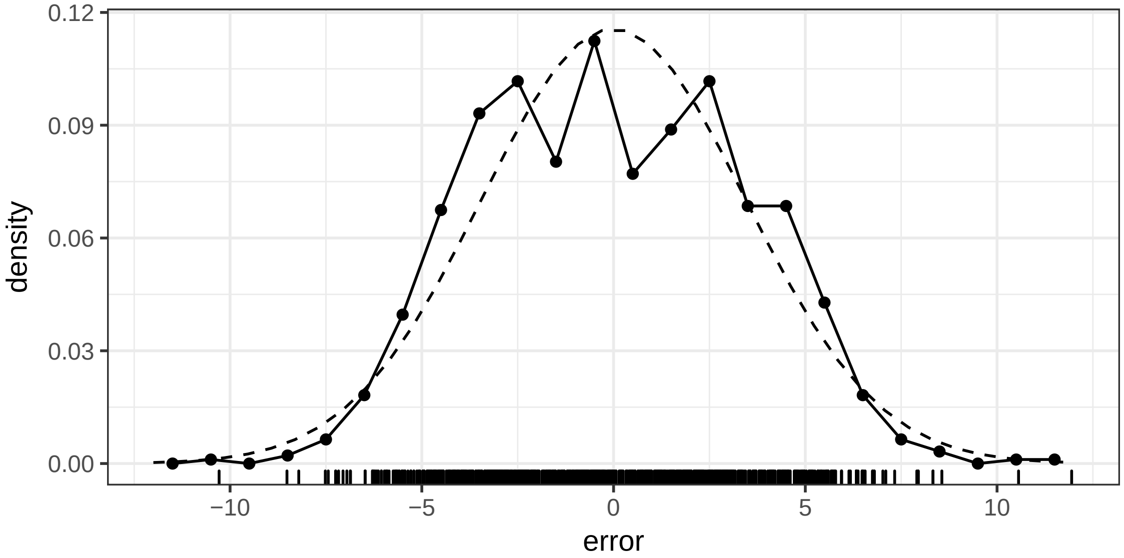


Figure 3.2: Residuals from `lm1` with density estimated from 1-inch bins (solid line) and the N(0, 3.45) density (dashed line).

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.13929    3.15991  12.703   <2e-16 ***
father       0.38451    0.04564   8.425   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.452 on 932 degrees of freedom
Multiple R-squared:  0.07078,Adjusted R-squared:  0.06978
F-statistic: 70.99 on 1 and 932 DF,  p-value: < 2.2e-16
```

Most of the summary does not concern us, but $\hat{\beta}_0$ and $\hat{\beta}_1$ are in the `Estimate` column of the `Coefficients` section. The `Residual standard error`, 3.452, is nearly $\hat{\sigma}$ but R, like most other statistical software, for reasons that don't concern us, divides $\sum(r_i)^2$ by the `degrees of freedom`, 932, instead of by the number of observations 934. So another way to calculate $\hat{\sigma}$ is to start with the residual standard error, multiply by the degrees of freedom, and divide by the number of observations.

Starting with (3.4) and (3.5) we used visual inspection of Figure 3.1 to justify modelling the residuals with a Normal distribution. It's a good idea to check whether Normal is a good model for the residuals. Figure 3.2 plots the residuals along with a density estimated from 1-inch bins as in Section 1.4 and the $N(0, 3.45)$ density. Normal does appear to be a good model.

Figure 3.2 was produced with the following R commands.

```
lm1 <- lm ( childHeight ~ father, data = GaltonFamilies )
tmp1in <- data.frame ( error = seq ( -11.5, 11.5, by = 1 ),
                       den = as.vector ( table ( cut ( residuals ( lm1 ),
                                                       breaks = seq ( -12, 12, by = 1 )
                                                     )
                                               )
                                       ) / 934
                     )
df <- data.frame ( x = seq ( -12, 12, length = 40 ) )
df$den <- dnorm ( df$x, mean = 0, sd = 3.45 )
# density with approximating Normal density
pdf ( "lm1_resid.pdf", height = 3, width = 6 )
p <- ggplot ( tmp1in, aes ( x = error, y = den ) )
p + geom_point() +
    geom_line() +
    geom_rug ( data = data.frame ( e = residuals ( lm1 ) ), mapping = aes ( x = e ),
               inherit.aes = FALSE ) +
    geom_line ( data = df, aes ( x = x, y = den ), lty = 2 ) +
    ylab ( "density" ) +
    theme_bw()
dev.off()
```

The main idea is that we want to plot the density of the residuals similarly to how we plotted the density of parents' heights in Example 1.4.

3.2.3 Profile Likelihood

One use of models is to compare how well different parameter values describe the data. For instance, in Example 1.2, the Slater School, we were interested in whether we could describe the number of cancers about as well with Pr[cancer] $\approx .03$, the national average, as with higher values of Pr[cancer]. If so, there is little evidence that proximity to high-voltage transmission lines promotes cancer. Such questions are answered *via* the likelihood function. Figure 2.7 shows the likelihood function for Example 1.2 and that $\Pr_{p=.03}$[8 cancers] is only a little bit smaller than $\Pr_{p=.055}$[8 cancers].

Much less was known about heredity in Galton's day than in ours so a question of scientific interest in his day would have been *"Is it plausible that β_1 in* (3.6) *is 0?"* If so, then the data do not provide much evidence for heretibility of height. On the other hand, if models with $\beta_1 > 0$ describe the data much better than models with $\beta_1 = 0$, then there is evidence that at least some part of a child's height is inherited from its father. We'll investigate that question *via* the likelihood function to see whether we can achieve nearly as high a likelihood with $\beta_1 = 0$ as with $\beta_1 > 0$.

So far we've worked with the likelihood function only to find the m.l.e.'s. Now we need to explore the likelihood function for other values of $(\beta_0, \beta_1, \sigma)$.

Let $\ell(\beta_0, \beta_1, \sigma)$ stand for the likelihood function.

$$\begin{aligned}
\ell(\beta_0, \beta_1, \sigma) &\equiv p_{\beta_0,\beta_1,\sigma}(y_1, \ldots, y_n) \\
&= p_{\beta_0,\beta_1,\sigma}(y_1) \times \cdots \times p_{\beta_0,\beta_1,\sigma}(y_n) \quad \text{(Why?)} \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{y_1-(\beta_0+\beta_1 x_1)}{2\sigma}\right)^2} \times \cdots \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{y_n-(\beta_0+\beta_1 x_n)}{2\sigma}\right)^2} \quad \text{(Why?)} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{2\sigma}\right)^2} \quad \text{(math notation)}
\end{aligned}$$

which, for a given $(\beta_0, \beta_1, \sigma)$, we could calculate in R with

```
beta0 <- 40.14
beta1 <- .385
prod ( dnorm ( GaltonFamilies$childHeight, mean = beta0 + beta1*GaltonFamilies$father, sd = sig ) )
```

In the previous expression, `GaltonFamilies$childHeight` and `beta0 + beta1*GaltonFamilies$father` are each a vector of length 934. So `dnorm` returns a vector of length 934. Each element of the vector is the `dnorm` for one child. Most values in that vector are less than 1 and multiplying them together with `prod` gives a result close to 0. In fact, to within the accuracy of my computer, the result is 0 for any combination of $(\beta_0, \beta_1, \sigma)$. That result is useless for comparing many values of $(\beta_0, \beta_1, \sigma)$ because we'd just be comparing 0 to 0 to 0, and so on.

Fortunately there's a trick to perform a more accurate calculation: calculate the log of the likelihood function instead.

$$\begin{aligned}
\log \ell(\beta_0, \beta_1, \sigma) &\equiv \log p_{\beta_0,\beta_1,\sigma}(y_1, \ldots, y_n) \\
&= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{2\sigma}\right)^2} \\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{2\sigma}\right)^2}
\end{aligned}$$

In R that would be

```
sum ( dnorm ( GaltonFamilies$childHeight,
              mean = beta0 + beta1*GaltonFamilies$father,
              sd = sig,
              log = TRUE
           )
    )
```

That calculation is much more accurate. If desired we can then calculate $\ell(\beta_0, \beta_1, \sigma) = e^{\log \ell(\beta_0, \beta_1, \sigma)}$.

However, we don't want to compare all values of $(\beta_0, \beta_1, \sigma)$; we want to see how well different values of β_1 can describe Galton's data. To do that, we match each value of β_1 to its own best values of β_0 and σ. That is, for each β_1 we find $\hat{\beta}_0(\beta_1)$

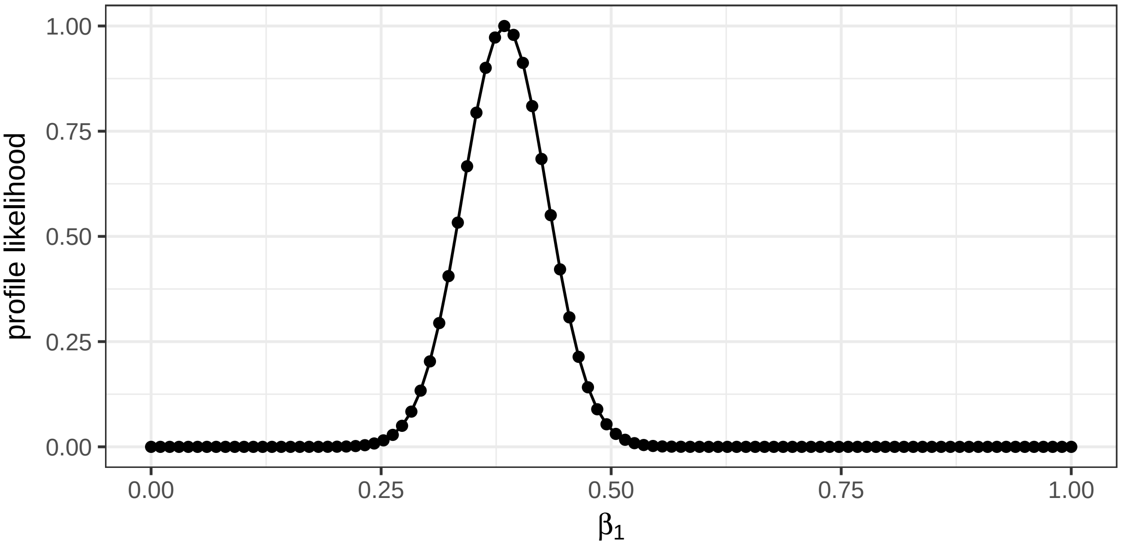


Figure 3.3: Profile likelihood for β_1 in (3.9).

and $\hat{\sigma}(\beta_1)$ where the notation means these are the best values of β_0 and σ to go with this value of β_1. In symbols,

$$(\hat{\beta}_0(\beta_1), \hat{\sigma}(\beta_1)) = \text{argmax}_{\beta_0,\sigma}\, \ell(\beta_0, \beta_1, \sigma).$$

Then to say how well each β_1 can describe the data we define the *profile likelihood* function

$$plik(\beta_1) \equiv \max_{\beta_0,\sigma} \ell(\beta_0, \beta_1, \sigma) = \ell(\hat{\beta}_0(\beta_1), \beta_1, \hat{\sigma}(\beta_1)).$$

$plik(\beta_1)$ is a function of β_1 alone, not of (β_0, σ). $plik(\beta_1)$ is the value of $\ell(\beta_0, \beta_1, \sigma)$ when β_1 is matched with its own best values of β_0 and σ.

Figure 3.3 plots $plik(\beta_1)$. We chose a sequence of 100 values of β_1 from 0 to 1. For each value of β_1 we found $\hat{\beta}_0(\beta_1)$ and $\hat{\sigma}(\beta_1)$. Then we found the profile likelihood $plik(\beta_1) = \ell(\hat{\beta}_0(\beta_1), \beta_1, \hat{\sigma}(\beta_1))$. Then we rescaled the profile likelihood values to have a maximum of 1 and plotted them. The figure shows that model (3.9) can describe the data much better when it uses values of $\beta_1 \in (.25, .50)$ than when it uses values $\beta_1 < .25$ or $\beta_1 > .5$. Thus, we have evidence that height is heritable or else there must be some other explanation why $plik(0) \ll plik(.37)$.

Figure 3.3 was produced with the following R code.

```
nvals <- 100 # how many values of beta_1 to try
# a data frame with values of beta_1 at which to evaluate the profile likelihood.
# leave place holders for beta_0, sigma, logplik, and plik.
plik <- data.frame ( b1 = seq ( 0, 1, length = nvals ),
                     b0hat   = NA,
                     sighat  = NA,
                     logplik = NA,
                     plik    = NA
                   )
for ( i in 1:nvals ) {
  tmp <- plik$b1[i]*GaltonFamilies$father
  plik$b0hat[i] <- mean ( GaltonFamilies$childHeight - tmp )
  pred <- plik$b0hat[i] + tmp
  resids <- GaltonFamilies$childHeight - pred
  plik$sighat[i] <- sqrt ( mean ( resids^2 ) )
  plik$logplik[i] <- sum ( dnorm ( GaltonFamilies$childHeight,
                                   mean = pred,
                                   sd = plik$sighat[i],
                                   log = TRUE
                                 )
                         ) # the logarithm trick

}
plik$logplik <- plik$logplik - max  ( plik$logplik )
plik$plik <- exp ( plik$logplik )
pdf ( "plik.pdf", height = 3, width = 6 )
p <- ggplot ( plik, aes ( x = b1, y = plik ) )
p + geom_point() +
    geom_line() +
    xlab ( expression ( beta[1] ) ) +
    ylab ( "profile likelihood" ) +
    theme_bw()
dev.off()
```

We begin by creating a dataframe `plik` with columns for β_1, $\hat{\beta}_0$, $\hat{\sigma}$, $\log(plik)$, and $plik$. We fill in just the values of β_1 for now and leave the other columns blank. The reason for $\log(plik)$ was explained in an earlier note. For each value of β_1, i.e. for each row of `plik`, we have to calculate several things. That's what the loop `for ( i in 1:nvals ) { ... }` does. Inside the loop we calculate the following things.

1. `tmp <- plik$b1[i]*GaltonFamilies$father`. This line calculates $\beta_1 x$. `b1[i]` is a number and `GaltonFamilies$father` is a vector with one entry for each child. So `tmp` is a vector with one entry for each child. The entry for the j'th child is `b1[i]*GaltonFamilies$father[j]`. If you don't see what R is doing you can set `i <- 1`, then run `tmp <- plik$b1[i]*GaltonFamilies$father` and examine the result. Try it again with other values of i until you understand it. The next step uses `tmp`.

2. `plik$b0hat[i] <- mean ( GaltonFamilies$childHeight - tmp )`. We saw earlier that our model says $y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma)$. Another way to say the same thing is $y_i - \beta_1 x_i \sim \mathrm{N}(\beta_0, \sigma)$. The advantage of writing it this way is that we now have a collection of objects $y_i - \beta_1 x_i$ that all have the same $\mathrm{N}(\beta_0, \sigma)$ distribution and we can use the formula on page 79 to find the m.l.e. $\hat{\beta}_0(\beta_1)$.

3. We also use page 79 to find the m.l.e. $\hat{\sigma}(\beta_1)$. To do that we have to

   (a) subtract $\hat{\beta}_0(\beta_1)$ from each object. That's what `pred <- plik$b0hat[i] + tmp; resids <- GaltonFamilies$childHeight - pred` does. `pred` stands for "predicted" and `resids` stands for "residuals."

(b) apply the formula on page 79. That's what
`plik$sighat[i] <- sqrt ( mean ( resids^2 ) )` does.

4. Now that we have $\hat{\beta}_0(\beta_1)$ and $\hat{\sigma}(\beta_1)$ we can calculate $p_{\hat{\beta}_0(\beta_1),\beta_1,\hat{\sigma}(\beta_1)}[y_i]$ for each child using the logarithm trick we saw in an earlier note.

However, there's still a problem. The values of $p_{\hat{\beta}_0(\beta_1),\beta_1,\hat{\sigma}(\beta_1)}[y_i]$ are large negative numbers and when we convert back from logarithms we get $plik(\beta_1) = e^{\text{-(large number)}} = 0$, as accurately as my computer can calculate. Fortunately, there's a solution. We say `plik$logplik <- plik$logplik - max ( plik$logplik )`, which changes the largest *plik* from less than about -2000 to 0. (With likelihood functions, multiplicative constants don't matter so, with loglikelihoods, additive constants don't matter.) Finally we can say `plik$plik <- exp ( plik$logplik )` and have something useful.

3.2.4 Model Sequences

Instead of describing a data set with a single model it may be useful to describe it with a sequence of models. Early models in the sequence describe big important features of the data while later models describe smaller less important features. We illustrate with a data set on soil respiration.

Example 3.1 (Soil Respiration)
Soil respiration refers to the exchange of carbon between the soil and the atmosphere. This example draws on data analyzed by Giasson et al. (2013) of over 100,000 measurements of soil respiration in the Harvard Forest, in north-central Massachusetts. To quote from Giasson et al.:

> "[S]oil respiration (R_s) [is] the sum of belowground autotrophic (roots and associated mycorrhizae) and heterotrophic (mainly microbes, microfauna, and mesofauna) respiration. Estimates of global R_s range from 68 to 98 Gt C yr^{-1} (Raich and Schlesinger, 1992; Schlesinger and Andrews, 2000; Bond-Lamberty and Thomson, 2010), or about two-thirds of all of the C emitted to the atmosphere by terrestrial ecosystems. The amount of C emitted through R_s is $\sim$10 times more than that released through fossil fuel combustion and cement manufacturing (IPCC, 2007; Peters et al., 2012), although, for the most part, R_s is closely coupled to a large photosynthetic uptake, leading to a much smaller net C exchange with the atmosphere (Schlesinger and Andrews, 2000)."

Because R_s is generated by metabolic processes it is affected by soil temperature T_s, and much interest centers on the relationship between T_s and R_s. Figure 3.4 shows that modeling $\log(R_s)$ as a roughly linear function of T_s is a good description of most of the data, with the exception of a small fraction of data points with very low R_s or high T_s. A good first description of the data is

Model A $\log(R_s)$ is approximately a linear function of T_s with slope about 0.127 or so and intercept about -0.7809 or so with Normally distributed residuals whose standard deviation is about 0.67 or so.

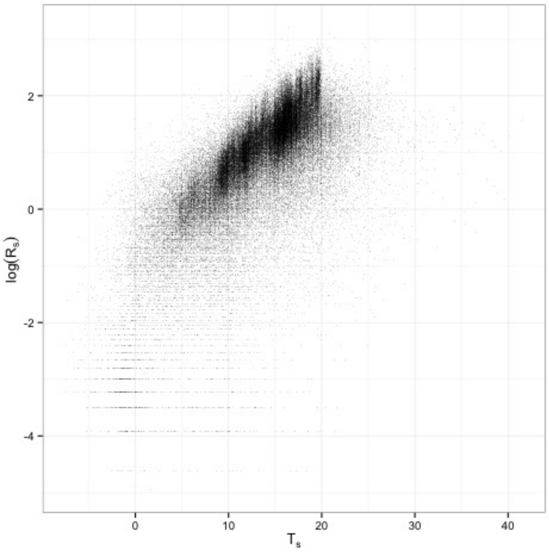


Figure 3.4: $\log(R_s)$ versus T_s in the Harvard Forest. Over 100,000 data points.

There are multiple enhancements of **Model A** that describe the data more accurately. We show several below.

Model B We could propose a nonlinear regression function such as, for example, a quadratic. Figure 3.4 doesn't suggest much curvature except for the points with low R_s or high T_s but it might be of interest to quantify how much better **Model B** describes the data than **Model A**.

Model C Figure 3.5 shows that the relationship between R_s and T_s varies slightly by the type of site, so we could propose a model in which different forest types have different slopes and intercepts. It might be of interest to quantify how much better **Model C** describes the data than **Model A** or **Model B**.

Model D Figure 3.6 shows that residuals from **Model A** are asymmetric, non-Normal, and vary by type of site, so we could propose a more accurate model for residuals. It might be of interest to quantify how much better **Model D** describes the data than **Model A**, **Model B**, or **Model C**.

Model E Giasson et al. (2013) further explain measurement of R_s:

> "R_s was measured in fixed locations on a given sampling day, generally where PVC or aluminum collars had been inserted and left in the soil, usually for the duration of the study. ... Four methods were used to measure R_s, in order of increasing measurement frequency: (1) soda-lime systems where pellets were left beneath a closed chamber for 24 hours to absorb CO_2 emitted from the soil, (2) static chamber systems where a chamber was placed on each collar and headspace air samples were taken at fixed intervals over 15 to 30 minutes and subsequently analyzed with an infrared gas analyzer (IRGA) or a gas chromatograph, (3) dynamic chamber systems in which a chamber was placed on each collar, chamber air was circulated to and from a portable IRGA system, and the rate of increase in CO2 concentration was measured in situ for a period of five minutes, and (4) automated chamber systems (herein autochambers), in which a datalogger-controlled system closed one chamber at a time and circulated the headspace air through an IRGA."

Figure 3.7 shows where the R_s measurements were taken. Each point in Figure 3.7 is a site where at least one collar was permanently installed and used to measure R_s on at least one occasion. In fact, there is much collar-to-collar variability in the response of R_s to T_s, and that was the main focus of Giasson et al. (2013). It is not our focus here, but we could propose a model in which the parameters differ from collar to collar. It might be of interest to quantify how much better **Model E** describes the data than **Model A**, **Model B**, **Model C**, or **Model D**.

The point of Example 3.1 is that we don't need to settle on just one model. Instead, it is often helpful to create a sequence of models. Early models in the sequence may be overly simple, but can describe the most salient features of the data. Further analysis shows the ways in which early models are imperfect and how they can be improved to describe less salient features, at the expense of making the models more complex. Different models in the sequence might be useful for different purposes.

3.2.5 Model Checking

It is generally a good idea to check whether a model accurately describes the data it's modelling. Anscombe (1973) created a data set to illustrate why. The data come with R; you can type `anscombe` to see them.

```
> anscombe
   x1 x2 x3 x4    y1   y2    y3    y4
1  10 10 10  8  8.04 9.14  7.46  6.58
2   8  8  8  8  6.95 8.14  6.77  5.76
3  13 13 13  8  7.58 8.74 12.74  7.71
4   9  9  9  8  8.81 8.77  7.11  8.84
5  11 11 11  8  8.33 9.26  7.81  8.47
6  14 14 14  8  9.96 8.10  8.84  7.04
7   6  6  6  8  7.24 6.13  6.08  5.25
8   4  4  4 19  4.26 3.10  5.39 12.50
9  12 12 12  8 10.84 9.13  8.15  5.56
10  7  7  7  8  4.82 7.26  6.42  7.91
11  5  5  5  8  5.68 4.74  5.73  6.89
```

There are really four data sets — (`x1, y1`), (`x2, y2`), (`x3, y3`), and (`x4, y4`) — each with 11 observations $(x_1, y_1), \ldots, (x_{11}, y_{11})$. We are to model $y1 \sim x1$; $y2 \sim x2$; $y3 \sim x3$; and $y4 \sim x4$. To fit linear models we would say

```
lm1 <- lm ( y1 ~ x1, data = anscombe )
lm2 <- lm ( y2 ~ x2, data = anscombe )
lm3 <- lm ( y3 ~ x3, data = anscombe )
lm4 <- lm ( y4 ~ x4, data = anscombe )
```

Then to examine the coefficients we would say

```
> coefficients ( lm1 )
(Intercept)          x1
  3.0000909   0.5000909
> coefficients ( lm2 )
(Intercept)          x2
   3.000909    0.500000
> coefficients ( lm3 )
(Intercept)          x3
  3.0024545   0.4997273
> coefficients ( lm4 )
```

```
(Intercept)          x4
  3.0017273   0.4999091
```

We see that all four models have about the same m.l.e.'s for the intercept and slope. We can check the four $\hat{\sigma}$'s too.

```
> sqrt ( mean ( residuals ( lm1 )^2 ) )
[1] 1.11855
> sqrt ( mean ( residuals ( lm2 )^2 ) )
[1] 1.119102
> sqrt ( mean ( residuals ( lm3 )^2 ) )
[1] 1.118286
> sqrt ( mean ( residuals ( lm4 )^2 ) )
[1] 1.117729
```

They're all about the same. To assess how well each model fits its data we'll calculate $p_{\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}}(y_1,\dots,y_{11})$ for each data set. Though we haven't said so, we're treating the y_i's as independent of each other and the residuals as Normal so

```
> sighat <- 1.12
> sum ( dnorm ( anscombe$y1, mean = fitted(lm1), sd = sighat, log = TRUE ) )
[1] -16.84071
> sum ( dnorm ( anscombe$y2, mean = fitted(lm2), sd = sighat, log = TRUE ) )
[1] -16.84613
> sum ( dnorm ( anscombe$y3, mean = fitted(lm3), sd = sighat, log = TRUE ) )
[1] -16.83812
> sum ( dnorm ( anscombe$y4, mean = fitted(lm4), sd = sighat, log = TRUE ) )
[1] -16.83265
```

does the calculations and shows the answer is about -16.84 for all four data sets. It seems the same linear model describes all four data sets about equally well.

Now we plot the data. I want to put each data set in a separate facet of the plot, so I make a new data frame with all the x's in one column, all the y's in another column, and a third column indicating which data set they come from. The R code is

```
anscombe2 <- with ( anscombe,
                    data.frame ( x = c ( x1, x2, x3, x4 ),
                                 y = c ( y1, y2, y3, y4 ),
                                 set = rep ( 1:4, each = 11 )
                               )
                  )
```

The with command tells R that the variables x1, ..., y4 are found in the dataframe anscombe. If you don't see what anscombe2 is, try the commands and examine the result.

Finally we're ready to plot. The result is Figure 3.8. It seems the same linear model does not describe all four data sets equally well. It describes the first data set well but not the second, third, and fourth. There is an apparent contradiction: $p_{\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}}(y_1,\ldots,y_{11})$ is about equal for all four data sets, but the model isn't equally good for all four data sets. The key to resolving the contradiction is to realize the question isn't "How good is the model" but rather "How good is the model compared to alternatives?" In data set 1 there is not a good alternative and the model describes the overall pattern of the points. In data set 2 there is a good alternative: y is a quadratic function of x. In data set 3 a good alternative description is that most of the points fall on line — not the line we have drawn — but one point is different. In data set 4 a good alternative description is that all points but one have the same value of x, so there is very little information about how the distribution of y varies with x.

Figure 3.8 was produced by

```
b0hat <- 3
b1hat <- .5
pdf ( "anscombe.pdf", height=4, width=7 )
p <- ggplot ( anscombe2, aes ( x = x, y = y ) )
p + geom_point() +
    facet_wrap ( ~ set ) +
    geom_abline ( intercept = b0hat, slope = b1hat, color = "blue" ) +
    theme_bw()
dev.off()
```

3.3 Exercises

1. What models are suggested by Figures 2.5 and 2.6?

2. R comes with many preloaded datasets. One is called `Loblolly`. You can learn about it by typing `?Loblolly`. It contains the heights of 14 Loblolly pine trees, measured at ages 3, 5, 10, 15, 20, and 25. Plot height as a function of age and connect the dots for each tree. You might use a command such as
`p <- ggplot ( Loblolly, aes ( x = age, y = height, group = Seed ) ).`
"`group = Seed`" groups data from the same Seed together.

   (a) When modelling the heights of trees, should we account for age?

   (b) Would a linear model be appropriate for height $\sim$ age?

   (c) Do the tall trees at young ages continue to be the tall trees when they're older? Does that match your experience with human characteristics?

   (d) Another data set in R is `Orange`, on the growth of orange trees. Plot the orange tree data. Are the answers to the previous questions different for orange trees than for Loblolly pines?

3. R comes with the dataset `Seatbelts` which gives the monthly totals of car drivers in Great Britain killed or seriously injured Jan 1969 to Dec 1984. Compulsory wearing of seat belts was introduced on 31 Jan 1983. We want to plot the data with `ggplot`. However, `ggplot` works on dataframes and `Seatbelts` is not a dataframe. Fortunately, you can convert `Seatbelts` to a dataframe by a command such as
`mySeatbelts <- as.data.frame ( Seatbelts ).`
Add a column called `month` to `mySeatbelts`. `month` should contain the numbers 1, 2, . . . , for however many months are in the data set.

   (a) The column `drivers` contains the number of drivers' deaths each month. Plot `drivers` vs. `month`.

   (b) Do some months typically have a large number of deaths and other months a low number? Such a pattern when repeated year after year is called *seasonal.*

   (c) Would it make sense to represent the data as a collection of independent observations all coming from the same distribution?

   (d) The column `law` indicates whether the seatbelt law was in effect. Does the pattern seem to shift after the law went into effect? Estimate the effect of the law. Does the total number of deaths seem to change? Does the seasonal pattern seem to change? A visual estimate is good enough.

(e) Would it make sense to represent the data as two sets of independent observations — one from before the law and one from after — with all the data in each set coming from the same distribution?

4. R's `PlantGrowth` dataset contains the weights of plants grown under three different treatments. Plot the data to show whether the distribution of weight varies from treatment to treatment. Which treatment produces the heaviest plants?

5. The dataset `UCBAdmissions` shows admission rates to six large departments at UC Berkeley in 1973 by gender. Use `is.data.frame` to find out whether `UCBAdmissions` is a dataframe. If it's not, use `as.data.frame` to convert it.

   (a) Grouping all six departments together, how many males applied for admission? How many females?

   (b) What fraction of males were admitted? What fraction of females?

   (c) We'll explore the binomial model for admissions in which each applicant has the same probability of admission and applicants are admitted independently of each other. Under this model the number of admissions has a Bin(n, p) distribution where n is the number of applicants and p is the probability of admission.

      i. Find and plot the likelihood function for p.

      ii. Find $\hat{p}$.

   (d) We'll explore a two-binomial model for admissions in which males and females have different probabilities of admission. Call them p_m and p_f.

      i. Find, plot, and compare the likelihood functions for p_m and p_f.

      ii. Find $\hat{p}_m$ and $\hat{p}_f$.

      iii. How much better does the two-binomial model describe the data than the one-binomial model?

      iv. Do there appear to be different admission rates for males and females? Which gender is favored, if any?

   (e) We'll explore models in which each department has its own admission rate.

i. For each department, make a one-binomial model for admissions. Find and plot the likelihood functions for p_A, …, p_F where the subscript indicates the department. Find $\hat{p}_A$, …, $\hat{p}_F$. Do departments seem to have the same admission rates?

ii. For each department, make a two-binomial model for admissions. The parameters will be denoted by symbols like $p_{A,F}$ where the first subscript indicates department and the second indicates gender. For each department plot the two likelihood functions $p_{\cdot,M}$ and $p_{\cdot,F}$ where $\cdot$ is a placeholder for department and M and F indicate gender. Find $\hat{p}_{\cdot,M}$ and $\hat{p}_{\cdot,F}$ for each department.

iii. Which departments seem to favor admitting males and which favor females?

(f) Match the parts of this exercise to the following descriptions: (i) the marginal distribution of admission, (ii) the conditional distribution of admission given gender, (iii) the conditional distribution of admission given department, (iv) the conditional distribution of admission given department and gender.

(g) Which parts of this exercise support or contradict the following statements?

i. There is no gender discrimination in admissions in these six departments.

ii. The university tends to give more money to support graduate students in the departments to which males tend to apply.

iii. Males tend to apply to departments that have high admission rates.

6. R's dataset `cars` contains data on the speed and stopping distance of cars.

(a) Plot distance vs. speed. Is a linear model for dist $\sim$ speed reasonable?

(b) Find the m.l.e.'s of the linear model.

(c) It seems reasonable that a car going 0 mph needs no distance to stop, so it seems reasonable that the linear model should go through the point $(0, 0)$. Write down the model for a line that goes through $(0, 0)$. Find the mle's of its parameters.

(d) Add two lines to the plot: the m.l.e. lines for both linear models. Does one line describe the data much better than the other?

7. Find the intercept and slope of the blue line in the mothers' panel of Figure 3.1. Do mothers or fathers appear to have a greater influence on their children's heights, or are they about the same? Does your answer agree with your understanding of genetics?

8. Make a figure similar to Figure 3.2 but using mothers' heights instead of fathers'.

9. Model (3.5) says that the standard deviation of children's height does not depend on their fathers' height. Fit Model (3.5) by finding $\hat{\beta}_0$ and $\hat{\beta}_1$. Plot the residuals vs. the fathers' height and say whether you see a trend in the spread of the residuals.

10. Figure 3.2 displayed the residuals from Model (3.5). Exercise 10 studies those residuals to see whether we can find a more descriptive model than (3.5).

    (a) Use R's `residuals` command to get the residuals from (3.5).

    (b) Display the residuals in a strip chart as in Figure 2.1H.

    (c) We'll check whether boys and girls tend to be the same height, even after accounting for their fathers' heights. Modify the strip chart in the previous part to use different colors for boys and girls. Modify it again to put the boys' and girls' residuals in different facets. What do you learn?

    (d) We'll fit a model for child's height that accounts for both the father's height and the child's gender. Use R's `lm` command to fit the model `childHeight ~ father + gender` in `GaltonFamilies`. Use R's `coefficients` command to get the coefficients from the model you fit. Does the result surprise you? Does it agree, at least roughly, with your stripcharts?

    (e) We'll see whether a model for a child's height should also account for its mother's height. Plot the residuals from the previous model in two ways. First, put them in a strip chart with color for the mother's height. Then make a plot with mother's height on the x-axis and residuals on the y-axis[1]. What do you see? Would it be useful to include mother's height as a predictor? Fit the model `childHeight ~ father + mother + gender` and print the coefficients. Do they make sense?

[1] This is not the best way to see the effect of mother's height. See the section on added variable plots in Weisberg (2013) for a better way.

The theory of fitting linear models with more than one predictor is beyond the scope of this book. Two good references are Weisberg (2013) and Kaplan (2011).

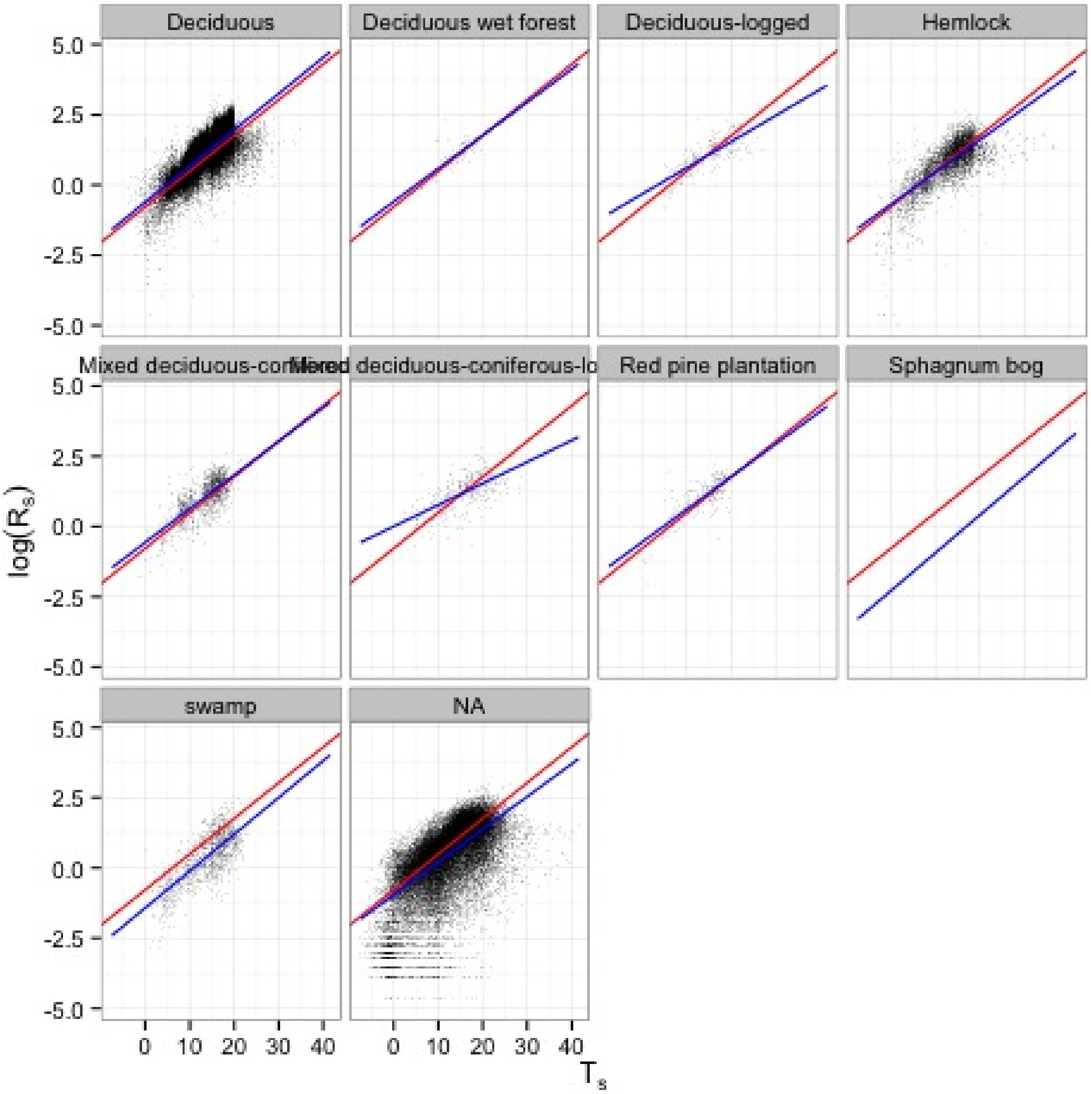


Figure 3.5: $\log(R_s)$ versus T_s in the Harvard Forest by type of site. The red line is the same in each facet and comes from **Model A**. The blue lines are different in each facet and come from **Model C**.

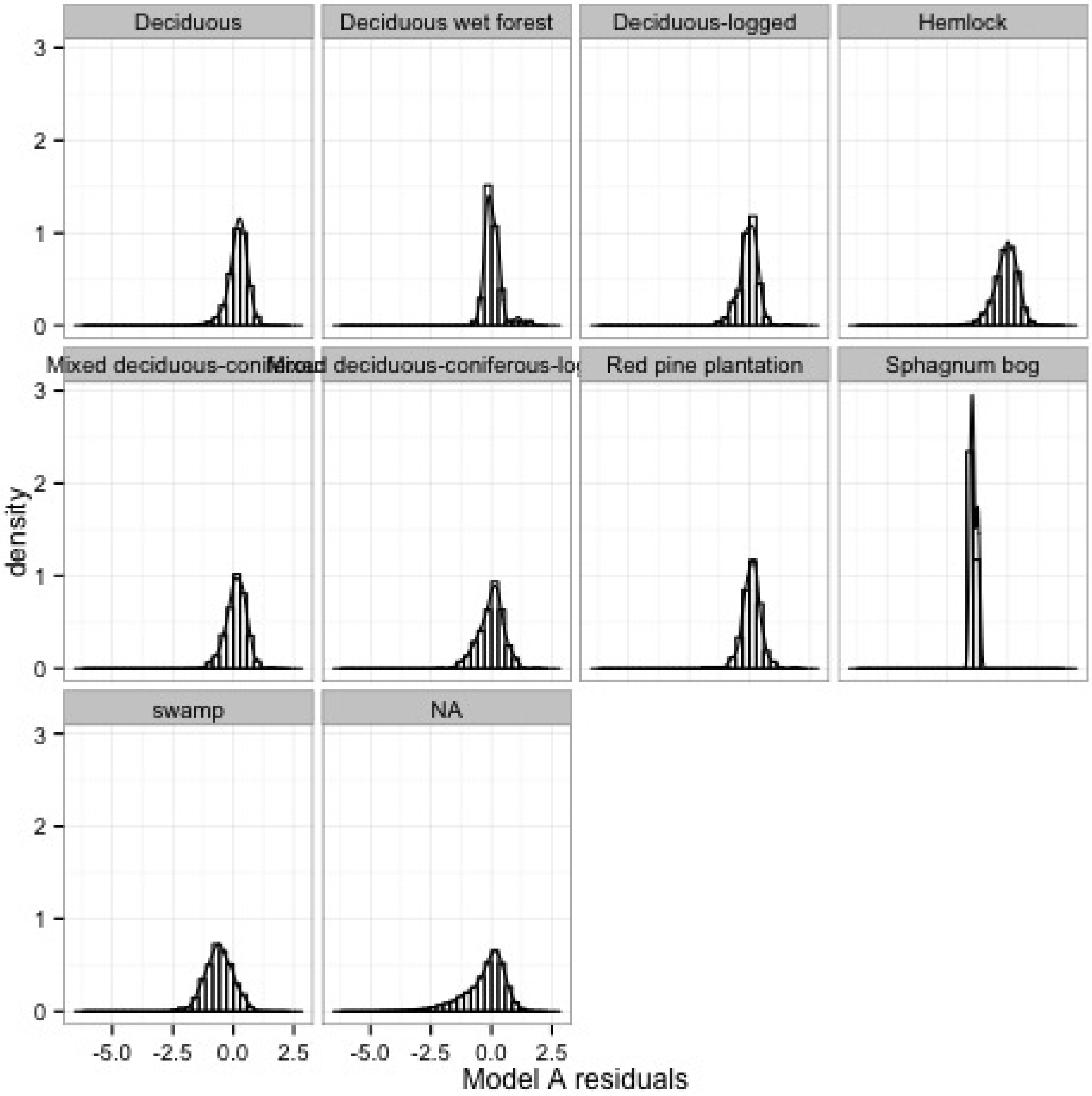


Figure 3.6: residuals from **Model A** by type of site.

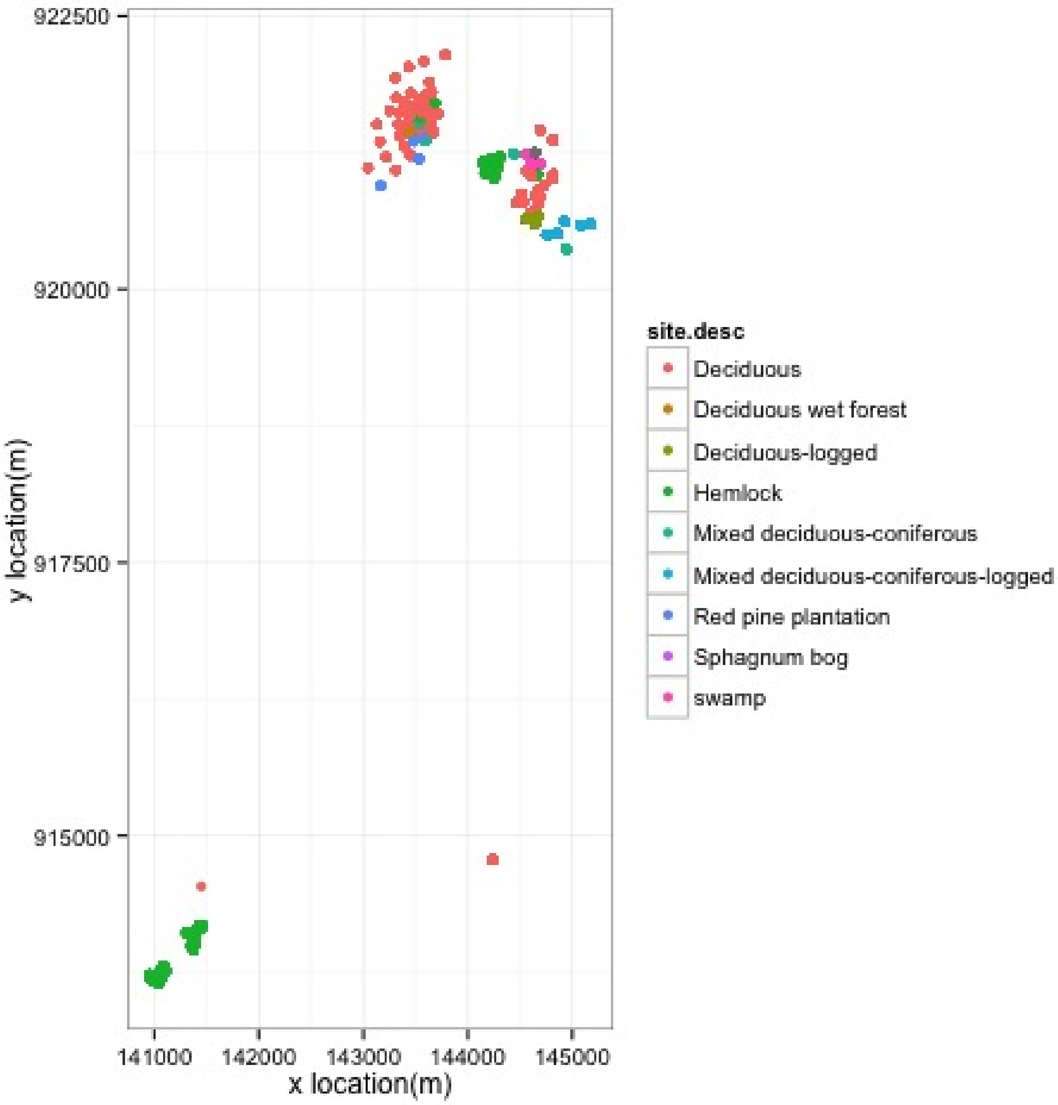


Figure 3.7: Locations of R_s collars in the Harvard Forest.

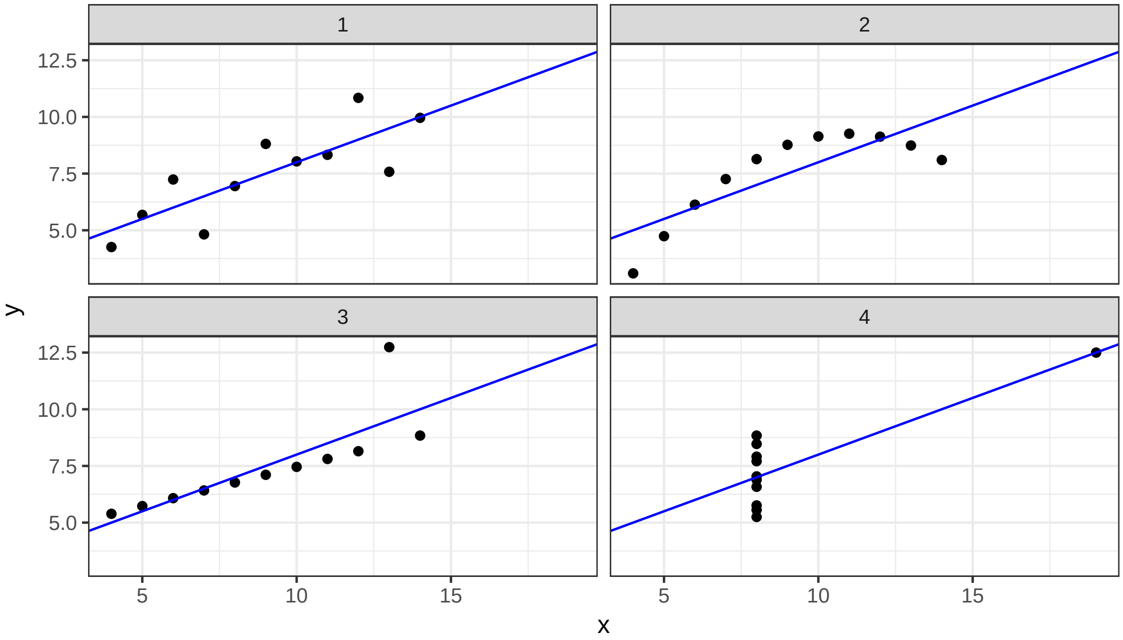


Figure 3.8: y vs. x for each of Anscombe's four data sets. The blue line has intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$. It is the same in each facet.

Index of Concepts

Index of Examples

Index of R Commands

Bibliography

Allison, T and D. V. Cicchetti (1976). “Sleep in mammals: Ecological and constitutional correlates”. In: *Science* 194, pp. 732–734.

Andrews, D. F. and A. M. Herzberg (1985). *Data*. ny: springer.

Anscombe, F. J. (1973). “Graphs in Statistical Analysis”. In: *The American Statistician* 27.1, pp. 17–21.

Bond-Lamberty, Ben and Allison Thomson (2010). “Temperature-associated increases in the global soil respiration record”. In: *Nature* 464.7288, pp. 579–582. URL: HTTP://DX.DOI.ORG/10.1038/NATURE08930.

Box, G. E. P. (1979). “Robustness in the strategy of scientific model building”. In: *Robustness in Statistics*. Ed. by R. L. Launer and G. N. Wilkinson. Academic Press, pp. 201–236.

Brodeur, Paul (1992). “Annals of Radiation, The Cancer at Slater School”. In: *The New Yorker* Dec. 7.

Giasson, M.-A. et al. (2013). “Soil respiration in a northeastern US temperate forest: a 22-year synthesis”. In: *Ecosphere* 4.11.

IPCC (2007). “The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC”. In: ed. by S. Solomon et al. Cambridge University Press. Chap. 2, section 2.3.1.

Kalbfleisch, John D. and Ross L. Prentice (2011). *The Statistical Analysis of Failure Time Data*. Wiley.

Kaplan, Daniel T. (2011). *Statistical Modeling: A Fresh Approach*. 2nd. Project Mosaic.

Lavine, Michael, Brian Beckage, and James S. Clark (2002). “Statistical Modelling of Seedling Mortality”. In: *Journal of Agricultural, Biological and Environmental Statistics* 7, pp. 21–41.

Lindley, D. V. (1993). “Discussion of *Exchangeability and Data Analysis (with Discussion)*”. In: *Journal of the Royal Statistical Society (Series A)* 156, pp. 29–30.

Peters, G et al. (2012). "CO2 emissions rebound after the Global Financial Crisis". In: *Nature Climate Change* 2, pp. 2–4.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Raich, J. W. and W. H. Schlesinger (1992). "The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate". In: *Tellus B* 44.2, pp. 81–99. ISSN: 1600-0889. DOI: `10.1034/J.1600-0889.1992.T01-1-00001.X`. URL: `HTTP://DX.DOI.ORG/10.1034/J.1600-0889.1992.T01-1-00001.X`.

Schlesinger, William H. and Jeffrey A. Andrews (2000). "Soil respiration and the global carbon cycle". English. In: *Biogeochemistry* 48.1, pp. 7–20. ISSN: 0168-2563. DOI: `10.1023/A:1006247623877`. URL: `HTTP://DX.DOI.ORG/10.1023/A:1006247623877`.

Utts, Jessica (1991). "Replication and Meta-Analysis in Parapsychology". In: *statsci* 4, pp. 363–403.

Venables, W. N., D. M. Smith, and the R Core Team (2018). *An Introduction to R.*

Weisberg, Sanford (2013). *Applied Linear Regression.* 4th. Wiley.