Extended solutions to Exercises for *Statistical Thinking from Scratch: A Primer for Scientists*

Exercise Set 2-2

1) Here is code to do the whole procedure. You'll see that petal length and width are strongly associated and that all three species can be reasonably clearly distinguished on the basis of petal length and width.

```
#Compute the sample mean and median
mean(iris$Petal.Length)
median(iris$Petal.Length)

#Histograms provide a visual summary of the distribution
hist(iris$Petal.Length, xlab = "Petal Length")
mean(iris$Petal.Length[iris$Species=="setosa"])

#We'll skip the conditional.mean function and use tapply()
#instead.
tapply(iris$Petal.Length, iris$Species, mean)
tapply(iris$Petal.Width, iris$Species, mean)

#A basic boxplot
boxplot(iris$Petal.Length ~ iris$Species)
title(xlab = "Species", ylab = "Petal Length") #Add axis labels

#A scatterplot
plot(iris$Petal.Length, iris$Petal.Width, pch =
as.numeric(iris$Species),
     xlab = "Petal Length", ylab = "Petal Width")
#And add a legend
legend("topleft", pch = c(1,2,3), legend = c("setosa",
"versicolor", "virginica"))
```

2) Code is in the question. The `gpairs()` plot includes scatterplots for every possible pair of variables in the `iris` dataset.

Exercise Set 3-1

1) a) Here's one way you could use equations 3.6 and 3.7 in `R`.

```
> x <- anscombe$x1
> y <- anscombe$y1
> n <- length(x) #number of pairs of observations
> #calculate the slope and intercept.
> b <- (sum(x*y) - (1/n) * sum(x) * sum(y))/(sum(x^2) - (1/n) *
sum(x)^2)
```

```
> a <- (sum(y) - b*sum(x))/n
> a
[1] 3.000091
> b
[1] 0.5000909
```

I assigned values values to `x`, `y`, and `n` to make the code a little easier to read, but you could calculate `a` and `b` without doing that, just replacing `x` with `anscombe$x1`, replacing `y` with `anscombe$y1`, and replacing `n` with `length(anscombe$x1)`.

To use equations 3.8 and 3.9, you could write

```
> x <- anscombe$x1
> y <- anscombe$y1
> b <- sum((x - mean(x))*(y - mean(y))) / sum((x - mean(x))^2)
> a <- mean(y) - b*mean(x)
> a
[1] 3.000091
> b
[1] 0.5000909
```

2) a) Start by plugging the expression for $\tilde{a}$ into the expression for $\tilde{b}$.

$$\tilde{b} = \frac{\sum_{i=1}^{n} x_i y_i - \tilde{a}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i - \left(\frac{\sum_{i=1}^{n} y_i - \tilde{b}\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

We can distribute the $\sum_{i=1}^{n} x_i$ on the right and then split up the expression as follows:

$$\tilde{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + \tilde{b}\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2} + \frac{\tilde{b}\left(\sum_{i=1}^{n} x_i\right)^2}{n\sum_{i=1}^{n} x_i^2}$$

Subtract $\frac{\tilde{b}\left(\sum_{i=1}^{n} x_i\right)^2}{n\sum_{i=1}^{n} x_i^2}$ from both sides.

$$\tilde{b} - \frac{\tilde{b}\left(\sum_{i=1}^{n} x_i\right)^2}{n\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2}$$

Factor out the $\tilde{b}$ on the left.

$$\tilde{b}\left(1 - \frac{1}{n}\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{\sum_{i=1}^{n} x_i^2}\right) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2}$$

Divide both sides by $\left(1 - \frac{1}{n}\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{\sum_{i=1}^{n} x_i^2}\right)$ to isolate the $\tilde{b}$

$$\tilde{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 \left(1 - \frac{1}{n}\frac{(\sum_{i=1}^{n} x_i)^2}{\sum_{i=1}^{n} x_i^2}\right)},$$

and simplify the denominator by distributing the $\sum_{i=1}^{n} x_i^2$, giving equation 3.7:

$$\tilde{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}.$$

b) Equation 3.6 can be split into

$$\tilde{a} = \frac{1}{n}\sum_{i=1}^{n} y_i - \tilde{b}\frac{1}{n}\sum_{i=1}^{n} x_i,$$

Which, by the definitions of $\bar{x}$ and $\bar{y}$, is equivalent to equation 3.8.

c) Following the hint in the problem, we work backwards, starting with the denominator of equation 3.9, which is $\sum_{i=1}^{n}[(x_i - \bar{x})^2]$. Expanding the square term gives

$$\sum_{i=1}^{n}[(x_i - \bar{x})^2] = \sum_{i=1}^{n}[x_i^2 - 2x_i\bar{x} + \bar{x}^2].$$

Breaking this into three sums and pulling out the terms that do not depend on i, we have

$$\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2\sum_{i=1}^{n} 1.$$

$\sum_{i=1}^{n} 1 = n$, and by definition, $\sum_{i=1}^{n} x_i = n\bar{x}$, so we have

$$\sum_{i=1}^{n}(x_i^2) - 2\bar{x}n\bar{x} + n\bar{x}^2 = \sum_{i=1}^{n}(x_i^2) - n\bar{x}^2,$$

Where the last step comes from noticing that $\bar{x}n\bar{x} = n\bar{x}^2$. Finally, to get the denominator of equation 3.9, replace $\bar{x}$ with its definition, $(1/n)\sum_{i=1}^{n} x_i$, giving

$$\sum_{i=1}^{n}[(x_i - \bar{x})^2] = \sum_{i=1}^{n} x_i^2 - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2.$$

Thus, the denominators of equations 3.7 and 3.9 are equivalent. Next, we need to show that the numerators of equations 3.7 and 3.9 are equivalent. The procedure is similar. We start with the numerator of equation 3.9,

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Expanding the product inside the sum gives

$$\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}).$$

Breaking this statement into four separate sums and pulling out the terms that do not depend on i gives

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y}$$

Recalling the definitions of $\bar{x}$ and $\bar{y}$ gives

$$\sum_{i=1}^{n}(x_i y_i) - n\bar{y}\bar{x} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^{n}(x_i y_i) - n\bar{y}\bar{x}.$$

Finally, to get the numerator of equation 3.7, replace $\bar{x}$ and $\bar{y}$ with their definitions,

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i y_i) - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right).$$

Having shown that the numerators and denominators of the expressions in equations 3.7 and 3.9 are equal, we have shown that the two expressions are equal.

Exercise Set 3-2

1) a) Suppose that we already know $\tilde{b}$. Then the sum of squared errors as a function of a is

$$g(a) = \sum_{i=1}^{n}(y_i^2 - 2ay_i - 2\tilde{b}x_i y_i + a^2 + 2a\tilde{b}x_i + \tilde{b}^2 x_i^2).$$

The derivative with respect to a is

$$g'(a) = \sum_{i=1}^{n}(- 2y_i + 2a + 2\tilde{b}x_i).$$

We find $\tilde{a}$ by finding any values of a for which $g'(a) = 0$. Through a set of steps parallel to those used for $\tilde{b}$ in the main text, we find the sole value of a for which $g'(a) = 0$:

$$\tilde{a} = \frac{\sum_{i=1}^{n} y_i - \tilde{b}\ \sum_{i=1}^{n} x_i}{n}.$$

b) If $x = \bar{x}$, then $\tilde{y} = \tilde{a} + \tilde{b}\bar{x}$. Replacing $\tilde{a}$ with $\bar{y} - \tilde{b}\bar{x}$ gives $\tilde{y} = \left(\bar{y} - \tilde{b}\bar{x}\right) + \tilde{b}\bar{x} = \bar{y}$, so when $x = \bar{x}$, the y-coordinate of the line is $\bar{y}$.

2) The slope is

$$\acute{b} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

One way to get the slope is to minimize the squared line errors for a line of the form $y_i = \acute{b}x_i$. A second way is to set $\tilde{a} = 0$ in equation 3.5, which works because we know that equation 3.5 gives the "best" (in the sense of minimizing the squared line errors) value of b given any provided value of a, including 0.

3) One could re-do the minimization of squared line errors, or one could simply switch x and y in the expressions for $\tilde{a}$ and $\tilde{b}$ to get

$$\tilde{c} = \frac{\sum_{i=1}^{n} x_i - \tilde{d}\sum_{i=1}^{n} y_i}{n}$$

and

$$\tilde{d} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2}$$

In general, $\tilde{c}$ and $\tilde{d}$ are not equal to $\tilde{a}$ and $\tilde{b}$. (Try `lm(anscombe$y1 ~ anscombe$x1)` and `lm(anscombe$x1 ~ anscombe$y1)`.) But they are equal if the x and y values are rescaled so that their sums are equal and the sums of their squares are equal. Sometimes people *standardize* variables—subtracting each variable's mean and dividing by the standard deviation—after which the least-squares slope is the same regardless of which variable is on which side of the linear equation.

4) a) Use

```
install.packages("quantreg")
library(quantreg)
```

b) Use

```
mod.fit.L1  <- rq(anscombe$y1 ~ anscombe$x1)
summary(mod.fit.L1)
```

The intercept is 3.24, and the slope is 0.48.

c) The least-squares and L1 lines are similar in this case.

e) The L1 line goes through almost all the points and misses the one that falls off the line. The least-squares line, in contrast, is pulled toward the outlying point. One way to understand this is to note that the sum of the squared line errors is sensitive to large line errors. As such, the least-squares line will be pulled toward individual outlying points in order to avoid making large errors. In contrast, the L1 line weights small line errors more heavily and large line errors less heavily than the least-squares line.

Exercise Set 4-1

1) The properties of complements say that $A \cup A^C = \Omega$ and that $A \cap A^C = \emptyset$. The second axiom of probability tells us that $\mathrm{P}(\Omega) = 1$, so we know that $\mathrm{P}(A \cup A^C) = 1$. In more familiar terms, A will either happen, or it will not. The fact that $A \cap A^C = \emptyset$ tells us that A and A^C are mutually exclusive—they share no events in common, so they cannot both happen. Because $A \cap A^C = \emptyset$, we can invoke the third axiom to write $\mathrm{P}(A \cup A^C) = 1 = \mathrm{P}(A) + \mathrm{P}(A^C)$. Thus, $\mathrm{P}(A^C) = 1 - \mathrm{P}(A)$. This is a useful fact: the probability of an event occuring is one minus the probability that the event does not happen.

2) We have to define the probability function for the roulette problem. We start with two facts, $\mathrm{P}(B) = \beta$ and $\mathrm{P}(R) = \rho$, plus the three axioms of probability. We need to find the probability of every event in $\mathcal{F}$:

$$\mathcal{F} = \big\{\emptyset, \{B\}, \{R\}, \{G\}, \{B,R\}, \{B,G\}, \{R,G\}, \{B,R,G\}\big\}.$$

Let’s start with what we can get just from the axioms: Axiom (ii) tells us that $\mathrm{P}(\{B,R,G\}) = 1$ because $\{B,R,G\} = \Omega$. Note that the empty set is the complement of $\{B,R,G\}$, so the result from exercise 1 above tells us that $\mathrm{P}(\emptyset) = 1 - \mathrm{P}(\Omega) = 1 - 1 = 0$.

Now we can use the facts that $\mathrm{P}(B) = \beta$ and $\mathrm{P}(R) = \rho$. First notice that $\{B\} \cap \{R\} = \emptyset$, so we can use axiom (iii) to learn that $\mathrm{P}(\{B,R\}) = \ \beta + \rho$. Exercise 1 then gives us that $\mathrm{P}(G) = 1 - \beta - \rho$. Similarly, exercise 1 tells us that $\mathrm{P}(\{B,G\}) = 1 - \rho$ and $\mathrm{P}(\{R,G\}) = 1 - \beta$.

3) If $\mathrm{P}(E_1 \cap E_2) = 0$, then the claim to be proven, $\mathrm{P}(E_1 \cup E_2) = \mathrm{P}(E_1) + \mathrm{P}(E_2) - \ \mathrm{P}(E_1 \cap E_2)$, is equivalent to the third axiom of probability. We need to find a solution that works when E_1 and E_2 share outcomes in common – when it is possible for both events to occur. Define the set difference $E_2 \backslash E_1$ as the set of elements that are members of E_2 but not members of E_1.

Notice that $E_1 \cup E_2 \backslash E_1 = E_1 \cup E_2$. This implies that $\mathrm{P}(E_1 \cup E_2) = \mathrm{P}(E_1 \cup E_2 \backslash E_1)$. Also notice that we have removed the elements of E_2 that are present in E_1, so $E_1 \cap E_2 \backslash E_1 = \emptyset$. We can invoke the third axiom to see that $\mathrm{P}(E_1 \cup E_2 \backslash E_1) = \mathrm{P}(E_1) + \mathrm{P}(E_2 \backslash E_1)$.

Putting these two statements together gives us $\mathrm{P}(E_1 \cup E_2) = \mathrm{P}(E_1) + \mathrm{P}(E_2 \backslash E_1)$. Thus, to show that $\mathrm{P}(E_1 \cup E_2) = \mathrm{P}(E_1) + \mathrm{P}(E_2) - \mathrm{P}(E_1 \cap E_2)$, we only have to show that $\mathrm{P}(E_2 \backslash E_1) = \mathrm{P}(E_2) - \mathrm{P}(E_1 \cap E_2)$.

We can apply the third axiom again. Notice that $(E_2 \backslash E_1) \cup (E_1 \cap E_2) = E_2$ and $(E_2 \backslash E_1) \cap (E_1 \cap E_2) = \emptyset$. Therefore, by the third axiom, $\mathrm{P}(E_2) = \mathrm{P}(E_2 \backslash E_1) + \mathrm{P}(E_1 \cap E_2)$. Solving for $\mathrm{P}(E_2 \backslash E_1)$ gives that $\mathrm{P}(E_2 \backslash E_1) = \mathrm{P}(E_2) - \mathrm{P}(E_1 \cap E_2)$. This implies that $\mathrm{P}(E_1 \cup E_2) = \mathrm{P}(E_1) + \mathrm{P}(E_2) - \mathrm{P}(E_1 \cap E_2)$: the probability that one of two events happens is the sum of the probability of each of them minus the probability that they both happen.

Exercise Set 4-2:

1) If events A and B are independent, then $\mathrm{P}(A|B) = \mathrm{P}(A)$. Replacing the conditional probability with its definition gives

$$\frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \mathrm{P}(A).$$

Multiplying both sides by $\mathrm{P}(B)$ gives $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. To prove the second part, divide both sides of $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$ by $\mathrm{P}(A)$. The left side becomes $\mathrm{P}(A \cap B)/\mathrm{P}(A)$, which, by the definition in equation 4.1, is $\mathrm{P}(B|A)$. Thus, $\mathrm{P}(B|A) = \mathrm{P}(B)$.

2) Start by examining the definitions.

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)},$$

and

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}.$$

If we can get $\mathrm{P}(A \cap B)$ on its own, then to get $\mathrm{P}(B|A)$, we would only need to divide by $\mathrm{P}(A)$. We can do this:

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} = \frac{\frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}\mathrm{P}(B)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|B)\mathrm{P}(B)}{\mathrm{P}(A)}.$$

The last step, $\mathrm{P}(A \cap B)/\mathrm{P}(B) = \mathrm{P}(A|B)$, follows from equation 4.1. The statement

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\mathrm{P}(B)}{\mathrm{P}(A)}$$

is called Bayes' Theorem, and we will have more to say about it in the next section.

Exercise Set 4-3:

1) The probability mass function is:

$f_X(0) = f_X(3) = \frac{1}{8}$, $f_X(1) = f_X(2) = \frac{3}{8}$, and $f_X(x) = 0$ for all other x.

2) The sum would be 1. Axiom (ii) of probability (in optional section 4.1) tells us that the probability of the event that includes all possible outcomes is 1.

3) $F_X(b) - F_X(a)$. Remember that $F_X(b) = \mathrm{P}(X \leq b)$ and $F_X(a) = \mathrm{P}(X \leq a)$. Thus, $F_X(b) - F_X(a) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a)$, or the probability that X is less than or equal to b but not less than or equal to a, $\mathrm{P}(a < X \leq b)$.

Exercise Set 4-4:

1) There are three parts to consider. For any $x < 0$, $\mathrm{P}(X \leq x) = 0$, so the c.d.f. starts with a flat line at height 0 extending from negative infinity to $x = 0$. Similarly, if $x > 1$, then $\mathrm{P}(X \leq x) = 1$, so the c.d.f. has another flat line, this one at height 1 and extending from $x = 1$ to positive infinity. The remaining interval is between 0 and 1, the values that X can take. The requirement that all equally-sized intervals in $[0,1]$ are equally likely to contain X results in a line of uniform slope connecting (0,0) and (1,1). In this case, the slope required is 1. You can simply draw the function, but here is code for plotting it in `R`:

```
x <- c(0,1)
Fx <- x
plot(x, Fx, type = "l", xlim = c(-1,2))
lines(c(-1, 0), c(0, 0))
lines(c(1,2), c(1, 1))
```

2) Just as was the case for discrete random variables, $\mathrm{P}(a \leq X < b) = F_X(b) - F_X(a)$. This means that if the probability of landing in two intervals of equal sizes differs, then the average slope of F_X in those two intervals must differ. Specifically, the interval with the higher probability must have the higher average slope. Here is R code for drawing one possible c.d.f. that meets the description in the problem. Specifically, $\mathrm{P}(0.4 \leq X < 0.6) = 0.4$:

```
x1 <- c(0,0.4)
x2 <- c(0.4, 0.6)
x3 <- c(0.6, 1)
Fx1 <- c(0,0.3)
Fx2 <- c(0.3, 0.7)
Fx3 <- c(0.7, 1)
plot(x1, Fx1, type = "l", xlim = c(-1,2), ylim = c(0,1), xlab =
"x", ylab = "Fx")
lines(x2, Fx2)
lines(x3, Fx3)
lines(c(-1, 0), c(0, 0))
lines(c(1,2), c(1, 1))
```

Exercise Set 4-5:

1) $\int_{-\infty}^{\infty} f_X(x)dx = 1$. Recall that for any cumulative distribution function, $F_X(\infty) = 1$, and recall that $F_X(\infty) = \int_{-\infty}^{\infty} f_X(x)dx$. Probability density functions always have an area under the curve of exactly 1.

2) Yes, this could be a density for continuous random variable. The area under this function is 1 (axiom ii), and the function is non-negative, which means that the probability of the random variable falling in any interval is non-negative (axiom i). This is different from a probability mass function because mass functions can never take values larger than 1, whereas this density is equal to 10 when $0 \leq x \leq 1/10$. The difference is that whereas mass functions must *sum* to 1, density functions must *integrate* to 1.

Exercise Set 4-6:

1) The probability mass function of the Poisson distribution is $\mathrm{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. Plugging in the appropriate values for k and λ gives: i) e^{-5}, ii) $5e^{-5}$, iii) $(25/2)e^{-5}$.

2) Use the probability mass function of the geometric distribution with parameter 1/2. If our first “heads” occurs on the 6th flip, then we have five tails before it. We plug $p = 1/2$ and $k = 5$ into $\mathrm{P}(X = k) = (1-p)^k p$ to get $\mathrm{P}(X = 5) = \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right) = 1/64$.

3) a) Use:

```
> x <- seq(-3, 3, length.out = 1000)
> plot(x, dnorm(x, mean = 0, sd = 1), type = "l")
```

b) Use:

```
> x <- seq(-3, 3, length.out = 1000)
> plot(x, pnorm(x, mean = 0, sd = 1), type = "l")
```

c) What value of x is at the 97.5th percentile of the standard normal distribution?

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

This value, 1.96, is actually a useful number to remember in applied statistics, for reasons that will be discussed in chapter 7.

4) a) Use:

```
> normsims <- rnorm(1000, mean = 0, sd = 1)
> hist(normsims)
```

b) Use:

```
> unifsims <- runif(1000, 0 , 1)
> hist(qnorm(unifsims, mean = 0, sd = 1))
```

We simulated uniform random draws, but we were able to transform them to draws from a normal distribution by feeding them through the normal quantile function, or the inverse of the normal distribution function.

Use this code to see a picture that shows how this works:

```
r <- seq(-3, 3, length.out = 1000)
cdf <- pnorm(r)

#Draw the normal cumulative distribution function.
plot(r, cdf, type = "l", xaxs = "i", yaxs = "i", xlim = c(-3,
3), xlab = expression(italic(x)), ylab =
expression(paste(italic(F[X]), "(", italic(x), ")", sep = "")),
lwd = 2)

#Draw light grey lines representing random samples from the
#standard normal distribution.
x <- rnorm(500)
for(i in x){
  lines(c(i,i), c(min(x), pnorm(i)), col = rgb(190, 190, 190,
     alpha = 60, max = 255))
  lines(c(min(x)-1,i), c(pnorm(i), pnorm(i)), col = rgb(190,
     190, 190, alpha = 60, max = 255))
}
```

The cumulative distribution function of X is drawn in solid black. The light grey lines represent 500 random draws from the distribution of X. Start on the horizontal axis. Each light grey line traces from the value of one of the random samples of X on the x-axis up to the cumulative distribution function. Once it hits the cumulative distribution function, it turns left until it hits the vertical axis. Notice that the positions where the lines hit the x-axis are centered on zero, symmetric, and concentrated near the middle—they look like a normal distribution. Entering `hist(x)` will confirm the suspicion. In contrast, the lines hit the y-axis with roughly uniform density from zero to one. `hist(pnorm(x))` will confirm.

To make the plot, we simulated x-values from the normal distribution and fed them into `pnorm()` to get uniformly distributed data. Effectively, we traced grey lines from the horizontal axis up to the cumulative distribution function and then to the left, ending with a uniform distribution along the vertical axis. But we can also go backwards, starting with uniformly distributed data on the vertical axis, tracing lines to the right until we get to the cumulative distribution function, and then drawing lines straight down to get normally distributed data. This is what happens when we apply `qnorm()` to uniformly distributed data.

This is a powerful idea, not just a curiosity. This approach lets us draw pseudorandom samples from any distribution with a known cumulative distribution function as long as it is possible to generate pseudorandom samples from a continuous uniform distribution.

Exercise Set 5-1:

1) a) If X is a Bernoulli random variable, then the mass function is $f_X(x) = \mathrm{P}(X = x) = p^x(1-p)^{1-x}$ for $x \in \{0,1\}$. Because there are only two possible values of X, summing is easy. The expectation is

$$\mathrm{E}(X) = \sum_{x=0}^{1} x f_X(x) = \sum_{x=0}^{1} x p^x (1-p)^{1-x} = 0p^0(1-p)^1 + 1p^1(1-p)^0 = p.$$

b) This is one example where the linearity of expectation comes in handy. If X is a binomial random variable, then $f_X(x) = \mathrm{P}(X = x) = \binom{n}{x} p^x(1-p)^{n-x}$ for $x \in \{0,1,2,\dots,n\}$. This means that the expectation is

$$\mathrm{E}(X) = \sum_{x=0}^{n} x f_X(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}.$$

It is possible to compute this directly, but there is an easier way. Recall that a binomial random variable is the number of successes out of n independent trials each with probability of success p. We already know that a Bernoulli random variable can model a single trial with probability of success p. Thus, we can re-imagine the binomial random variable as the sum of n independent Bernoulli random variables, which we can label $X_1, X_2, \dots, X_n$. So if $X = \sum_{i=1}^{n} X_i$, where the X_i are independent draws from a Bernoulli distribution with success probability p, then X is distributed as a binomial random variable with parameters n and p. We can now write

$$\mathrm{E}(X) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right).$$

And here is where linearity helps. Because the expectation is linear, the expectation of the sum is equal to the sum of the expectations. This lets us quickly finish the job using the result from part (a):

$$\mathrm{E}(X) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \sum_{i=1}^{n} p = np.$$

c) Let X be a discrete uniform random variable. Then the mass function is $f_X(x) = \mathrm{P}(X = x) = \frac{1}{b-a+1}$ for all for $x \in \{a, a+1, a+2, \dots, b\}$. (See Table 4-5). The expectation is

$$\mathrm{E}(X) = \sum_{x=a}^{b} x f_X(x) = \sum_{x=a}^{b} \frac{x}{b-a+1} = \frac{1}{b-a+1} \sum_{x=a}^{b} x.$$

Where the last step, moving $\frac{1}{b-a+1}$ outside the sum, is allowed because the sum is with respect to x and $\frac{1}{b-a+1}$ does not depend on x. Using the formula given in the hint to the problem,

$$\mathrm{E}(X) = \frac{1}{b-a+1} \sum_{x=a}^{b} x = \frac{(a+b)(b-a+1)/2}{b-a+1} = \frac{a+b}{2}.$$

d) If X is a continuous uniform random variable, then the density is $f_X(x) = \frac{1}{b-a}$ for $x \in [a, b]$. The expectation is

$$\mathrm{E}(X) = \int_a^b x f_X(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \int_a^b x dx.$$

Remembering that the integral $\int x \, dx = \frac{x^2}{2} + C$ and evaluating at a and b gives

$$\mathrm{E}(X) = \frac{(b^2 - a^2)/2}{b-a}.$$

The hint given in the problem lets us write this as

$$\mathrm{E}(X) = \frac{(b-a)(b+a)/2}{b-a},$$

which simplifies to

$$\mathrm{E}(X) = \frac{(a+b)}{2}.$$

2) a) When `n = 1`, the histogram ought to resemble the density function for a normal random variable with expectation 0 and standard deviation 1—we are just taking individual samples from a normal distribution and plotting them. That is to say it ought to be symmetric, centered around

0, and roughly bell-shaped. The great majority of the data should fall between -2 and 2 in this case. As we draw larger samples and take their means, the law of large numbers suggests that the sample means should generally be closer to the expectation than the individual observations are. Indeed, as we increase the sample size, we find that fewer samples have means that are far from the expectation. The shape of the distribution continues to look roughly normal, but the spread decreases.

b) Here is a modified version of the code that uses `rexp()` to simulate exponential random variables:

```
samp.size <- 20
n.samps <- 1000
samps <-matrix(rexp(samp.size*n.samps, rate = 1), ncol =
n.samps)
samp.means <- colMeans(samps)
hist(samp.means)
```

The shape of the exponential density is markedly different from the shape of the normal distribution. No observations smaller than 0 are allowed. When the expected value is set to 1, most of the observations are near 0, and the observations trail off far to the right. We say the distribution is “skewed right.” Again, when we take means of samples, we find that the sample means get closer to the expectation as the sample size increases. You ought to notice something odd here, though. The sample means cluster more tightly around the expectation as the sample size grows, but the shape of the distribution also changes. Namely, it starts to look more symmetric and bell-like—more normal. This is a preview of the central limit theorem, which will appear soon.

Exercise Set 5-2:

1) Our definition says that $\mathrm{Var}(X) = \mathrm{E}([X - \mathrm{E}(X)]^2)$. We expand the expression to get:

$$\mathrm{Var}(X) = \mathrm{E}(X^2 - 2X\mathrm{E}(X) + [\mathrm{E}(X)]^2).$$

Applying the linearity of expectation,

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(2X\mathrm{E}(X)) + \mathrm{E}([\mathrm{E}(X)]^2).$$

$\mathrm{E}(X)$ is a constant, as is $[\mathrm{E}(X)]^2$. The expectation of a constant is just the constant itself, so we have:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - 2\mathrm{E}(X)\mathrm{E}(X) + [\mathrm{E}(X)]^2.$$

Finally, we collect the $[\mathrm{E}(X)]^2$ terms to get

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$

2) a) This is one of the rare situations where it is easier to start with the definition of the variance in equation 5.6 than with the identity in equation 5.7. Using the definition, we have

$$\mathrm{Var}(X+c)=\mathrm{E}([X+c-\mathrm{E}(X+c)]^2).$$

Linearity of expectation (equation 5.4) lets us write this as

$$\mathrm{Var}(X+c)=\mathrm{E}([X+c-\mathrm{E}(X)-c]^2),$$

which is

$$\mathrm{Var}(X+c)=\mathrm{E}([X-\mathrm{E}(X)]^2)=\mathrm{Var}(X).$$

Thus, adding a constant to a random variable does not change the variance of the random variable.

b) This time, we'll start with equation 5.7, which gives

$$\mathrm{Var}(cX)=\mathrm{E}([cX]^2)-[\mathrm{E}(cX)]^2.$$

Linearity of expectation lets us pull the constants out of the expectations, which gives

$$\mathrm{Var}(cX)=c^2\mathrm{E}(X^2)-[c\mathrm{E}(X)]^2=c^2(\mathrm{E}(X^2)-[\mathrm{E}(X)]^2)=c^2\mathrm{Var}(X).$$

3) a) We want $f_{X,Y}(x,y)=\mathrm{P}(X=x\cap Y=y)$. If X and Y are independent, then $\mathrm{P}(X=x\cap Y=y)=\mathrm{P}(X=x)\mathrm{P}(Y=y)$. We already have $\mathrm{P}(X=x)$ and $\mathrm{P}(Y=y)$: these are $f_X(x)$ and $f_Y(y)$, respectively. Thus, if X and Y are independent, then $f_{X,Y}(x,y)=f_X(x)f_Y(y)$. That is, the joint mass function of independent discrete random variables is just the product of the *marginal* mass functions of the random variables. (Here, read "marginal" as "ignoring any other random variables.") You can use an analogous argument with cumulative distribution functions to prove a similar claim for continuous random variables.

b) Using equation 5.7,

$$\mathrm{Var}(X+Y)=\mathrm{E}[(X+Y)^2]-[\mathrm{E}(X+Y)]^2.$$

Expanding the first term and invoking the linearity of expectation on the second term gives

$$\mathrm{Var}(X+Y)=\mathrm{E}(X^2+2XY+Y^2)-[\mathrm{E}(X)+\mathrm{E}(Y)]^2,$$

Invoking linearity on the first term and expanding the second gives

$$\mathrm{Var}(X+Y)=\mathrm{E}(X^2)+2\mathrm{E}(XY)+\mathrm{E}(Y^2)-[\mathrm{E}(X)]^2-2\mathrm{E}(X)\mathrm{E}(Y)-[\mathrm{E}(Y)]^2.$$

Recognizing $\mathrm{Var}(X)=\mathrm{E}(X^2)-[\mathrm{E}(X)]^2$, $\mathrm{Var}(Y)=\mathrm{E}(Y^2)-[\mathrm{E}(Y)]^2$, and $\mathrm{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)$ lets us finish the proof:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2[\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)] = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

c) Remember that in part (a), we found that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. This means that

$$\mathrm{E}(XY) = \sum_x \sum_y xy\, f_X(x)f_Y(y).$$

Because $xf_X(x)$ is not a function of y and $yf_Y(y)$ is not a function of x, we can rewrite this as

$$\mathrm{E}(XY) = \left[\sum_x xf_X(x)\right]\left[\sum_y y\, f_Y(y)\right] = \mathrm{E}(X)\mathrm{E}(Y).$$

This result means that when X and Y are independent discrete random variables, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. You can prove the same result for continuous random variables using integrals over the joint density instead of sums over the joint mass function.

d) The trick is to notice that $X - Y = X + cY$, where $c = -1$. If X and Y are independent, then X and cY are also independent. This means that $\mathrm{Var}(X+cY) = \mathrm{Var}(X) + \mathrm{Var}(cY)$. We already proved that $\mathrm{Var}(cY) = c^2\mathrm{Var}(Y)$. If $c = -1$, then $\mathrm{Var}(cY) = \mathrm{Var}(Y)$, and so $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = \mathrm{Var}(X+Y)$ if X and Y are independent.

4) a) Bernoulli random variables only take two values, 0 and 1. Note that because $0^2 = 0$ and $1^2 = 1$, if X is a Bernoulli random variable, then $X = X^2$ in every case, which implies that $\mathrm{E}(X^2) = \mathrm{E}(X) = p$. By equation 5.7, this means that the variance is $\mathrm{Var}(X) = p - p^2 = p(1-p)$.

b) A binomial random variable with parameters n and p is just the sum of n independent Bernoulli random variables with parameter p. Because the variance of the sum of independent random variables is just the sum of the variances of the random variables, the variance of the binomial random variable is $\mathrm{Var}(X) = np(1-p)$. (This is so much easier than it is to calculate $\mathrm{E}(X^2) = \sum_{x=0}^{n} x^2\binom{n}{x}p^x(1-p)^{n-x}$.)

5) First, using equation 5.8, we note that

$$\mathrm{Var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n)$$

because $1/n$ is a constant. We also know that because the Xs are independent, the variance of the sum is the sum of the individual variances (equation 5.9), which is $n\sigma^2$. Thus, the variance is

$$\mathrm{Var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

The standard deviation is the square root of the variance:

$$\mathrm{SD}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{\sigma}{\sqrt{n}}$$

This is an important result in statistics. As we add observations to a sample, the sample mean becomes less and less variable, as long as we are taking observations from a distribution with finite variance. The result about the standard deviation of the sample mean shows that our increases in precision per additional sample—when considered in the original units—diminish as the sample size increases. Every time we double the sample size, the standard deviation decreases by a factor of $\sqrt{2} \approx 1.41$.

6) a) We define Z such that $Z = 1$ if $X \geq c$ and $Z = 0$ if $X < c$. X may have any distribution so long as it cannot take negative values. Z is a Bernoulli random variable with parameter $p = \mathrm{P}(X \geq c)$. Because the expectation of a Bernoulli random variable with parameter p is p (Exercise set 5-1, problem 1), we have

$$\mathrm{P}(X \geq c) = \mathrm{E}(Z).$$

If $X \geq c$, then $X/c \geq 1$ and $Z = 1$, so $Z \leq X/c$. Similarly, if $X < c$, then $Z = 0$, and because X is nonnegative and c is positive, once again, $Z \leq X/c$. This means that $Z \leq X/c$ in all cases, so

$$\mathrm{E}(Z) \leq \mathrm{E}(X/c).$$

Using equation 5.4 to pull c out of the expectation and combining gives

$$\mathrm{P}(X \geq c) = \mathrm{E}(Z) \leq \frac{\mathrm{E}(X)}{c},$$

which proves Markov's inequality.

b) $(Y - \mu)^2$ is a non-negative random variable, so Markov's inequality applies to it. $\mathrm{E}[(Y - \mu)^2] = \mathrm{Var}(Y)$ by definition, so Markov's inequality gives us

$$\mathrm{P}((Y - \mu)^2 \geq c) \leq \frac{\mathrm{Var}(Y)}{c}.$$

Now define another constant d as $d = \sqrt{c}$. Notice that the statement $(Y - \mu)^2 \geq c$ is equivalent to $|Y - \mu| \geq d$: whenever one statement is true, the other statement is guaranteed to be true. This means that $\mathrm{P}((Y - \mu)^2 \geq c) = \mathrm{P}(|Y - \mu| \geq d)$. Making this replacement and replacing the c on the right with d^2 gives Chebyshev's inequality:

$$\mathrm{P}(|Y - \mu| \geq d) \leq \frac{\mathrm{Var}(Y)}{d^2}.$$

c) Chebyshev's inequality gets us most of the way there. We need to prove that for positive δ, $\lim_{n\to\infty} \mathrm{P}(|\bar{X}_n - \mu| > \delta) = 0$. Chebyshev's inequality gives

$$\mathrm{P}(|\bar{X}_n - \mu| > \delta) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\delta^2}.$$

Because the expressions on both sides of the inequality are positive, if $\lim_{n\to\infty} \frac{\mathrm{Var}(\bar{X}_n)}{\delta^2} = 0$, then $\lim_{n\to\infty} \mathrm{P}(|\bar{X}_n - \mu| > \delta) = 0$. (This follows from the squeeze theorem.) Thus, we can prove the weak law of large numbers by proving that

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\bar{X}_n)}{\delta^2} = 0.$$

The result of problem 5 lets us write

$$\frac{\mathrm{Var}(\bar{X}_n)}{\delta^2} = \frac{\sigma^2}{n\delta^2}$$

Because σ^2 and δ^2 are finite constants, $(\sigma^2/\delta^2)/n$ approaches 0 as n approaches infinity, which implies that $\mathrm{P}(|\bar{X}_n - \mu| > \delta)$ approaches 0 and proves the weak law of large numbers.

Exercise Set 5-3:

1) a) We have to find $\mathrm{E}(XY)$, $\mathrm{E}(X)$, and $\mathrm{E}(Y)$. We do this by summing over the joint mass function. We can do this with a table like the following:

Outcome	Probability	X	Y	XY
$X = -1, Y = 1$	1/3	-1	1	-1
$X = 1, Y = 1$	1/3	1	1	1
$X = 0, Y = -2$	1/3	0	-2	0
		$\mathrm{E}(X) = 0$	$\mathrm{E}(Y) = 0$	$\mathrm{E}(XY) = 0$

In the bottom row, we compute the expectations by taking an average weighted by the probabilities, which, in this case, are all equal. Thus, $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = 0$.

b) $\mathrm{Cov}(X, Y) = 0$, but X and Y are not independent. For example, $\mathrm{P}(X = 0 \cap Y = -2) = 1/3 \neq \mathrm{P}(X = 0)\mathrm{P}(Y = -2) = 1/9$. You can compute similar examples for the other outcomes, but only one is needed to show that the two random variables are not independent.

2) First, the covariance:

$$\mathrm{Cov}(Z, Y) = \mathrm{Cov}(a + bX, Y) = \mathrm{E}[(a + bX)Y] - \mathrm{E}(a + bX)\mathrm{E}(Y)$$

Remember that expectation is linear (equation 5.4) and that the expectation of a constant is just that constant:

$$= \mathrm{E}[aY + bXY] - \mathrm{E}(a)\mathrm{E}(Y) - b\mathrm{E}(X)\mathrm{E}(Y) = a\mathrm{E}(Y) + b\mathrm{E}(XY) - a\mathrm{E}(Y) - b\mathrm{E}(X)\mathrm{E}(Y)$$
$$= b[\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)] = b\gamma.$$

To get the correlation, we take the covariance and divide by $\sqrt{\mathrm{Var}(Z)\mathrm{Var}(Y)}$. Remember that if X is a random variable a and b are constants, then $\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$. Thus,

$$\mathrm{Cor}(Z, Y) = \frac{b\gamma}{\sqrt{\mathrm{Var}(Z)\mathrm{Var}(Y)}} = \frac{b\gamma}{\sqrt{\mathrm{Var}(a + bX)\mathrm{Var}(Y)}} = \frac{b\gamma}{\sqrt{b^2\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

The b^2 can come outside the square root as b, where it will cancel with the b in the numerator, leaving

$$\mathrm{Cor}(Z, Y) = \frac{\gamma}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

From the definition of correlation and the fact that $\gamma = \mathrm{Cov}(X, Y)$, it follows that this is

$$\mathrm{Cor}(Z, Y) = \mathrm{Cor}(X, Y) = \rho.$$

This shows the covariance changes with linear scaling of one (or both) of the random variables being considered, but the correlation does not change with scaling.

3) Start by distributing the joint distribution and then splitting up the sum,

$$\mathrm{E}(X + Y) = \sum_{i=1}^{k}\sum_{j=1}^{m}(x_i + y_j)f_{X,Y}(x_i, y_j) = \sum_{i=1}^{k}\sum_{j=1}^{m}[x_i f_{X,Y}(x_i, y_j) + y_j f_{X,Y}(x_i, y_j)]$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{m} x_i f_{X,Y}(x_i, y_j) + \sum_{i=1}^{k}\sum_{j=1}^{m} y_j f_{X,Y}(x_i, y_j)$$

Now in the first double-sum, the x_i can be pulled outside the sum over j (because it does not depend on j), and by equation 5.13, $\sum_{j=1}^{m} f_{X,Y}(x_i, y_j) = f_X(x_i)$, giving

$$\sum_{i=1}^{k}\sum_{j=1}^{m} x_i f_{X,Y}(x_i, y_j) = \sum_{i=1}^{k} x_i \sum_{j=1}^{m} f_{X,Y}(x_i, y_j) = \sum_{i=1}^{k} x_i f_X(x_i) = \mathrm{E}(X).$$

We are allowed to switch the order of the two sums in the second double-sum, after which we can make the analogous computation,

$$\sum_{j=1}^{m}\sum_{i=1}^{k} y_j f_{X,Y}(x_i, y_j) = \sum_{j=1}^{m} y_i f_Y(y_i) = \mathrm{E}(Y).$$

Replacing the two double sums in the first expression gives the result.

Exercise Set 5-4:

2) Here's a set of commands that will let you explore the parameter set (1,1):

```
dosm.beta.hist(1, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(2, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(3, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(4, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(5, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(10, 10000, shape1 = 1, shape2 =  1)
dosm.beta.hist(50, 10000, shape1 = 1, shape2 =  1)
```

In this case, the normal distribution is an acceptable approximation for the distribution of means of samples of size 5, and it's a great approximation for samples of 10 and 50. The other values of the shape parameters suggest that for distributions that are more skewed or more U-shaped, larger samples are required before the normal distribution is a good approximation, but for the parameter sets seen here, the means of samples of size 50 are well-approximated by a normal distribution. Try to find more extreme parameter sets and see how large the samples need to be before their distribution looks normal.

3) Below is commented code that performs the simulations and makes the comparisons. With the parameters and sample size requested, the distribution of sample means is a good fit to the normal within about 2 standard deviations of the expectation, and the histogram looks kind of normal. Beyond two standard deviations, though, the Pareto sample mean distribution has much heavier tails than the normal—extreme observations are much more likely than normal theory predicts. For example, there are about 100 times as many observations beyond 5 standard deviations from the expectation as would be predicted by the normal distribution, and there are thousands of times as many observations beyond 6 standard deviations as the normal distribution predicts. Thus, with this distribution and $n = 1{,}000$, convergence in the center of the distribution is good, but convergence in the tails is poor. If the probability of an extreme event (such as, say, an earthquake of Richter magnitude >8) is important to know, then the central limit theorem can lead to spectacularly poor predictions. Convergence is worse with smaller shape parameters and smaller sample size.

```
#Sample size per simulation (n) and number of simulations.
n <- 1000
n.sim <- 100000

#Pareto parameters. Variance is finite, and so
#CLT applies, if a > 2. For large a, convergence to
#normal is better. With small a, convergence is slow,
#especially in the tails.
a <- 3
b <- 1
```

```
#Compute the expectation and variance of the distribution
#of the sample mean. a must be above 2 for these expressions
#to hold.
expec.par <- a*b/(a-1)
var.par <- a*b^2 / ((a-1)^2 * (a-2))
sd.mean <- sqrt(var.par / n)

#Simulate data, compute sample means.
sim <- matrix(rpareto(n*n.sim, a, b), nrow = n.sim)
means.sim <- rowMeans(sim)

#Draw a histogram of the sample means along with the approximate
#normal pdf that follows from the CLT.
hist(means.sim, prob = TRUE)
curve(dnorm(x, expec.par, sd.mean), add = TRUE)

compare.tail.to.normal(means.sim, 1/2, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 1, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 2, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 3, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 4, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 5, expec.par, sd.mean)
compare.tail.to.normal(means.sim, 6, expec.par, sd.mean)
```

Exercise Set 5-5:

1) From equation 5.16, the correlation coefficient is

$$\rho_{X,Y} = \beta \frac{\sigma_X}{\sigma_Y} = \frac{\beta \sigma_X}{\sqrt{\beta^2 \sigma_X^2 + \sigma_\epsilon^2}}.$$

Squaring the correlation coefficient gives

$$\rho_{X,Y}^2 = \left(\frac{\beta \sigma_X}{\sqrt{\beta^2 \sigma_X^2 + \sigma_\epsilon^2}}\right)^2 = \frac{\beta^2 \sigma_X^2}{\beta^2 \sigma_X^2 + \sigma_\epsilon^2} = 1 - \frac{\sigma_\epsilon^2}{\beta^2 \sigma_X^2 + \sigma_\epsilon^2} = 1 - \frac{\mathrm{Var}(Y|X = x)}{\mathrm{Var}(Y)}.$$

If the relationship between X and Y is linear—as it is in the model here—then the square of the correlation coefficient of X and Y is equal to one minus the proportion of the variance in Y that remains after conditioning on X. This is why people refer to the square of the correlation coefficient as “the proportion of variance explained.” This phrase is justified when the relationship between X and Y has the properties of the model we developed in the previous section.

2) When one repeatedly simulates with the same parameters, the resulting data vary. The degree to which they vary depends on the values of the parameters—in particular, making `var.eps`

larger both decreases the apparent strength of the relationship between `x` and `y` and increases the extent to which the results of different simulations vary. Varying `a` changes the numbers on the y-axis but little else. Varying `b` changes the strength and direction of the relationship between `x` and `y`—large absolute values give apparently stronger relationships, and changing the value from positive to negative changes the direction of the relationship. Changing `var.x` changes the spread of the observations on both the x- and y-axes.

Exercise Set 6-1:

1) The expectation of the sample mean is

$$\mathrm{E}[\hat{\theta}_n(D)] = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) = \mathrm{E}(X_1) = \theta.$$

The first step follows from applying the expectation operator to the definition in equation 6.1. The second step comes from the linearity of expectation (equation 5.4). The third step comes from the fact that all the X_i have the same expectation, and the fourth step comes from the fact that the first parameter of a normal distribution is equal to its expectation (equation 5.21). Because the expectation of the estimator is equal to the quantity we are trying to estimate, the estimator has a bias of zero—we say that it is “unbiased.”

If we do not know that the data are drawn from a normal distribution, the sample mean is still an unbiased estimator of the population expectation. The argument is the same as the above, but omitting the final step.

2) Use the `norm.samps()` function to draw a set of samples:

```
> s.mat <- norm.samps(0, 1, 25, 10000)
```

Notice that the “25” in the function call specifies the size of each sample; the “10000” specifies how many samples we draw.

We can calculate the mean and median of each sample using `apply()` (or using the `for()` loop given in the question):

```
> ests.mean <- apply(s.mat, 1, mean)
> ests.median <- apply(s.mat, 1, median)
```

You can use the `hist()` function to plot the means and medians of each sample, and you can also take their mean. The mean of the sample means will approach the expectation of the sample mean by the law of large numbers. The median does too, but the law of large numbers doesn’t tell us that.

```
> hist(ests.mean)
> hist(ests.median)
> mean(ests.mean)
> mean(ests.median)
```

You should see that the histogram of the sample median is centered around θ—in this case, set to 0 by the `norm.samps()` call—and that the mean of the sample medians is very close to zero. Repeating the procedure gives similar results. The results correctly suggest that the sample median is an unbiased estimate of θ when the data are independent samples from a $\mathrm{Normal}(\theta, 1)$ distribution. These results do not constitute proof, but they do suggest what turns out to be the right answer in this case.

Exercise Set 6-2:

1) We have seen this before. Because each observation is independent, it follows from equations 5.8 and 5.9 that

$$\mathrm{Var}(\hat{\theta}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n\mathrm{Var}(X_1) = \frac{1}{n}.$$

The second step comes from equation 5.8; the third step comes from equation 5.9 and the fact that the X_i are independent and identically distributed, and the fourth step comes from cancelling the n and remembering that $\mathrm{Var}(X_1) = 1$ under our model.

Notice that we only appealed to properties of the variance and to the fact that the observations are independent and identically distributed. We did not use the normality assumption, and we only used the known variance of the observations in the last step. Thus, if the data $X_1, X_2, \dots, X_n$ are independent and identically distributed with $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \cdots = \mathrm{Var}(X_n) = \sigma^2$, then the variance of the sample mean as an estimator of $\mathrm{E}(X_1)$ is

$$\mathrm{Var}(\hat{\theta}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}.$$

2) Use the `norm.samps()` function (from exercise set 6-1, problem 2) to draw a set of samples:

```
> s.mat <- norm.samps(0, 1, 25, 10000)
```

We can calculate the mean and median of each sample using `apply()` (or using the `for()` loop given in exercise set 6-1, problem 2):

```
> ests.mean <- apply(s.mat, 1, mean)
> ests.median <- apply(s.mat, 1, median)
```

Use the `var()` command to estimate the variance of the sample mean and sample median:

```
> var(ests.mean)
> var(ests.median)
```

When I computed these numbers, I found that the variance of the sample mean was 0.04 whereas the variance of the sample median was 0.06. You can also use `boxplot()` to see that the sample median is less precise:

```
> boxplot(ests.mean, ests.median)
```

Repeating this procedure gives similar results. Try it with different sample sizes by changing the `n` argument—set to 25 above—in the `norm.samps()` command. You should see that though the variance of both the sample mean and the sample median decrease when the size of each sample increases, the sample median has a larger variance—is less precise—than the sample mean. This turns out to be true in general when samples are drawn from a normal distribution.

Exercise Set 6-3:

1) For notational compactness, we write $\hat{\theta}_n(D)$ as $\hat{\theta}_n$, remembering implicitly that estimators are functions applied to random data. Starting with the definition of mean squared error given in equation 6.4, we have

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right].$$

Expanding the squared term in the definition gives

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{E}\left({\hat{\theta}_n}^2 - 2\theta\hat{\theta}_n + \theta^2\right).$$

By the linearity of expectation (equation 5.4), this is

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{E}\left({\hat{\theta}_n}^2\right) - 2\theta\mathrm{E}(\hat{\theta}_n) + \theta^2.$$

By the identity for the variance given in equation 5.7, the first term is $\mathrm{E}\left({\hat{\theta}_n}^2\right) = \mathrm{Var}\left({\hat{\theta}_n}^2\right) + [\mathrm{E}(\hat{\theta}_n)]^2$, letting us write

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{Var}\left({\hat{\theta}_n}^2\right) + [\mathrm{E}(\hat{\theta}_n)]^2 - 2\theta\mathrm{E}(\hat{\theta}_n) + \theta^2.$$

Noticing that $[\mathrm{E}(\hat{\theta}_n)]^2 - 2\theta\mathrm{E}(\hat{\theta}_n) + \theta^2 = \left(\mathrm{E}(\hat{\theta}_n) - \theta\right)^2 = \mathrm{B}(\hat{\theta}_n)^2$ completes the proof:

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{B}(\hat{\theta}_n)^2 + \mathrm{Var}(\hat{\theta}_n).$$

2) We have already seen that the sample mean and sample median are unbiased estimators of the first parameter of a normal distribution (exercise set 6-1, problem 2). Because the bias of each estimator is zero, the mean squared error of each estimator is equal to its variance by equation

6.5. We saw that the mean of a sample of normally distributed data has lower variance than the median of a sample of normally distributed data (exercise set 6-2, problem 2). Thus, the sample mean has lower mean squared error than the sample median as an estimator of the first parameter of a normal distribution.

Exercise Set 6-4:

1) The sample mean is a consistent estimator of the expectation of a random variable's distribution, regardless of the distribution family of the random variable. This follows from the weak law of large numbers (equation 5.3)—just note that $\bar{X}_n$ is the estimator and μ is the quantity being estimated, and equation 5.3 is equivalent to equation 6.6. The sample mean is also a consistent estimator of the first parameter of a normal distribution because if $X \sim \text{Normal}(\theta, 1)$, then $\text{E}(X) = \theta$ (equation 5.21).

2) The sample median is unbiased, so by equation 6.7, we only need to convince ourselves that the variance of the sample median decreases to zero as the sample size increases.

Use the `norm.samps()` function (from exercise set 6-1, problem 2) to draw a set of samples:

```
> s.mat <- norm.samps(0, 1, 25, 10000)
```

Calculate the median of each sample using `apply()` (or using the `for()` loop given in exercise set 6-1, problem 2):

```
> ests.median <- apply(s.mat, 1, median)
```

Use the `var()` command to estimate the variance of the sample median:

```
> var(ests.median)
```

Now try increasing the sample size, set to 25 in the above call. As you increase the size of the sample, you should see that the variance of the sample median gets smaller. Here is a set of commands that will do the trick:

```
#Generate 5 sets of normal samples with 10,000 samples of each
#of these sizes: 25, 50, 100, 500, 1000.
s.mat.25 <- norm.samps(0, 1, 25, 10000)
s.mat.50 <- norm.samps(0, 1, 50, 10000)
s.mat.100 <- norm.samps(0, 1, 100, 10000)
s.mat.500 <- norm.samps(0, 1, 500, 10000)
s.mat.1000 <- norm.samps(0, 1, 1000, 10000)

#Calculate the median of each sample generated above.
ests.median.25 <- apply(s.mat.25, 1, median)
ests.median.50 <- apply(s.mat.50, 1, median)
ests.median.100 <- apply(s.mat.100, 1, median)
ests.median.500 <- apply(s.mat.500, 1, median)
ests.median.1000 <- apply(s.mat.1000, 1, median)

#Estimate the variance of the sample median at each of the
#specified sample sizes.
var(ests.median.25)
var(ests.median.50)
var(ests.median.100)
var(ests.median.500)
var(ests.median.1000)

#Look at the variability of the sample median for each sample
#size.
boxplot(ests.median.25, ests.median.50, ests.median.100,
ests.median.500, ests.median.1000)
```

You should see that the variance of the sample median decreases as the size of the sample increases. We cannot prove it rigorously by simulation, but the variance of the sample median continues to approach a limit of zero as the sample size increases.

3) a) The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is both unbiased and consistent. We proved that it is unbiased in problem 1 of exercise set 6-1, and we proved that it is consistent in exercise 1 of this set.

b) The shifted sample mean is biased. $\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i + 1\right) = \theta + 1$. The shifted sample mean is also inconsistent. The sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ is consistent, meaning that

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \theta\right| > \epsilon\right) = 0.$$

This means that the shifted sample mean converges in probability to $\theta + 1$:

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i + 1 - (\theta + 1)\right| > \epsilon\right) = 0.$$

Because the shifted sample mean converges in probability to $\theta + 1$, it does not converge to θ, and therefore, it is inconsistent.

c) The first observation is unbiased; $\mathrm{E}(X_1) = \theta$. This follows from equation 5.21. However, the first observation is not a consistent estimator of θ. No matter how large the sample gets, the first observation is normally distributed with variance 1. If we draw a large sample, this estimator throws out most of the sample and uses only the first observation. Say we set $\epsilon = 1$. Then the limiting probability of X_1 being more than ϵ away from θ is

$$\lim_{n\to\infty} \mathrm{P}(|X_1 - \theta| > 1) \approx 0.32.$$

(To see this, look in a table of a normal distribution or use `2*pnorm(-1)` in `R`.) Because this limiting probability is not zero, the first observation is not a consistent estimator of θ.

d) The "shrunk" sample mean is biased; $\mathrm{E}\left(\frac{1}{n+1}\sum_{i=1}^{n} X_i\right) = \frac{n}{n+1}\theta$, which implies that

$$B\left(\frac{1}{n+1}\sum\nolimits_{i=1}^{n} X_i\right) = \left(\frac{n}{n+1} - 1\right)\theta = -\frac{\theta}{n+1}.$$

At the same time, the shrunk sample mean is consistent. Invoking the assumption that the true variance of each observation is 1, the variance of the shrunk sample mean is

$$\mathrm{Var}\left(\frac{1}{n+1}\sum\nolimits_{i=1}^{n} X_i\right) = \frac{n}{(n+1)^2}.$$

Using equation 6.5, the mean squared error of the shrunk sample mean is then

$$\mathrm{MSE}\left(\frac{1}{n+1}\sum\nolimits_{i=1}^{n} X_i\right) = \frac{\theta^2 + n}{(n+1)^2}.$$

As the sample size n increases, the denominator becomes much larger than the numerator, and so

$$\lim_{n\to\infty} \mathrm{MSE}\left(\frac{1}{n+1}\sum\nolimits_{i=1}^{n} X_i\right) = 0,$$

which, by equation 6.7, implies that the shrunk sample mean is a consistent estimator of θ.

Together, parts a-d show that it is possible for an estimator to be unbiased and consistent, biased and inconsistent, unbiased and inconsistent, or biased and consistent.

4) (Optional) For an estimator $\hat{\theta}_n$ of a quantity θ, we start by assuming that $\lim_{n\to\infty} \mathrm{MSE}(\hat{\theta}_n) = 0$. We want to prove that if the mean squared error converges to zero, then for positive δ,

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\hat{\theta}_n(D) - \theta\right| > \delta\right) = 0.$$

Notice that because $\left|\hat{\theta}_n(D) - \theta\right|$ and δ are positive, $\left|\hat{\theta}_n(D) - \theta\right| > \delta$ if and only if $(\hat{\theta}_n(D) - \theta)^2 > \delta^2$. Thus,

$$\mathrm{P}\left(\left|\hat{\theta}_n(D) - \theta\right| > \delta\right) = \mathrm{P}\left[(\hat{\theta}_n(D) - \theta)^2 > \delta^2\right].$$

$(\hat{\theta}_n(D) - \theta)^2$ is guaranteed to be non-negative, and δ^2 is positive, so we can apply Markov's inequality to get

$$\mathrm{P}\left(\left|\hat{\theta}_n(D) - \theta\right| > \delta\right) = \mathrm{P}\left[(\hat{\theta}_n(D) - \theta)^2 > \delta^2\right] \le \frac{\mathrm{E}\left[\left(\hat{\theta}_n(D) - \theta\right)^2\right]}{\delta^2}.$$

Notice that the numerator on the right, $\mathrm{E}\left[\left(\hat{\theta}_n(D)-\theta\right)^2\right]$, is $\mathrm{MSE}(\hat{\theta}_n)$ (equation 6.4). By the assumption that $\lim_{n\to\infty}\mathrm{MSE}(\hat{\theta}_n)=0$, we therefore have

$$\lim_{n\to\infty}\mathrm{P}\left(\left|\hat{\theta}_n(D)-\theta\right|>\delta\right)\leq\lim_{n\to\infty}\frac{\mathrm{E}\left[\left(\hat{\theta}_n(D)-\theta\right)^2\right]}{\delta^2}=0.$$

Because probabilities cannot be less than zero, we can replace the less-than-or-equal sign with an equal sign, giving equation 6.6. Thus, if equation 6.7 holds for an estimator, then equation 6.6 holds as well, which is what we wanted to prove.

Exercise Set 6-5

1) a) Here is some `R` code that estimates the relative efficiency of the sample mean and median as an estimator of the first parameter of a normal distribution using a sample of five observations. Use the `norm.samps()` function (from exercise set 6-1, problem 2) to draw a set of samples.

```
mu <- 0
s.mat <- norm.samps(mu, 1, 25, 10000)
ests.mean <- apply(s.mat, 1, mean)
ests.median <- apply(s.mat, 1, median)
```

Here’s the new part:

```
#The relative efficiency is estimated as the quotient of the
#MSEs. The relative efficiency of the sample mean vs. the
#sample median has the MSE of the sample mean in the
#denominator.
re <- mean((ests.med - mu)^2)/mean((ests.mean - mu)^2)
re
```

When I ran this code, I obtained a relative efficiency of 1.4. For samples of size five from a normal distribution, the sample mean is a more efficient estimator of the first parameter of a normal distribution than the sample median is. The sample mean’s mean squared error is lower.

b) Here is some `R` code that computes the requested estimates of relative efficiency and makes a basic plot of them:

```
n <- c(2,5,10,20,50,100,200,500)
nsims <- 10000
mu <- 0
sigma <- 1

re <- numeric(length(n))
```

```
for(i in 1:length(n)){
  x <- matrix(rnorm(n[i]*nsims, mu, sigma), nrow = nsims, ncol =
n[i])
  ests.med <- apply(x, 1, median)
  ests.mean <- apply(x, 1, mean)
  re[i] <- mean((ests.med - mu)^2)/mean((ests.mean - mu)^2)
}

plot(n, re, xlab = "sample size", ylab = "RE of sample mean vs.
median for normal data")
```

When I run this code, the relative efficiency appears to level off between 1.5 and 1.6. This agrees with theoretical results—a little math (beyond our scope) shows that the true asymptotic relative efficiency is $\pi/2 \approx 1.57$.

2) Once you have `rlaplace()` defined, you can complete this exercise by replacing `norm.samps()` in the solution to problem 1 with `laplace.samps()`. (Also remember to change the y axis label of the plot to note that you're using Laplace-distributed data.)

You'll see that things have reversed—if the data are Laplace distributed, then the median is actually a more efficient /lower variance estimator than the mean is, particularly for large samples. The point is that efficiency is not a property of a statistic, it is a property of an estimator under a model. If the model changes, then the relative efficiency of estimators may also change.

Exercise Set 6-6

1) Here is `R` code to draw the plots and some notes about each plot. (You don't have to use `R` to draw your own plots, but you can use this code to check hand-drawn plots.)

a)
```
x <- c(0, 1000, 2000)
y <- c(1000, 0 , 1000)
plot(x, y, pch = "", xlab = "Estimate", ylab = "Loss")
lines(x,y)
```

Relative to the maximum profit, we lose no money if our estimate is exactly right. Thus, $\lambda(1000{,}1000) = 0$. For every bushel by which our estimate is wrong, in either direction, we lose \$1—we lose \$1 in possible profit if we grow too little and we lose \$1 in the cost of growing a bushel of wheat we cannot sell if we grow too much. Thus, the loss function is

$$\lambda(\theta, \hat{\theta}_n) = |\hat{\theta}_n - \theta|,$$

which is also called "absolute error" loss.

b) The code is almost the same as in part (a), with one small change:

```
x <- c(0, 1000, 2000)
y <- c(1000, 0, 2000)
plot(x, y, pch = "", xlab = "Estimate", ylab = "Loss")
lines(x,y)
```

Because it now costs $2 to grow a bushel of wheat, it is costlier to overestimate the baker's demand by a given amount than it is to underestimate the baker's demand by that same amount. The loss function is

$$\lambda(\theta, \hat{\theta}_n) = \begin{cases} \theta - \hat{\theta}_n & \hat{\theta}_n \leq \theta \\ 2(\hat{\theta}_n - \theta) & \hat{\theta}_n > \theta \end{cases}.$$

c) In this case, $\hat{\theta}_n$ can take only integer values from 1 to 6. If we pick any number other than 3, we lose a dollar compared with what we would have won if we had picked correctly.

```
x <- 1:6
y <- c(1,1,0,1,1,1)
plot(x, y, pch = 19, xlab = "Estimate", ylab = "Loss")
```

The loss function is

$$\lambda(\theta, \hat{\theta}_n) = \begin{cases} 0 & \hat{\theta}_n = \theta \\ 1 & \hat{\theta}_n \neq \theta \end{cases}.$$

This loss function is called 0-1 loss. (Pronounced "zero-one loss.")

Exercise Set 6-7

1) a) You already have simulations to approximate the risk of the median of 100 independent normal samples under squared error loss. Because the sample median is unbiased, the risk under squared error loss is equal to the variance (exercise set 6-3, problem 2). We can use the `norm.samps()` function from exercise set 6-1, problem 2 to draw 100,000 samples of size 100, then use code from exercise set 6-4, problem 2 to compute the sample medians and check their variance:

```
> s.mat <- norm.samps(0, 1, 100, 100000)
> ests.median <- apply(s.mat, 1, median)
> var(ests.median)
```

When I run this code, I get an estimate of about 0.0154, which is larger than the risk of the sample mean.

b) After running the code in part (a) to simulate independent samples from a normal distribution, calculate the mean of each sample:

```
> ests.mean <- apply(s.mat, 1, mean)
```

Then, if you simulated with $\theta = 0$, you can calculate the approximate risk under absolute error loss with

```
> mean(abs(ests.mean))
> mean(abs(ests.median))
```

If you set θ to be a value other than 0, you need to subtract it from every entry in your vector of means before taking the absolute value. For example, if you set $\theta = 5$, then you would approximate the absolute error risk with

```
> mean(abs(ests.mean - 5))
```

When I run this code, I find that the risk for the sample mean is about 0.08 and that the risk for the sample median is about 0.10, still larger than the risk for the sample mean. Note that though the risk of the mean and median are larger and more similar with absolute error loss than with squared-error loss, absolute error loss and squared error loss are in different units—original units vs. squared units. Thus, the fact that the risks are larger and more similar isn't necessarily meaningful.

c) After running the code in parts (a) and (b), assuming that you set $\theta = 0$, the approximate risk is given by

```
mean(abs(ests.mean^3))
mean(abs(ests.median^3))
```

I get an approximate risk of 0.0016 for the sample mean and 0.0031 for the sample median.

2) Here is some `R` code that draws all four risk functions on the same plot. You can use it to check your answers for parts (a-d). Justification for each risk function is given below:

```
n <- 3
theta <- seq(4,8,length.out = 1000)
r.sm <- rep(1/n, length(theta))
r.fo <- rep(1, length(theta))
r.6 <- (theta - 6)^2
r.td <- (3 - theta/2)^2 + 1/(4*n)

plot(theta, r.6, pch = "", xlab = "theta", ylab = "risk")
lines(theta, r.sm)
lines(theta, r.fo, lty = 2)
lines(theta, r.6, lty = 3)
lines(theta, r.td, lty = 4)
```

a) Under squared error loss, the risk is the mean squared error. Because the sample mean is unbiased (exercise set 6-1, problem 1), the mean squared error of the sample mean is equal to its variance (equation 6.5). The variance of the mean of a sample of n independent observations from a distribution with variance 1 is $1 / n$ (exercise set 6-2, problem 1).

b) Using similar reasoning as in part (a), the risk, in this case, is equal to the variance. The variance of the first observation is 1, so this risk is 1 for all θ.

c) The risk is obtained by plugging the value of the estimator—in this case, 6—into the loss function. The risk is then $R\left(\theta, \breve{\theta}\right) = (6 - \theta)^2$.

d) The bias of the estimator is

$$\mathrm{B}\left(\dot{\theta}\right) = \mathrm{E}\left(\dot{\theta}\right) - \theta = \mathrm{E}\left(\frac{1}{2n}\sum_{i=1}^{n} X_i + 3\right) - \theta = 3 - \frac{\theta}{2},$$

and the variance of the estimator is

$$\mathrm{Var}\left(\dot{\theta}\right) = \mathrm{Var}\left(\frac{1}{2n}\sum_{i=1}^{n} X_i + 3\right) = \frac{1}{4n}.$$

Using the hint, we have $R\left(\theta, \dot{\theta}\right) = \mathrm{B}\left(\dot{\theta}\right)^2 + \mathrm{Var}\left(\dot{\theta}\right) = (3 - \theta/2)^2 + 1/(4n)$.

e) The only dominated estimator is the first observation, X_1, which is dominated by the sample mean. Thus, the first observation is inadmissible as an estimator of θ under squared-error loss. The other three estimators are all potentially admissible, as seen on the plot from parts a-d.

f) The sample mean is the only estimator of the four we examined that is a candidate minimax estimator. The maximum risk of the sample mean is $1/n$—in this example, $1/3$. All the other estimators have maximum risks greater than $1/3$. For example, when $\theta = 4$, the risks of all the other estimators in the problem are greater than or equal to 1. This result does not prove that the sample mean is the minimax estimator, but it turns out that in this context, it is. Thus, it is also admissible.

Exercise Set 6-8

1) Here is code to examine the first set of specified parameters:

```
dat <- rnorm.out(1000, 100, 0.001, lambda = 3)
means <- apply(dat,2,mean)
medians <- apply(dat,2,median)
mean(means)
var(means)
mean(medians)
var(medians)
hist(means)
hist(medians)
```

You can examine the other parameter sets by replacing the 0.001 in the above function call with the desired γ and replacing the 3 with the desired Λ. You will notice that when γ and Λ are large, both the median and the mean are biased upward, and they both increase in variance. The median, however, is much less affected by the aberrant observations than the mean is. When γ is small, the median is almost unaffected, regardless of how large Λ is. For example, $\gamma = 0.001$ and $\Lambda = 100$, the median is almost unbiased and its variance has scarcely changed at all. But the variance of the sample mean has increased by a factor of about 10. When the situation is truly awful—$\gamma = 0.2$ and $\Lambda = 100$—the sample median's variance increases by a factor of less than 2 while the sample mean's variance goes haywire, increasing by a factor of more than 1,500. Though both the mean and the median are biased, the sample mean's bias is about 60 times larger than the sample median's.

This exercise demonstrates the median's robustness against outliers. When some of the observations may reflect processes or populations that are not the target of study, the median will continue to give roughly correct answers, but the mean may not. One way to formalize this property is to define a statistic's *breakdown point*: roughly, the proportion of observations from a contaminating distribution required to make the statistic perform arbitrarily badly. The mean has a breakdown point of zero—in principle, a single observation from a different distribution can mess up the sample mean as much as you want, provided that the contaminating distribution is far enough removed from the real target. In contrast, the median has a breakdown point of .5, which is the maximum possible. As long as more than 50% of the data come from the distribution of interest, the sample median will at least be within the range of observations drawn from the correct distribution.

Exercise Set 6-9

1) a) In this scenario, the least-squares estimators are unbiased and consistent. (You will have a chance to prove this in some optional exercises in chapter 9.) At each sample size, the means of the estimates are close to the true values, and the variances of the estimates decrease as the sample size increases.

b) The results for the least-absolute-errors estimators are similar to those in part (a), though the variances are somewhat larger.

c) In this scenario (normally distributed disturbances of constant variance), the least-squares estimators are more efficient than the least-absolute-errors estimators—both sets of estimators appear close to unbiased in the simulations (and they are in fact unbiased), and the variances of the least-squares estimators are smaller at each sample size.

2) a) The cloud of observations is more vertically dispersed when the disturbances are Laplace distributed, but the effect is too subtle to detect reliably just by looking. (Statistical tests like those in the `gvlma` package [see the postlude chapter] are more sensitive than our eyes.)

b) Both sets of estimators appear to be approximately unbiased, and the simulations suggest that they may be consistent. However, the relative efficiency is reversed—now the least-squares estimators are less efficient than the least-absolute-errors estimators. Once again, efficiency is a property of an estimator under a specific model, not of the statistic itself. Under the model here (Laplace-distributed disturbances), the least-absolute-errors line is actually a maximum-likelihood estimator (see chapter 9), which explains its strong performance.

3) a) The command shown shows a cloud of points centered around a line (not drawn) with intercept 3 and slope 1/2. In some trials, there are some points in the lower-right corner that are far removed from the rest of the data. These are outliers both in the sense of being removed from the rest of the data and from being actually created by a different process—their disturbances are from the contaminating distribution.

b) With data contamination / outliers, neither set of estimators is unbiased or consistent. Both tend to produce slope estimates that are too low—the line is being "pulled down" by the outlying points in the lower right. However, the least-absolute-errors estimators are much more robust than the least-squares estimators—they are closer to the true values on average and have lower variance.

4) a-b) In this case, $\mathrm{E}(\tilde{\beta}) = \beta + \gamma\rho$, where ρ is the correlation of X and Z. This means that if either Z has no effect on Y ($\gamma = 0$) or X and Z are uncorrelated ($\rho = 0$), then $\tilde{\beta}$ is unbiased. Otherwise, if the data analysis ignores Z, then the estimate of β is biased, and the size and direction of the bias depend on γ and ρ. This problem is *not* specific to least-squares estimators—model misspecification threatens every estimator. Econometricians call this form of bias "omitted variable bias"—the bias in the estimation of β occurs because Z is omitted from the model being estimated—which is a special case of a statistical ailment called endogeneity. People in other fields often call it "confounding." In words, the problem is that part of the causal effect of Z on Y is being wrongly attributed to X.

c) The omitted variable bias is a problem if we want to interpret the estimate of β causally—for example, if we would like to make claims like, "increasing X by one produces a value of Y that is β units larger." If variables that are correlated with X—sometimes called confounds—have effects on Y independently of X's effects on Y but are excluded from the data analysis, then such causal claims will be incorrect. $\tilde{\beta}$ retains its interpretation as the slope of the least-squares line, and it may be useful for prediction. But it will mislead us if we want to manipulate X to produce a desired change in Y. For example, in the fertilizer example, we estimate $\tilde{\beta} = 0.5$. If confounding is a problem, then increasing a country's fertilizer use by 1 kg/hectare would not increase its cereal yield by 50 kg/hectare. This problem remains even in the largest samples.

You may have been taught that randomized experiments are the best way to infer causation. The formula $\mathrm{E}(\tilde{\beta}) = \beta + \gamma\rho$ is a way to explain and defend this claim.

Consider our example. We expect that cereal yield would be a function of fertilizer application, but it might also depend on many other factors: latitude, crops grown, irrigation, availability of labor and equipment, etc. These confounds can cause omitted variable bias if we fail to account for them.

In an experiment, the treatment (X) is assigned randomly to the units being studied. For example, if we wanted to do a country-level experiment on fertilizer yield, then we would randomly decide how much fertilizer each country ought to apply to its fields. If we assign the levels of fertilizer use randomly, then we ensure that the confounds are each uncorrelated with the level of fertilizer use—that is, we set $\rho = 0$. If $\rho = 0$ for all possible confounds, then the omitted variable bias disappears, and we can get unbiased estimates of β—the causal effect of X on Y—without accounting for confounds in the data analysis.

Exercise Set 7-1

1) a) The standard deviation of the observations is σ. This comes from the fact that the second parameter of the normal distribution is equal to the distribution's variance (equation 5.22).
b) The standard error of the estimate is $\text{SE}(\hat{\theta}_n) = \sigma/\sqrt{n}$. Recall from exercise set 6-2, problem 1, that $\text{Var}(\hat{\theta}_n) = \sigma^2/n$. The standard error follows immediately from this fact and equation 7.1.
c) Plugging in the numbers in the problem gives $\text{SE}(\hat{\theta}_n) = 1/5 = 0.2$.
d) Use the `norm.samps()` function (from exercise set 6-1, problem 2, and reprinted here)

```
norm.samps <- function(mu = 0, sigma = 1, n = 25, nsamps =
10000){
  samps <- rnorm(n*nsamps, mu, sigma)
  samp.mat <- matrix(samps, nrow = nsamps, ncol = n)
  return(samp.mat)
}
```

to draw a set of samples:

```
> s.mat <- norm.samps(0, 1, 25, 10000)
```

Calculate the median of each sample using `apply()` (or using the `for()` loop given in exercise set 6-1, problem 2):

```
> ests.median <- apply(s.mat, 1, median)
```

Use the `sd()` command to estimate the standard error of the sample median. (Or use `sqrt(var())`):

```
> sd(ests.median)
```

You will get an answer of approximately 0.25, which is larger than the standard error of the sample mean.

2) We seek the probability $\text{P}(\hat{\theta}_n - \omega < \theta < \hat{\theta}_n + \omega)$. θ is fixed, and in this problem, so is ω, so this is really a probability statement about $\hat{\theta}_n$, the only random variable in the expression. The statement

$$\hat{\theta}_n - \omega < \theta < \hat{\theta}_n + \omega$$

is equivalent to

$$\theta - \omega < \hat{\theta}_n < \theta + \omega,$$

so $\mathrm{P}\big(\hat{\theta}_n - \omega < \theta < \hat{\theta}_n + \omega\big) = \mathrm{P}\big(\theta - \omega < \hat{\theta}_n < \theta + \omega\big) = \mathrm{P}\big(\hat{\theta}_n < \theta + \omega\big) - \mathrm{P}(\hat{\theta}_n < \theta - \omega)$. Recalling that normally distributed random variables are continuous, we know that $\mathrm{P}\big(\hat{\theta}_n < \theta + \omega\big) = \mathrm{P}(\hat{\theta}_n \leq \theta + \omega)$, and therefore, if the cumulative distribution function of $\hat{\theta}_n$ is written $F_{\hat{\theta}_n}$, then the probability we seek is

$$F_{\hat{\theta}_n}(\theta + \omega) - F_{\hat{\theta}_n}(\theta - \omega).$$

Because $\hat{\theta}_n$ is normally distributed with expectation θ and standard deviation ω, this is the probability that a normally distributed random variable falls within 1 standard deviation of its mean. You can look up this probability in a table or use `R`'s `pnorm()` function, which evaluates the cumulative distribution function of the normal distribution:

```
> pnorm(1)-pnorm(-1)
[1] 0.6826895
```

Thus, the interval $(\hat{\theta}_n - \omega, \hat{\theta}_n + \omega)$ would contain θ about 68% of the time. There is a subtle but important point here, which will be emphasized in the next section: the probability statement is about the *interval*, which is random, and not about θ, which is fixed.

b) Reasoning along similar lines as in part (a), we evaluate the necessary probability in `R` as

```
> pnorm(2)-pnorm(-2)
[1] 0.9544997
```

Thus, the interval $(\hat{\theta}_n - 2\omega, \hat{\theta}_n + 2\omega)$ will contain θ about 95% of the time.

3) The first question is to evaluate $\mathrm{E}\left(\left[\hat{\theta}_a - \hat{\theta}_b\right]^2\right)$. One way to do it is to expand the squared term and use the linearity of expectation to write

$$\mathrm{E}\left(\left[\hat{\theta}_a - \hat{\theta}_b\right]^2\right) = \mathrm{E}\left({\hat{\theta}_a}^2 - 2\hat{\theta}_a\hat{\theta}_b + {\hat{\theta}_b}^2\right) = \mathrm{E}\left({\hat{\theta}_a}^2\right) - 2\mathrm{E}\big(\hat{\theta}_a\hat{\theta}_b\big) + \mathrm{E}\left({\hat{\theta}_b}^2\right).$$

Remember that $\hat{\theta}_a$ and $\hat{\theta}_b$ are independent, which guarantees that $\mathrm{E}\big(\hat{\theta}_a\hat{\theta}_b\big) = \mathrm{E}\big(\hat{\theta}_a\big)\mathrm{E}(\hat{\theta}_b)$ (exercise set 5-2, problem 3c). Because the samples are identically distributed and of equal size, and because the same estimator is applied to each sample, the two estimates are identically distributed. This means that $\mathrm{E}\big(\hat{\theta}_a\big) = \mathrm{E}\big(\hat{\theta}_b\big)$, which implies that $\mathrm{E}\big(\hat{\theta}_a\big)\mathrm{E}\big(\hat{\theta}_b\big) = \mathrm{E}\big(\hat{\theta}_a\big)^2 = \mathrm{E}\big(\hat{\theta}_b\big)^2$, and it also implies that $\mathrm{E}\left({\hat{\theta}_a}^2\right) = \mathrm{E}\left({\hat{\theta}_b}^2\right)$. Combining these insights, we have

$$\mathrm{E}\left(\left[\hat{\theta}_a - \hat{\theta}_b\right]^2\right) = 2\mathrm{E}\left({\hat{\theta}_a}^2\right) - 2\mathrm{E}\big(\hat{\theta}_a\big)^2 = 2\mathrm{Var}\big(\hat{\theta}_a\big) = 2\omega^2.$$

Thus, the expected squared difference between two estimates derived by applying the same estimator to two independent, equally-sizes samples is twice the square of the standard error of the estimate. Similarly, the square root of the expected squared difference is $\sqrt{2}\omega$.

Exercise Set 7-2

1) a) $z_{a/2} \approx 0.674$, which you can verify with the `R` command `qnorm(0.75)`. The following `R` code plots an appropriate picture.

```
a <- 0.5 #parameter: 1 minus the confidence level.

#plot a standard normal distribution
x <- seq(-3.5, 3.5, length.out = 10000)
z.a2 <- qnorm(1 - a/2)
fx <- dnorm(x)
plot(x, fx, type = "l")

#Shade in the appropriate area
x.zs <- seq(-z.a2, z.a2, length.out = 10000)
fx.zs <- dnorm(x.zs)
polygon(c(-z.a2, x.zs, z.a2), c(0, fx.zs, 0), col = "grey",
border = FALSE)
```

The area of the shaded region is $1/2$.

b) By equation 7.3, the confidence interval was constructed as $(\hat{\theta} - \omega z_{a/2}, \hat{\theta} + \omega z_{a/2})$. Thus, $\hat{\theta}$ is at the midpoint between the boundaries of the interval—in this case, $\hat{\theta} = 3$.

c) The lower bound of the interval is 2, and $\hat{\theta} = 3$, giving the equation $3 - \omega z_{a/2} = 2$, which implies that $\omega z_{a/2} = 1$. By part a, $z_{a/2} \approx 0.674$, and it follows that the standard error is $\omega = 1/z_{a/2} \approx 1/0.674 \approx 1.48$.

d) For a 95% confidence interval, we need $z_{0.025}$, which is approximately 1.96. The 95% confidence interval is then $(\hat{\theta} - \omega z_{a/2}, \hat{\theta} + \omega z_{a/2}) \approx (3 - 1.48 * 1.96, 3 + 1.48 * 1.96) \approx (0.09, 5.91)$. This is a larger range than that covered by the 50% confidence interval.

e) $3/1.48 \approx 2.02$. The estimate was about two standard errors away from zero.

f) We need the value of a that solves the equation $\hat{\theta} - \omega z_{a/2} = 3 - 1.48 * z_{a/2} = 0$. That is, we need the value of a that gives $z_{a/2} = 3/1.48 \approx 2.02$. The necessary value of a is thus one minus the probability that a random variable drawn from a normal distribution will fall within 2.02 standard deviations of its expectation. The probability that such a random variable will fall more than 2.02 standard deviations *above* its expectation is $1 - \phi(2.02)$, or in `R`, `1-pnorm(2.02)`. Because the normal distribution is symmetric, the probability that such a random variable will

fall more than 2.02 standard deviations *above* its expectation *or* more than 2.02 standard deviations *below* its expectation is therefore double this quantity, or approximately 0.043. This is the necessary value of a.
g) If $\theta = 0$, then the probability of observing $|\hat{\theta}| > 3$ is the probability of observing a normal random variable more than $(\hat{\theta} - \theta)/\omega = 3/1.48 = 2.02$ standard deviations away from its expectation. By the reasoning in the solution to part (f), this probability is 0.043.

Exercise Set 7-3

1) There are many possible responses. It's a good idea to divide responses into two: first, are there reasons why, if the hypothesis were true, you would not be convinced that the theory is true? Second, are there reasons why, if the hypothesis were false, you would not be convinced that the theory is false? So first, suppose that I had found that Arizonan women were shorter than other women from the US. There are many reasons not to take such a finding as evidence in favor of the theory that smaller people are attracted to hotter climates. The most important is *confounding*. Arizona has many properties besides hot weather—for example, it is dry, it borders Mexico, its population includes a higher-than-average proportion of members of the Church of Jesus Christ of Latter-Day Saints, it contains most of the Navajo Nation, its population is less dense than average, it is a popular destination for retirees, and it has historically been a major center of copper mining. When one chooses to study Arizonans, one does not get a set of people who live in a place that is hotter than average and otherwise just like everywhere else. One instead gets a set of people who live in a place with a slew of unique or unusual properties. Any difference between Arizonans and other people could in principle—though perhaps not always plausibly—be due to any of these properties rather than to the Arizonan heat. That is confounding.

Secondly, a failure to find a significant difference between the heights of Arizonan women and other women does not kill the theory. For one thing, I only sampled 25 women. As we will see soon, a sample of 25 is not large enough to detect small or moderate differences between population means. The assumption that Arizona is a good test case for the effects of a hot climate is crude. Though the major population centers—Phoenix and Tucson—are hot, the northern part of the state is at high altitude and experiences reasonably cold winters. Sampling from across the state thus dilutes the effects of heat, possibly masking effects that would be apparent if only the hottest parts of the state were sampled. Using height—an imperfect proxy of body size—as the measured variable has the same effect. Confounding can also mask true effects in addition to generating spurious ones. It may be that smaller people are genuinely attracted to hotter climates but that some other feature of Arizona has attracted tall people, masking the effect.

This list of problems is not exhaustive but gives a sense of the many difficulties facing any empirical researcher.

2) a) The standard error is the standard deviation divided by the square root of the sample size, or in this case, 1mm.

b) From equation 7.3, a confidence interval is $\left(\hat{\theta} - \omega z_{a/2}, \hat{\theta} + \omega z_{a/2}\right)$, where $\hat{\theta}$ is the estimator, ω is the standard error, and $z_{a/2} = \phi^{-1}(1 - a/2)$, the $(1 - a/2)$th quantile of the standard normal distribution. For a 95% confidence interval, $z_{a/2}$ can be found using the R command qnorm(0.975), and it is equal to 1.96. Thus, because $\omega = 1$, a 95% confidence interval is given by $(\bar{x} - 1.96, \bar{x} + 1.96)$, where $\bar{x}$ is the sample mean.

c) We start by finding the values of $\bar{x}$ that would give two-sided $p = 0.05$. The lower of these values is the same as would give a one-sided p of 0.025. Thus, we want to find the value of $\bar{x}$ that satisfies the equation $\Phi_{100,1}(\bar{x}) = 0.025$, where $\Phi_{100,1}$ is the cumulative distribution function of the Normal(100,1) distribution, or equivalently, $\mathrm{P}(\bar{X} \leq \bar{x} | \mu = 100) = 0.025$. That is, we need $\Phi_{100,1}^{-1}(0.025) = \bar{x}$. In R, we use the command qnorm(0.025, mean = 100, sd = 1) to get $\bar{x} = 98.04$, or $\bar{x} = 100 - 1.96$. By symmetry, the two-sided p is 0.05 if $\bar{x} = 100 + 1.96$. Thus, the two-sided p of a test of the null hypothesis that $\mu = 100$ will be less than 0.05 if either $\bar{x} < 98.04$ or $\bar{x} > 101.96$. Notice that, by the result of part (b), these are exactly the cases in which a 95% confidence interval for μ excludes 100.

d) Here is a function that computes a two-sided p for a test of a mean from a normal sample:

```
twotailed.p.normal <- function(x.bar, mu, stand.err){
  abs.diff <- abs(x.bar - mu)
  2 * pnorm(mu - abs.diff, mean = mu, sd = stand.err)
}
```

This function takes a sample mean (x.bar), the mean under the null hypothesis (mu), and the standard error of the sample mean (stand.err).

To simulate the means of 10,000 samples of size four, we have two options. We can either simulate the samples (here, storing them in a matrix) and take their means:

```
sim.mat <- matrix(rnorm(40000, mean = 100, sd = 2), ncol = 4,
     nrow = 10000)
sim.means <- rowMeans(sim.mat)
```

Or we can simulate the means directly, remembering that the means are normally distributed with expectation 100 and variance 1:

```
sim.means <- rnorm(10000, mean = 100, sd = 1)
```

To get the distribution of ps, we could use a for() loop:

```
ps <- numeric(10000)
for(i in 1:10000){
  ps[i] <- twotailed.p.normal(sim.means[i], 100, 1)
}
```

Or, even better, we could use sapply(), a version of apply() that takes vectors as input:

```
ps <- sapply(sim.means, FUN = twotailed.p.normal, mu = 100,
     stand.err = 1)
```

We can plot the distribution of the p values with `hist(ps)`. It is approximately uniform. To find the proportion of p values less than 0.05, use `mean(ps < 0.05)`. It should be approximately 0.05. Similarly, the proportion of p values less than 0.10 should be about 0.10. This is a good result—it means that when the null hypothesis is true, the test works approximately as advertised.

e) To simulate normal samples of size 4 from a $\text{Normal}(101,4)$ distribution and test the null hypothesis that $\mu = 100$, use the following `R` code:

```
sim.mat <- matrix(rnorm(40000, mean = 101, sd = 2), ncol = 4,
     nrow = 10000)
sim.means <- rowMeans(sim.mat)
ps <- sapply(sim.means, FUN = twotailed.p.normal, mu = 100,
     stand.err = 1)
```

The distribution of p values is no longer uniform—it has a concentration of low p values, representing samples that would be unlikely to be drawn if μ were in fact 100. I find that about 17% of the p values are less than 0.05 and that about 26% are less than 0.10.

f) To simulate normal samples of size 4 from a $\text{Normal}(102,4)$ distribution and test the null hypothesis that $\mu = 100$, use the following `R` code:

```
sim.mat <- matrix(rnorm(40000, mean = 102, sd = 2), ncol = 4,
     nrow = 10000)
sim.means <- rowMeans(sim.mat)
ps <- sapply(sim.means, FUN = twotailed.p.normal, mu = 100,
     stand.err = 1)
```

Again, the distribution of p values shows a concentration of low values, even more pronounced than in part (e). I find that about 51% of the p values are less than 0.05 and that about 64% are less than 0.10.

g) To simulate normal samples of size 16 from a $\text{Normal}(101,4)$ distribution and test the null hypothesis that $\mu = 100$, use the following `R` code:

```
sim.mat <- matrix(rnorm(160000, mean = 101, sd = 2), ncol = 16,
     nrow = 10000)
sim.means <- rowMeans(sim.mat)
ps <- sapply(sim.means, FUN = twotailed.p.normal, mu = 100,
     stand.err = 1/2)
```

Again, the distribution of p values shows a concentration of low values, even more pronounced than in part (e). I find that about 51% of the p values are less than 0.05 and that about 64% are less than 0.10. Notice that these are the same values as in part (f): doubling the difference between the true mean and the mean under the null hypothesis had the same effect on the distribution of p values as quadrupling the sample size. There is a good reason for this—both changes have the effect of doubling the number of standard errors separating the true parameter from the value postulated by the null hypothesis.

Exercise Set 7-4

1) a) The proportion of the time that the null hypothesis is true is

$$\tau = \frac{t_n + f_p}{t_n + f_p + f_n + t_p} = t_n + f_p.$$

The simplification comes from the fact that the denominator $t_n + f_p + f_n + t_p = 1$.

b) The proportion of the time the null hypothesis is false is $\varphi = f_n + t_p$.

c) The power of the test is

$$\pi = \mathrm{P}(R|H_0^C) = \frac{t_p}{t_p + f_n}.$$

By part (b), this quantity is also equal to t_p/φ.

d) In the notation of the table, the false discovery rate is

$$\mathrm{P}(H_0|R) = \frac{f_p}{t_p + f_p}.$$

One way to express the false discovery rate in terms of τ, φ, γ, and π is to use Bayes' theorem. By Bayes' Theorem,

$$\mathrm{P}(H_0|R) = \mathrm{P}(R|H_o)\frac{\mathrm{P}(H_0)}{\mathrm{P}(R)}.$$

$\mathrm{P}(R|H_o) = \gamma$ by the definition in equation 7.4, and $\mathrm{P}(H_0) = t_n + f_p$ by part a. To get the remaining term, use

$$\mathrm{P}(R) = \mathrm{P}(R|H_0)\mathrm{P}(H_0) + \mathrm{P}(R|H_0^C)\mathrm{P}(H_0^C) = \gamma\tau + \pi\varphi.$$

Applying these identities gives

$$\mathrm{P}(H_0|R) = \frac{\gamma\tau}{\gamma\tau + \pi\varphi}.$$

You can verify that this expression is equivalent to $\mathrm{P}(H_0|R) = f_p/(f_p + t_p)$ by replacing τ, φ, γ, and π with their definitions in terms of t_n, f_p, f_n, and t_p. Because $\varphi + \tau = 1$, it is also possible to remove either φ or τ from the expression. For example, we could remove φ by writing

$$\mathrm{P}(H_0|R) = \frac{\gamma\tau}{\gamma\tau + \pi(1-\tau)}.$$

Because τ, γ, and π are all between 0 and 1, the false discovery rate decreases as the power of the test, π, increases. This is one excellent reason for valuing tests with high power to reject false null hypotheses.

e) In the notation of the table, the negative predictive value is

$$\mathrm{P}(H_0|R^C) = \frac{t_n}{t_n + f_n}.$$

Using reasoning parallel to that used in part (d), the negative predictive value is equal to

$$\mathrm{P}(H_0|R^C) = \frac{(1-\gamma)\tau}{(1-\gamma)\tau + (1-\pi)\varphi} = \frac{(1-\gamma)\tau}{(1-\gamma)\tau + (1-\pi)(1-\tau)}.$$

Because τ, γ, and π are all between 0 and 1, the negative predictive value increases as the power of the test, π, increases.

2) a) If the research group adopted the proposed procedure, then they would falsely reject the null hypothesis about 11% of the time, which is more than twice the nominal level of each of their tests. This is true even though each of the individual tests rejects the null hypothesis at the correct rate.

b) Other things equal, increasing the number of measurements increases the probability that at least one of the tests leads to an incorrect rejection of the null hypothesis, also called the type I error rate. Increasing the degree of correlation between the measurements toward 1 tends to decrease the probability that at least one of the tests leads to a rejection of the null hypothesis.

One way to control the familywise error rate is with Bonferroni correction, in which each p value is compared to the value γ/k, where γ is the desired familywise error rate and k is the number of hypothesis tests being conducted.

3) The proposed procedure leads to an incorrect rejection of the null hypothesis about 11-12% of the time, which grows worse with more repeated testing.

```
ps <- serial.testing.sim()

sigs <- ps < .05
colMeans(sigs)
mean(rowMeans(sigs) > 0)
```

Exercise Set 7-5

1) The following block of code provides one way to produce the necessary plot, assuming that the `ps.1sz()` function has been defined:

```
n <- 25
d <- seq(-2, 2, length.out = 101)
pow <- numeric(length(d))
for(i in 1:length(d)){
  pow[i] <- ps.1sz(d[i], n)
}
plot(d, pow, ylim = c(0,1), type = "l", ylab = "Power")
```

2) a) When d and n are large enough, power is near 1, and the estimated effect size from studies that rejected the null hypothesis is approximately correct. But for smaller d and n, the bias produced by the winner's curse can be substantial. For example, when $d = 0.25$ and $n = 25$, the mean estimated effect size in studies that reject the null hypothesis is approximately twice the true value. Here are some possible parameter choices:

```
wc.1sz( .3, 50, .05)
wc.1sz( .5, 50, .05)
wc.1sz( .1, 50, .05)
wc.1sz( .3, 25, .05)
wc.1sz( .3, 50, .01)
```

b) The following code is one way to produce the requested plots:

```
true.d <- 0.3
ns <- seq(5, 200, by = 5)
pows <- numeric(length(ns))
est.ds <- numeric(length(ns))

#Save power and estimated effect sizes.
for(i in 1:length(ns)){
  wc <- wc.1sz(true.d, ns[i])
  est.ds[i] <- wc[2]
  pows[i] <- wc[3]
}

#First Plot: Cursed effect size estimate as a function of sample
#size.
plot(ns, est.ds, type = "l", lty = 2, lwd = 2, ylim = c(0,
max(est.ds)), ylab = "d", xlab = "n")
lines(ns, rep(true.d, length(ns)), lwd = 2)
legend("topright", lwd = c(2,2), lty = c(2,1), legend =
c("Cursed", "True"))
```

```
#Second Plot: Size of the winner's curse effect as a function of
#power.
curse.size <- est.ds - true.d
plot(pows, curse.size, type = "l", lwd = 2, xlab = "Power", ylab
= "Size of Winner's Curse Effect")
```

Exercise Set 8-1

1) a) As the sample size `n` increases, the empirical distribution function matches the true cumulative distribution function increasingly closely. For large sample sizes (say, `n = 10^5`), the empirical distribution function and true cumulative distribution function are not visibly distinguishable in the figure.
b) Here's a sample script to make a similar plot with the exponential distribution:

```
min.x <- 0 # min value to plot
rate.x <- 1 #exponential parameter
max.x <- 5/rate.x #max value to plot
n <- 20 #size of sample for ecdf.

x.vals <- seq(min.x, max.x, length.out = 10000)
Fx <- pexp(x.vals, rate.x)
plot(x.vals, Fx, xlab = "z", ylab = "F(z)", type = "l")

x <- rexp(n, rate.x)
lines(ecdf(x), verticals = TRUE, do.points = FALSE, lty = 2)
```

And here's the Poisson distribution:

```
min.x <- 0 # min value to plot
rate.x <- 5 #Poisson parameter
max.x <- 2.2*rate.x #max value to plot
n <- 20 #size of sample for ecdf.

x.vals <- seq(min.x, max.x, length.out = 10000)
Fx <- ppois(x.vals, rate.x)
plot(x.vals, Fx, xlab = "z", ylab = "F(z)", type = "l")

x <- rpois(n, rate.x)
lines(ecdf(x), verticals = TRUE, do.points = FALSE, lty = 2)
```

The Poisson distribution looks different because it is discrete, so its true cumulative distribution function looks like a step function. You can try this with other distribution families.

2) a) $\mathrm{P}(I_{X_i \le z} = 1)$ is the probability that $X_i \le z$, which is also the value of the cumulative distribution function of X evaluated at z. Thus, $\mathrm{P}\left(I_{X_i \le z} = 1\right) = \mathrm{P}(X_i \le z) = F_X(z)$ (see equation 4.5), and so $I_{X_i \le z}$ is a Bernoulli random variable with parameter $F_X(z)$ (see Table 4-5). Applying

the solution to exercise 1c in exercise set 5-1, $\mathrm{E}\big(I_{X_i \le z}\big) = F_X(z)$. Similarly, by the solution to exercise 4a in exercise set 5-2, $\mathrm{Var}\big(I_{X_i \le z}\big) = [1 - F_X(z)]F_X(z)$. You could also compute the expectation and variance directly without realizing that $I_{X_i \le z}$ is a Bernoulli random variable.

b) By the linearity of expectation (equation 5.4), $\mathrm{E}\big[\widehat{F_n}(z)\big] = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} I_{X_i \le z}\right) = \frac{1}{n} n\mathrm{E}(I_{X_i \le z}) = F_X(z)$. Similarly, by equations 4.8-4.9 and the independence of the $I_{X_i \le z}$, $\mathrm{Var}\big[\widehat{F_n}(z)\big] = \mathrm{Var}\big([1/n]\sum_{i=1}^{n} I_{X_i \le z}\big) = (1/n^2)n\mathrm{Var}\big(I_{X_i \le z}\big) = (1/n)[1 - F_X(z)]F_X(z)$. $\widehat{F_n}(z)$ is an unbiased estimator of $F_X(z)$. The mean squared error is $(1/n)[1 - F_X(z)]F_X(z)$, and so the mean squared error approaches 0 as n increases to infinity. Thus, $\widehat{F_n}(z)$ is a consistent estimator of $F_X(z)$. Alternatively, we could have obtained the consistency result by appealing directly to the weak law of large numbers.
This result proves that the empirical distribution function (weakly) converges *pointwise* to the true cumulative distribution function. There is a stronger result, the Glivenko-Cantelli theorem, which shows both that the empirical cdf converges in a stronger sense (called “almost surely), and that it converges *uniformly*, which roughly means that it converges everywhere.

Exercise Set 8-2

1) a) There are two expressions. Using the identity in equation 5.7, the plug-in estimator is

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Equivalently, we can use the definition of the variance (equation 5.6) directly to get

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

These two expressions are equivalent. The proof is parallel to the one used to prove equation 5.7. Start by expanding $\sum_{i=1}^{n}(X_i - \bar{X})^2$ and breaking up the sum,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2.$$

Combine the last two terms by remembering that $\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{2}{n}\left(\sum_{i=1}^{n} X_i\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2.$$

Multiplying both sides of this equation by $1/n$ shows that the two plug-in estimators of the variance are equal, completing the proof.

b) The plug-in estimator of the standard deviation is the square root of the plug-in estimator of the variance. That is, it is either

$$\tilde{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 + \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}$$

or

$$\tilde{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

c) As with the variance, there are two equivalent expressions for the plug-in estimator of the covariance, depending on which of the expressions in equation 5.15 is used as a basis. The first is

$$\widetilde{\sigma_{X,Y}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}),$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. The second expression is

$$\widetilde{\sigma_{X,Y}} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right).$$

The proof that these two expressions are equal is similar to the proof that the two expressions in part (a) are equal.

d) The correlation is the covariance scaled by the product of the standard deviations of the two variables. We can thus construct expressions for the plug-in estimator of the correlation—often denoted r—from the plug-in estimators of the covariance and standard deviation. For example,

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

This is one of several equivalent expressions, with other versions using different equivalent forms of the plug-in estimators of the covariance and standard deviation.

2) a) Here is one way to do it using a `for()` statement:

```
n <- 5
nsamps <- 100000
vars.pi <- numeric(nsamps)
for(i in 1:nsamps){
  samp <- rnorm(n, 0, 1)
  var.pi <- sum((samp-mean(samp))^2)/length(samp)
  vars.pi[i] <- var.pi
}
mean(vars.pi)
```

Here is another approach that uses `apply()` instead of `for()`. Remember that other things equal, most `R` programmers consider `apply()` statements to be better `R` style than `for()` statements:

```
n <- 5
nsamps <- 100000
x <- rnorm(nsamps*n,0,1)
samps <- matrix(x, nrow = nsamps, ncol = n)
var.pi <- function(vec){
  return(sum((vec-mean(vec))^2)/length(vec))
}
vars.pi <- apply(samps, 1, var.pi)
mean(vars.pi)
```

Executing either block of code gives answers very close to 0.8, not 1.

b) Your simulated answers will be very close to those in the following table:

n	2	3	4	5	6	7	8	9	10
Plug-in variance	1/2	2/3	3/4	4/5	5/6	6/7	7/8	8/9	9/10

If σ^2 is the true variance and $\widetilde{\sigma_n^2}$ is the plug-in estimator of the variance using a sample of n independent observations, then

$$\mathrm{E}\left(\widetilde{\sigma_n^2}\right) = \frac{n-1}{n}\sigma^2.$$

Thus, $\widetilde{\sigma_n^2}$ is biased downward, especially for small sample sizes. (If n is large, then $(n-1)/n$ is close to 1, and the bias is small.) We can obtain an unbiased estimator of σ^2 by multiplying $\widetilde{\sigma_n^2}$ by $n/(n-1)$, making it slightly larger to correct its downward bias. This yields what is called the “sample variance,” s^2:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1},$$

where $\bar{X}$ is the sample mean. The `var()` function in `R` computes the sample variance. Try repeating your simulations using `var()` to confirm that it is unbiased.

The "sample standard deviation" is the square root of the sample variance. Though the sample *variance* is unbiased, the sample *standard deviation* is biased slightly downward, but less biased than the plug-in estimator of the standard deviation.

c) Here is one approach to the proof—there are many. We start by taking the expectation of the plug-in estimator and making the simplifications suggested by the linearity of expectation (equation 5.4),

$$\mathrm{E}\left(\widetilde{\sigma_n^2}\right) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i - \bar{X})^2.$$

Because all the X_i are identically distributed, $\mathrm{E}(X_i - \bar{X})^2$ will be the same for all i. Thus, we can proceed by identifying $\mathrm{E}(X_1 - \bar{X})^2$. Expanding the expression gives

$$\mathrm{E}(X_1 - \bar{X})^2 = \mathrm{E}(X_1^2 + 2X_1\bar{X} + \bar{X}^2) = \mathrm{E}(X_1^2) - 2\mathrm{E}(X_1\bar{X}) + \mathrm{E}(\bar{X}^2).$$

By equation 5.7, we know that $\mathrm{E}(X_1^2) = \sigma^2 + \mu^2$, where $\mathrm{E}(X_1) = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Similarly, $\mathrm{E}(\bar{X}^2) = \sigma^2/n + \mu^2$, remembering that the variance of the mean of n independent and identically distributed observations is the variance of each observation divided by n, which follows from equations 4.8 and 4.9. The middle term is a little trickier. To identify $\mathrm{E}(X_1\bar{X})$, notice that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ and break this up as $\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}\sum_{i=2}^{n}X_i$—the summation now runs from $i = 2$ to n. Now we work as follows:

$$\mathrm{E}(X_1\bar{X}) = \mathrm{E}\left[X_1\left(\frac{1}{n}X_1 + \frac{1}{n}\sum_{i=2}^{n}X_i\right)\right] = \frac{1}{n}\mathrm{E}(X_1^2) + \frac{1}{n}\mathrm{E}\left(X_1\sum_{i=2}^{n}X_i\right).$$

We already have an expression for $\mathrm{E}(X_1^2)$. For the second term, recall that X_1 is independent of each of the other X_i. If two random variables X and Y are independent, then $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ (see exercise set 5-2, problem 3, part c). Therefore

$$\frac{1}{n}\mathrm{E}\left(X_1\sum_{i=2}^{n}X_i\right) = \frac{1}{n}\mathrm{E}(X_1)\mathrm{E}\left(\sum_{i=2}^{n}X_i\right) = \frac{1}{n}\mu(n-1)\mu = \frac{n-1}{n}\mu^2,$$

meaning that

$$\mathrm{E}(X_1\bar{X}) = \frac{1}{n}(\sigma^2 + \mu^2) + \frac{n-1}{n}\mu^2 = \frac{\sigma^2}{n} + \mu^2.$$

Putting together the expressions for $\mathrm{E}(X_1^2)$, $\mathrm{E}(X_1\bar{X})$, and $\mathrm{E}(\bar{X}^2)$, we have

$$\mathrm{E}(X_1-\bar{X})^2 = \mathrm{E}(X_1^2) - 2\mathrm{E}(X_1\bar{X}) + \mathrm{E}(\bar{X}^2) = \sigma^2+\mu^2 - 2\left(\frac{\sigma^2}{n}+\mu^2\right) + \frac{\sigma^2}{n}+\mu^2 = \frac{n-1}{n}\sigma^2.$$

Because the X_i are independent and identically distributed, $\mathrm{E}(X_1-\bar{X})^2 = \mathrm{E}(X_i-\bar{X})^2$ for all i. Thus,

$$\mathrm{E}\left(\widetilde{\sigma_n^2}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i-\bar{X})^2 = \frac{1}{n}n\mathrm{E}(X_1-\bar{X})^2 = \mathrm{E}(X_1-\bar{X})^2 = \frac{n-1}{n}\sigma^2.$$

This is the expression we wanted to prove. Notice that we did not rely on any distributional assumptions. We only assumed that the observations were independent and identically distributed with finite variance.

3) Here is some `R` code that estimates the requested moments using the sample moments. I used a sample of ten million observations, but you could increase the precision further by taking a larger sample or by aggregating the results from several samples with a larger total size:

```
x <- rnorm(10^7, 0, 1)
mean(x^4)
mean(x^5)
mean(x^6)
mean(x^7)
mean(x^8)
```

The estimates I obtained were close to the true values, which are $\mathrm{E}(X^4)=3$, $\mathrm{E}(X^5)=0$, $\mathrm{E}(X^6)=15$, $\mathrm{E}(X^7)=0$, and $\mathrm{E}(X^8)=105$. Here are some interesting observations that are not important at all to the rest of the book: For odd k, $\mathrm{E}(X^k)=0$. This makes sense because the distribution of X is symmetric around 0, and when k is odd, X^k has the same sign as X. Thus, every positive value of X^k is cancelled by some negative value of X^k. For even k, there is a pattern. Remember that $\mathrm{E}(X^2)=1$. From there, $\mathrm{E}(X^4)=3=1*3$, $\mathrm{E}(X^6)=15=1*3*5$, and $\mathrm{E}(X^8)=105=1*3*5*7$. This pattern continues for all larger even moments.

Exercise Set 8-3

1) a) The following R code simulates a sample of m independent Binomial(n,p) observations and estimates n and p shown in the main text. Increasing the value of m produces estimates closer to the true values. When $m=10^7$, for example, the estimates are very close to the true values.

```
m <- 100
n <- 50
p <- 0.3
```

```
x <- rbinom(m, n, p)
n.est <- ((sum(x)/m)^2)/((sum(x)/m)^2 + sum(x)/m - sum(x^2)/m)
p.est <- ((sum(x)/m)^2 + sum(x)/m - sum(x^2)/m)/(sum(x)/m)
n.est
p.est
```

b) Start by isolating p. Notice that the quotient $\mathrm{E}(X^2)/\mathrm{E}(X)$ contains an isolated p,

$$\frac{\mathrm{E}(X^2)}{\mathrm{E}(X)} = 1 - p - np,$$

so we can get an expression for p if we can get expressions for 1 and np with $\mathrm{E}(X)$ in the denominator. We can do so: $\mathrm{E}(X)/\mathrm{E}(X) = 1$, and $\mathrm{E}(X) = [\mathrm{E}(X)]^2/\mathrm{E}(X) = np$. Thus,

$$p = \frac{[\mathrm{E}(X)]^2 + \mathrm{E}(X) - \mathrm{E}(X^2)}{\mathrm{E}(X)}.$$

To obtain an expression for n, we take the quotient np/p, remembering that $\mathrm{E}(X) = np$ and p is given by the expression above. This gives the expression for n shown in the main text.

2) From exercise set 5-1, problem 1b, the expectation of X if X has a continuous $\mathrm{Uniform}(a, b)$ distribution is $\mathrm{E}(X) = (a + b)/2$. Here, $a = 0$, so $\mathrm{E}(X) = b/2$. Thus, $b = 2\mathrm{E}(X)$, and the method-of-moments estimator is

$$\tilde{b} = \frac{2}{n}\sum_{i=1}^{n} X_i.$$

3) The appropriate expression is

$$\tilde{\beta} = r_{X,Y}\frac{\widetilde{\sigma_Y}}{\widetilde{\sigma_X}},$$

where r_{XY} is the plug-in estimator of the correlation of X and Y, $\widetilde{\sigma_Y}$ is the plug-in estimator of the standard deviation of Y, and $\widetilde{\sigma_X}$ is the plug-in estimator of the standard deviation of X. $\widetilde{\sigma_Y}$ and $\widetilde{\sigma_X}$ can be replaced by the sample standard deviations, s_Y and s_X (see solution to exercise set 6-9, problem 4b). Notice that this solution is parallel to a rearrangement of equation 5.30,

$$\beta = \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}.$$

Exercise Set 8-4

1) Running `wrap.bm()` repeatedly with each of the combinations of n and B values suggested shows that both n and B have to be reasonably large in order for the bootstrap distribution of the sample mean to approximate the true distribution of the sample mean. When n is 5, the bootstrap

distribution does not look normal, and its mean and standard deviation vary widely around the true value, even when B is very large. Similarly, when B is 10, the bootstrap distribution is a poor approximation of the true distribution, even if n is large. For this problem, the bootstrap distribution of the sample mean starts to be a reasonable approximation of the true distribution of the sample mean when $n \geq 20$ and $B \geq 1{,}000$. The approximation continues to improve noticeably as n increases above 20, but the improvement that comes with increasing B above 1,000 is hard to see. The values of n and B that lead to useful answers will vary in different settings. This setting, that of estimating the sampling distribution of the mean of a normal sample, is less demanding than many problems one will see in practice.

2) The bootstrap standard error is imprecise for the midrange with the values of n and B specified in problem 1. Even increasing n to 1,000 still gives variable standard errors. The midrange is sensitive to small changes in the data because it is a function of two values—the maximum and the minimum of the sample. Hence, small differences between the empirical distribution function and the true cumulative distribution function can lead to large differences in the bootstrap vs. sampling distribution of the midrange.

3) Imagine that we have a distribution function $F_X(x)$. Because it is a distribution function, $F_X(x)$ is monotonically increasing in x, its minimum value is zero, and its maximum value is 1. (A true cumulative distribution function may only asymptote toward 0 and 1 without ever reaching them, but an empirical distribution function actually reaches 0 and 1.)

Graphically, drawing one observation from a distribution function is equivalent to the following procedure:

i) Sample a $\text{Uniform}(0,1)$ random number $Y = y$.
ii) Draw a horizontal line with height equal to y.
iii) Where the horizontal line with height y intersects $F_X(x)$, draw a vertical line down to the axis.
iv) The value on the horizontal axis immediately below the intersection of $F_X(x)$ and the horizontal line with height y is a sample from the distribution function $F_X(x)$.

To get more independent samples from $F_X(x)$, one repeats this procedure with independent $\text{Uniform}(0,1)$ random numbers.

Suppose we construct an empirical distribution function for a set of observations $x_1, x_2, \dots, x_n$. The empirical distribution function is horizontal everywhere except at points directly above the observations $x_1, x_2, \dots, x_n$, where there will be vertical line segments of length $1/n$. (If there are sets of observations with equal value among $x_1, x_2, \dots, x_n$, then some of the vertical sections with have length k/n, where k is the number of observations that share the same value.) This means that, for every uniform random number $Y = y$ drawn in step (i), there is a $1/n$ probability that the horizontal line at height y intersects the empirical distribution function in the vertical strip associated with each observation. If the horizontal line at height y intersects the empirical distribution function in the vertical strip associated with a given observation, then a value equal to that observation is added to the bootstrap sample.

Thus, each time we sample from the empirical distribution function, every observation from the original dataset has a $1/n$ probability of being copied into the bootstrap sample, regardless of whether it has been chosen before. This is equivalent to sampling from the original sample with replacement.

Exercise Set 8-5

1) The first step is to choose a test statistic. One sensible choice is the difference in mean wheat yield between the fields that received substance Z and the fields that did not. We might also choose the difference in median wheat yields, or something else.

The next step is to identify a way of permuting the data. We can lean on what we have already done. Suppose that Y represents a field's wheat yield. Further assume that $Z = 1$ if a field is randomly assigned to receive substance Z and that $Z = 0$ if the field is randomly assigned not to receive it. Then one model for the data is

$$Y_i = \alpha + \beta Z_i + \epsilon_i,$$

where α and β are constants and ϵ is a random variable with $\mathrm{E}(\epsilon) = 0$. The subscript i identifies a particular field, and all fields are assumed to be independent of each other, meaning that the ϵ_i terms are independent. This is exactly the model we developed in the main text, with Z in place of X. We can therefore test it in a similar way, shuffling the labels Z_i independently of the wheat yields Y_i. If $\beta = 0$, then every such permutation leads to a hypothetical dataset that is exactly as probable as the original data. For every permutation we try, we compute the mean difference in wheat yield between fields associated with $Z = 1$ and fields associated with a label $Z = 0$. These differences form a permutation distribution, to which we compare the original mean difference we observed.

As to the null hypothesis being tested: this is a tricky question. One is tempted to claim that we are testing the null hypothesis that substance Z does not affect the expected wheat yield. But we are really testing the null hypothesis that substance Z does not affect the *distribution* of the wheat yield—that is, that $\beta = 0$ *and* that nothing about the distribution of ϵ_i depends on Z_i. For example, if substance Z changes the variance of the wheat yield but does not change the expected wheat yield, then our permutation procedure might still reject the null hypothesis with a probability higher than the nominal significance level.

2) To call `sim.perm.B()`, assess the rate at which the null hypothesis is rejected with a significance level of 0.05, and plot a histogram of the permutation p values for $n = 10$ and $\beta = 0$, use the following commands:

```
ps <- sim.perm.B(0, 0, n = 10)
mean(ps < 0.05)
hist(ps)
```

When I ran these simulations, I arrived at the following results. Your exact results may differ slightly.

	$n = 10$	$n = 50$	$n = 100$
$\beta = 0$	.034	0.046	0.052
$\beta = 0.1$	0.086	0.300	0.486
$\beta = 0.2$	0.182	0.760	0.978

The top row of the table, in which $\beta = 0$, is encouraging. When we set the significance level to 0.05, the Type I error rate is in fact approximately 0.05. In the next two rows, we can see, unsurprisingly, that as the sample size increases (left to right), and as the effect size increases ($\beta = 0.2$ vs. $\beta = 0.1$), the power of the permutation test increases.

Exercise Set 9-1

1) This statement is false. In frequentist statistics, θ is not random, so we cannot make probability statements about it. In Bayesian statistics, the value of θ that maximizes $L(\theta)$ is not generally the most probable value of θ given the data, but sometimes it is (see chapter 10). A better statement is "*The value of θ that maximizes $L(\theta)$ is the one that maximizes the probability of obtaining the observed data*." This statement is correct for discrete random variables; for continuous random variables, it could be modified to "*The value of θ that maximizes $L(\theta)$ is the one that maximizes the joint probability density associated with the observed data*."

2) a) The density is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

b) The log-likelihood is the log of the density,

$$l(\mu) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right).$$

This expression can by simplified. First, notice that it is a product and that the log of a product is the sum of the logs of the terms being multiplied,

$$l(\mu) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \ln\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

The second term contains the log of an exponent. Raising e to a power and taking a natural log are inverse operations, so

$$l(\mu) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x-\mu)^2}{2\sigma^2}.$$

We could simplify the first term further, but this is enough for us.

c) Because the two random variables are independent, the joint density function is the product of the two marginal density functions,

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} * \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_2-\mu)^2}{2\sigma^2}}.$$

This expression simplifies to

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{\sigma^2 2\pi}e^{-\frac{(x_1-\mu)^2+(x_2-\mu)^2}{2\sigma^2}}.$$

d) The log-likelihood for two observations is the log of the density of two observations. We could either take the log of the expression in part (c) directly, or we could start with part (b), remembering that the log of a product is the sum of the logs of the terms being multiplied. Either way, we obtain an expression equivalent to

$$l(\mu) = 2\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x_1-\mu)^2}{2\sigma^2} - \frac{(x_2-\mu)^2}{2\sigma^2}.$$

e) Extending the result in part (d), we have

$$l(\mu) = n\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}.$$

Exercise Set 9-2

1) a) The likelihood function is equal to the joint probability mass function. Because the observations are assumed to be independent, their joint probability mass function is the product of their individual probability mass functions. Thus, it is

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}.$$

b) The log-likelihood is the natural log of the likelihood function, which in this case is

$$l(p) = \ln[L(p)] = \ln\left[\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\right] = \sum_{i=1}^{n}\ln[p^{x_i}(1-p)^{1-x_i}]$$
$$= \sum_{i=1}^{n} x_i\ln(p) + (1-x_i)\ln(1-p).$$

c) Here is some `R` code that draws the requested samples and plots the likelihood and log-likelihood functions.

```
#Draw a sample
n <- 10
p <- 0.6
x <- rbinom(n,1,p)

#Given a vector of values for p and a vector of Bernoulli trials
#x, this function computes the
#likelihood for each value of p.
Ln.Bern <- function(p, x){
  k <- sum(x)
  n <- length(x)
  Ln <- numeric(length(p))
  for(i in 1:length(p)){
    Ln[i] <- prod(p[i]^k * (1-p[i])^(n-k))
  }
  return(Ln)
}

#a set of values of p to plot
p <- seq(0.001, 0.999, length.out = 999)

#The likelihood.
Ln <- Ln.Bern(p, x)
#The log-likelihood
ln <- log(Ln)
plot(p, Ln, type = "l")
plot(p, ln, type = "l")
```

d) The maximum-likelihood estimate of p is the mean of the sample, or the proportion of “1” outcomes.

2) a) The likelihood function, remembering that the observations are independent and identically distributed, is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\theta)^2}{2\sigma^2}} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2}.$$

The log-likelihood function is then

$$\begin{aligned} l(\theta) = \ln[L(\theta)] &= n\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 \\ &= n\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^2 - 2x_i\theta + \theta^2). \end{aligned}$$

To maximize the log-likelihood, we take the derivative with respect to θ:

$$\frac{\partial l(\theta)}{\partial \theta} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(\theta - x_i) = -\frac{n}{\sigma^2}(\theta - \bar{x}),$$

where $\bar{x}$ is the sample mean, $\frac{1}{n}\sum_{i=1}^{n} x_i$. Because n and σ^2 are both positive, the only value of θ that sets the derivative equal to 0 is $\theta = \bar{x}$. To check that this solution maximizes, rather than minimizes, the log-likelihood function, we confirm that the second derivative is negative (exercise 2 of exercise set A-1):

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} = -\frac{n}{\sigma^2}.$$

Setting $\theta = \bar{x}$ maximizes the log-likelihood function, and thus the likelihood function. The maximum-likelihood estimator of θ is therefore $\hat{\theta} = \bar{x}$, the sample mean.

b) The natural logs of the observations have a normal distribution, $\ln(Y_i) \sim \mathrm{Normal}(\theta, \sigma^2)$ for all i. Thus, by part a, the maximum-likelihood estimator of θ is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \ln(Y_i).$$

By the functional invariance property of maximum-likelihood estimators, the maximum likelihood estimator of e^{θ} is $e^{\hat{\theta}}$. Notice that this is different from the method-of-moments estimator, which would estimate $e^{\theta} = \mathrm{E}(Y)$ as the sample mean, $\frac{1}{n}\sum_{i=1}^{n} Y_i$.

Exercise Set 9-3

1) To find the expectation of $\hat{\beta}$, start by making the substitution indicated in the problem's hint:

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \mathrm{E}\left(\frac{\sum_{i=1}^{n} x_i Y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}\right) \\ &= \mathrm{E}\left(\frac{\sum_{i=1}^{n} x_i(\alpha + \beta x_i + \epsilon_i) - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n}(\alpha + \beta x_i + \epsilon_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}\right).\end{aligned}$$

Now, expanding the terms in the numerator, using the linearity of expectation (equation 5.4), and noticing that the only random variables are the disturbance terms, we rewrite as

$$\mathrm{E}(\hat{\beta}) = \frac{\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i \, \mathrm{E}(\epsilon_i) - n\alpha \frac{1}{n} \sum_{i=1}^{n} x_i - \beta \frac{1}{n} (\sum_{i=1}^{n} x_i)^2 + \frac{1}{n} \sum_{i=1}^{n} x_i \, \mathrm{E}(\epsilon_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2}.$$

From here, we can make several simplifications. The two α terms in the numerator exactly cancel each other, and the two terms with the expectation of the disturbances ϵ_i disappear because $\mathrm{E}(\epsilon_i) = 0$, leaving

$$\mathrm{E}(\hat{\beta}) = \frac{\beta \sum_{i=1}^{n} x_i^2 - \beta \frac{1}{n} (\sum_{i=1}^{n} x_i)^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2} = \beta \frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} (\sum_{i=1}^{n} x_i)^2} = \beta.$$

Because $\mathrm{E}(\hat{\beta}) = \beta$, $\hat{\beta}$ is an unbiased estimator of β.

We find the expectation of $\hat{\alpha}$ similarly, beginning by making the substitution in the hint,

$$\mathrm{E}(\hat{\alpha}) = \mathrm{E}\left(\frac{\sum_{i=1}^{n} Y_i - \hat{\beta} \; \sum_{i=1}^{n} x_i}{n}\right) = \mathrm{E}\left(\frac{\sum_{i=1}^{n} (\alpha + \beta x_i + \epsilon_i) - \hat{\beta} \; \sum_{i=1}^{n} x_i}{n}\right).$$

Applying equation 5.4 (linearity of expectation) gives

$$\mathrm{E}(\hat{\alpha}) = \frac{n\alpha + \beta \; \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \mathrm{E}(\epsilon_i) - \mathrm{E}(\hat{\beta}) \; \sum_{i=1}^{n} x_i}{n}.$$

Remembering that $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{E}(\hat{\beta}) = \beta$, the expression simplifies to

$$\mathrm{E}(\tilde{\alpha}) = \frac{n\alpha}{n} = \alpha.$$

Because $\mathrm{E}(\hat{\alpha}) = \alpha$, $\hat{\alpha}$ is an unbiased estimator of α.

The unbiasedness of the least-squares estimators is not guaranteed by their status as method-of-moments or maximum-likelihood estimators, so we had to show it directly. Notice also that we did not rely on the assumptions of normality of disturbances or constant variance of the disturbances. We did not even invoke independence of the disturbance terms—we just used $\mathrm{E}(\epsilon_i) = 0$ for all x. Thus, the least-squares estimator is also unbiased under the weaker assumptions used in chapter 8.

2) To identify the variance, start with the substitution indicated in the problem's hint:

$$\begin{aligned}\operatorname{Var}(\hat{\beta}) &= \operatorname{Var}\left(\frac{\sum_{i=1}^n x_i Y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}\right) \\ &= \operatorname{Var}\left(\frac{\sum_{i=1}^n x_i(\alpha+\beta x_i+\epsilon_i) - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n(\alpha+\beta x_i+\epsilon_i)}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}\right).\end{aligned}$$

Because the x_i are not random, many of these terms are constants and thus do not influence the variance (equation 5.8). Dropping the constants leaves

$$\operatorname{Var}(\hat{\beta}) = \operatorname{Var}\left(\frac{\sum_{i=1}^n x_i\epsilon_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n \epsilon_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}\right).$$

In the numerator, we have two functions of the disturbances. Making the substitution $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and moving $\bar{x}$ inside the sum lets us combine them,

$$\operatorname{Var}(\hat{\beta}) = \operatorname{Var}\left(\frac{\sum_{i=1}^n x_i\epsilon_i - \bar{x}\sum_{i=1}^n \epsilon_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}\right) = \operatorname{Var}\left(\frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}\right).$$

The variance of $\hat{\beta}$ is thus the sum of functions of the variance of the disturbances. Because the disturbances are independent, the variance of the sum is the sum of the variances of the individual terms (equation 5.9). Applying this insight and equation 5.8 ($\operatorname{Var}(a + cX) = c^2\operatorname{Var}(X)$) gives

$$\operatorname{Var}(\hat{\beta}) = \frac{1}{\left[\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2\right]^2}\sum_{i=1}^n (x_i - \bar{x})^2\operatorname{Var}(\epsilon_i).$$

For all i, $\operatorname{Var}(\epsilon_i) = \sigma^2$ by assumption. We can therefore pull it out of the sum, leaving

$$\operatorname{Var}(\hat{\beta}) = \sigma^2\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2\right]^2}.$$

The last step is to notice that, by the identity given in the hint, $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2$, the denominator of the fraction on the right is the square of the numerator. Applying this insight yields two equivalent simplified expressions for the variance:

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The second form makes clear that the variance of $\hat{\beta}$ decreases as the number of observations increases. In this proof, we relied on the independence and constant variance of the disturbances, but we did not invoke their normality.

3) Start by writing the likelihood:

$$L(\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y_i-\alpha-\beta x_i)^2}{2\sigma^2}}.$$

We define $v = \sigma^2$ and make the appropriate substitutions, as suggested in the hint:

$$L(v) = \prod_{i=1}^{n} \frac{1}{\sqrt{v}\sqrt{2\pi}} e^{\frac{-(y_i-\alpha-\beta x_i)^2}{2v}}.$$

The log of this expression is the log-likelihood,

$$l(v) = n \ln\left(\frac{1}{\sqrt{v}\sqrt{2\pi}}\right) - \frac{1}{2v}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

We need to take the derivative of the log-likelihood with respect to v. For the first term, notice that by equation 9.3,

$$n \ln\left(\frac{1}{\sqrt{v}\sqrt{2\pi}}\right) = n \ln\left(\frac{1}{\sqrt{v}} * \frac{1}{\sqrt{2\pi}}\right) = n\left[\ln\left(\frac{1}{\sqrt{v}}\right) + \ln\left(\frac{1}{\sqrt{2\pi}}\right)\right] = -n \ln(\sqrt{v}) - n \ln(\sqrt{2\pi}).$$

We make this substitution to take the derivative,

$$\frac{\partial}{\partial v} l(v) = \frac{\partial}{\partial v}\left[-n \ln(\sqrt{v}) - n \ln(\sqrt{2\pi}) - \frac{1}{2v}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right]$$
$$= -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

The first term of the derivative comes from the problem hint. To set the derivative equal to zero, we need to solve the equation

$$0 = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

Adding $n/(2v)$ to both sides and then multiplying both sides by $2v^2/n$ gives the unique solution,

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

We omit the step of showing that the second derivative of the log-likelihood is negative. The maximum-likelihood estimator of σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2,$$

This is what we would do if we *knew* α and β already and only needed to estimate σ^2. This situation is rare in practice. When α and β are unknown, we replace them with their maximum-likelihood estimators, and the maximum-likelihood estimate of σ^2 becomes

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2,$$

or the average of the squared line errors. However, if α and β are unknown and must be estimated, then the expectation of the maximum-likelihood estimator of the disturbance variance is

$$\mathrm{E}\left(\widehat{\sigma^2}\right) = \frac{n-2}{n}\sigma^2.$$

Thus, the maximum likelihood estimator is biased downward. One way to understand this bias is to notice that the maximum-likelihood estimator of σ^2 is directly proportional to the sum of the squared line errors, and $\hat{\alpha}$ and $\hat{\beta}$ are chosen to make the sum of the squared line errors *as small as possible*. (They are the *least-squares* estimates!) Thus, to the extent that our estimates of α and β err, they will err in ways that make the sum of the squared line errors smaller than they would be if the true values of α and β were known.

One unbiased estimator of the variance of the disturbances is

$$\widetilde{\sigma^2} = \frac{n}{n-2}\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2.$$

When n is large, $n/(n-2) \approx 1$, and the two estimators are nearly identical. In practice, the unbiased estimator is used more often.

4) Starting from the definition of $\hat{\beta}$ and making the substitution $Y_i = \alpha + \beta x_i + \epsilon_i$ gives

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i(\alpha + \beta x_i + \epsilon_i) - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n}(\alpha + \beta x_i + \epsilon_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}.$$

Making the substitution $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and collecting the sums in the numerator gives

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i + \epsilon_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}.$$

We split this sum into two components, a random one and a non-random one,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,\epsilon_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}.$$

The first term is fixed, and by (i), it does not affect whether $\hat{\beta}$ is normally distributed. We label it c and ignore it.

$$\hat{\beta} = c + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,\epsilon_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}.$$

The second term on the right is the sum of n independent random variables, each of which is the product of a non-random term and ϵ_i. By (i), each of the individual random variables is normally distributed. Because the individual random variables are independent and normally distributed, (ii) guarantees that their sum is also normally distributed. Thus, $\hat{\beta}$ is normally distributed. In combination, you have proven in exercises 1, 2, and 4 that given the assumptions in this section,

$$\hat{\beta} \sim \text{Normal}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Exercise Set 9-4

1) a) The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!}.$$

The log-likelihood is

$$l(\theta) = \sum_{i=1}^{n}[-\theta + x_i \ln\theta - \ln(x_i!)] = -n\theta + n\bar{x}\ln\theta - \sum_{i=1}^{n}\ln(x_i!),$$

where $\bar{x}$ is the sample mean. Taking the derivative of the log-likelihood with respect to θ gives

$$\frac{\partial}{\partial\theta} l(\theta) = -n + \frac{n\bar{x}}{\theta}.$$

Setting the log-likelihood to 0 gives

$$0 = n\left(\frac{\bar{x}}{\theta} - 1\right) \Rightarrow \ \theta = \bar{x}.$$

Setting $\theta = \bar{x}$ maximizes the likelihood; thus $\hat{\theta} = \bar{x}$ is the maximum-likelihood estimator of θ.

b) $\hat{\theta} = \bar{x}$, the sample mean, is the maximum-likelihood estimator of θ. Each observation averaged in $\bar{x}$ is independent, and because they are distributed as $\mathrm{Poisson}(\theta)$, their variance is θ. By equations 5.8 and 5.9,

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{n\theta}{n^2} = \frac{\theta}{n}.$$

We could estimate the variance of $\hat{\theta}$ by plugging the estimator $\hat{\theta}$ in for θ.

c) Picking up where we left off in part (a), the first derivative of the log-likelihood is

$$\frac{\partial}{\partial\theta} l(\theta) = -n + \frac{n\bar{x}}{\theta}.$$

The second derivative is then

$$\frac{\partial^2}{\partial\theta^2} l(\theta) = -\frac{n\bar{x}}{\theta^2}.$$

The Fisher Information is the negative expectation of

$$-\frac{n\bar{X}}{\theta^2},$$

where $\bar{X}$ is a random variable representing the mean of a sample of independent, identically distributed $\mathrm{Poisson}(\theta)$ random variables. The expectation of $\bar{X}$ is θ, so

$$\mathcal{I}(\theta) = \frac{n}{\theta}.$$

By equation 9.17, the asymptotic variance of $\hat{\theta}$ is

$$\text{Var}(\hat{\theta}) = \frac{\theta}{n},$$

which we would estimate by plugging in $\hat{\theta}$ for θ. The Fisher Information method only gives us the asymptotic variance of $\hat{\theta}$, but we know from the direct method (part b) that this is also the small-sample variance.

Exercise Set 9-5

1) a) The Wald test statistic (equation 9.20) is

$$W \approx W^* = \frac{\hat{\beta} - \beta_0}{\sqrt{\widehat{\text{Var}(\hat{\beta})}}}.$$

The maximum-likelihood estimate $\hat{\beta}$ is the least-squares slope, which we have already computed as $\hat{\beta} = 0.5$. (See, for example, chapter 3.) Because of the null hypothesis specified in the problem, $\beta_0 = 0$. The estimated variance of $\hat{\beta}$ is

$$\widehat{\text{Var}(\hat{\beta})} = \frac{\widetilde{\sigma^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where $\widetilde{\sigma^2}$ is as given in equation 9.14b. Computing this value for the example data gives a result of approximately 0.014. The following `R` function computes the test statistic:

```
#Function to compute Wald statistic for slope in simple
#linear regression.
wald.stat.slr <- function(x, y, B0 = 0){
  n <- length(x)
  #compute MLEs of beta and alpha
  B.hat <- (sum(x*y)-sum(x)*sum(y)/n)/( sum(x^2) - sum(x)^2/n)
  A.hat <- (sum(y) - B.hat*sum(x))/n
  #Compute estimated variance of MLE of beta
  vhat.dists <- sum((y - A.hat - B.hat*x)^2)/(n-2)
  vhat.Bhat <- vhat.dists/sum((x - mean(x))^2)
  #Wald statistic
  wald <- (B.hat - B0)/sqrt(vhat.Bhat)
  return(wald)
}
```

The `lm()` function also computes the Wald statistic, but it labels it “t.” Applying the function to the agricultural data in the running example gives $W^* \approx 4.24$. By equation 9.21, $p = 2\varphi(-|W^*|) \approx 0.00002$. Comparing the test statistic against the appropriate t distribution gives a p value of 0.002 (using the t distribution with 9 degrees of freedom, because there are 11 data points minus two parameters being estimated), which is in close agreement with the permutation test from chapter 8.

2) Under the null hypothesis, the Wald test statistic is distributed as $\text{Normal}(0,1)$. The square of the Wald statistic is thus distributed as the square of a $\text{Normal}(0,1)$ random variable—in other words, it is distributed as $\chi^2(1)$.

3) Modifying the function for simulating permutation-test p values from exercise set 8-5, problem 2, the below function will return Wald-test p values from simulated datasets with independent, normally distributed disturbances (by default). The x values are here drawn from a normal distribution by default:

```
sim.Wald.B <- function(a, b, B0 = 0, n.sim = 1000, var.eps = 1,
n = 50,
                       mu.x = 8, var.x = 4, rdist = rnorm, rx =
rnorm, pfun = pnorm, ...){
  #Initialize variables.
  ps <- numeric(n.sim)
  for(i in 1:n.sim){
    #Simulate data and compute p value.
    dat <- sim.lm(a, b, var.eps, n, mu.x, var.x, rdist = rdist,
rx = rx)
    x <- dat[,1]
    y <- dat[,2]
    #compute MLEs of beta and alpha
    B.hat <- (sum(x*y)-sum(x)*sum(y)/n)/( sum(x^2) - sum(x)^2/n)
    A.hat <- (sum(y) - B.hat*sum(x))/n
    #Compute estimated variance of MLE of beta
    vhat.dists <- sum((y - A.hat - B.hat*x)^2)/(n-2)
    vhat.Bhat <- vhat.dists/sum((x - mean(x))^2)
    #Wald statistic
    wald <- (B.hat - B0)/sqrt(vhat.Bhat)
    ps[i] <- 2*pfun(-abs(wald), ...)
  }
  #Return the p values
  return(ps)
}
```

To call the function, assess the rate at which the null hypothesis is rejected with a significance level of 0.05, and plot a histogram of the permutation p values for $n = 10$ and $\beta = 0$, use the following commands:

```
> ps <- sim.Wald.B(0, 0, n = 10)
> mean(ps < 0.05)
> hist(ps)
```

When I ran these simulations, I arrived at the following results. Your exact results may differ slightly.

	$n = 10$	$n = 50$	$n = 100$
$\beta = 0$	0.088	0.060	0.039
$\beta = 0.1$	0.127	0.317	0.530
$\beta = 0.2$	0.257	0.789	0.956

The top row of the Wald test table, in which $\beta = 0$, is somewhat unsettling. When $n = 10$, the Wald test produces $p < 0.05$ in 9% of the simulations. This problem is attributable to the fact that the variance of the disturbances is unknown and has to be estimated. As mentioned in the main text, the normal null distribution for the Wald statistic depends on the assumption that the variance of the disturbances is known. Increasing n allows for better estimation of the unknown variance, and the problem is ameliorated. One solution—the one adopted by `R`'s `lm()` function—is to compare the Wald statistic to an appropriate t distribution rather than a standard normal distribution. (You can simulate the type I error rate when the t distribution is used for comparison with, for example,

```
mean(sim.Wald.B(0, 0, n = 10, pfun = pt, df = 8) < .05)
```

The `df` parameter should be set to two less than `n`. When comparing with the t distribution, the table looks much like the one for the permutation test.)

For comparison, here are the results I obtained with the permutation test:

	$n = 10$	$n = 50$	$n = 100$
$\beta = 0$	.034	0.046	0.052
$\beta = 0.1$	0.086	0.300	0.486
$\beta = 0.2$	0.182	0.760	0.978

In these simulations, the power of the Wald test is similar to that of the permutation test for $n = 50$ and $n = 100$. At $n = 10$, the Wald test has greater power than the permutation test, but this is of little use—the Wald test is untrustworthy for $n = 10$ (because of the high type I error rate).

Exercise Set 9-6

1) a) There are several ways to show this, but here is one. Recall that the log-likelihood is

$$l(\alpha, \beta) = n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

Under the null hypothesis that $\beta = 0$, the log-likelihood becomes

$$l(\alpha) = n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha)^2 = n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\alpha - \alpha^2).$$

The derivative with respect to α is

$$\frac{\partial l(\alpha)}{\partial \alpha} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(2y_i - 2\alpha) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha) = -\frac{n}{\sigma^2}(\bar{y} - \alpha),$$

Where the last step follows because $\bar{y} = \sum_{i=1}^{n} y_i / n$. Setting the derivative of the log-likelihood to zero maximizes the log-likelihood and is achieved by setting $\alpha = \bar{y}$. Thus, when β is constrained to be zero, $\hat{\alpha} = \bar{y}$.

c) The likelihood-ratio test statistic (equation 9.20) is

$$\Lambda = 2\ln\left(\frac{L(\hat{\theta})}{L(\widehat{\theta_0})}\right) = 2\left(l(\hat{\theta}) - l(\widehat{\theta_0})\right).$$

Here, $l(\hat{\theta})$ is the value of the log-likelihood at its maximum when all parameters are free, and $l(\widehat{\theta_0})$ is the value of the log-likelihood at its maximum assuming that $\beta = 0$. In part (a), you showed that if $\beta = 0$, then $\hat{\alpha} = \bar{y}$. Thus, Λ becomes

$$\Lambda = 2\left[\left(n\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2\right) - \left(n\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\right].$$

Noticing that the natural log terms cancel, as do the 2’s in front and in the denominator of the remaining terms, this simplifies to

$$\Lambda = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2\right],$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum-likelihood estimates of α and β (equations 9.10 and 9.11). The expression can be simplified further, as in part b of problem 2 below. But this version is easy enough to calculate. The below `R` code computes the likelihood-ratio statistic, substituting an estimate of σ^2 (the one given by the estimator in equation 9.14b) for σ^2 itself.

```
lr.stat.slr <- function(x, y){
  n <- length(x)
  #compute MLEs of beta and alpha
  B.hat <- (sum(x*y)-sum(x)*sum(y)/n)/( sum(x^2) - sum(x)^2/n)
  A.hat <- (sum(y) - B.hat*sum(x))/n
  #Compute estimated variance of MLE of beta
  vhat <- sum((y - A.hat - B.hat*x)^2)/(n-2)
  #likelihood-ratio statistic
  lr <- (sum((y - mean(y))^2) - sum((y - A.hat - B.hat*x)^2))
/vhat
```

```
  return(lr)
}
```

Applying the function to the agricultural data in the running example gives $\Lambda^* \approx 17.99$. (We add the asterisk to indicate that this value of Λ has been calculated using an estimate of σ^2.) Because we held exactly one parameter constant—namely, β—we compare Λ^* to a $\chi^2(1)$ distribution. The p value of 0.00002 is found using `pchisq(17.99, 1)` in `R`.

c) The p values from part (b) and from problem 1 of Exercise Set 9-5 are identical. Moreover, the test statistic from part (b) is the square of the test statistic from problem 1 of Exercise Set 9-5. This relationship explains the identity of the p values. The test statistic in part (b) is compared to a $\chi^2(1)$ distribution, which is the distribution of the square of a single draw from the standard normal distribution, and we compare the test statistic from problem 1 of Exercise Set 9-5 to a standard normal distribution. The agreement between the two tests is not a coincidence—in the simple linear regression case, the Wald test and the likelihood-ratio test are equivalent, even for small sample sizes.

2) a) Starting from the hint, we write

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}([Y_i - \hat{Y}_i] + [\hat{Y}_i - \bar{Y}])^2.$$

Expanding the squared term on the right gives

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}).$$

Notice that this statement implies that the claim we want to show is true if and only if

$$2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0.$$

We now set out to prove that, in fact, $2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$. Start by replacing $\hat{Y}_i$ with its definition, $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$, giving

$$2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 2\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)(\hat{\alpha} + \hat{\beta}x_i - \bar{Y})$$

Now replace $\hat{\alpha}$ with its definition, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$, giving

$$2\sum_{i=1}^{n}(Y_i - [\bar{Y} - \hat{\beta}\bar{x}] - \hat{\beta}x_i)([\bar{Y} - \hat{\beta}\bar{x}] + \hat{\beta}x_i - \bar{Y}).$$

Cancelling and grouping terms as appropriate gives

$$2\sum_{i=1}^{n}(Y_i - \bar{Y} - \hat{\beta}[x_i - \bar{x}])(\hat{\beta}[x_i - \bar{x}])$$

Split this into two sums by distributing the term on the right to give

$$2\hat{\beta}\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) - 2\hat{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

We're getting close. Now replace $\hat{\beta}$ with its definition, $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}[(x_i-\bar{x})^2]}$, to give

$$2\frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}[(x_i - \bar{x})^2]}\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) - 2\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}[(x_i - \bar{x})^2]}\right]^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

We're home. Carrying through the multiplications gives

$$2\frac{[\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})]^2}{\sum_{i=1}^{n}[(x_i - \bar{x})^2]}(1 - 1) = 0.$$

Thus, by showing that

$$2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0,$$

We have also proven the claim we set out to show,

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2.$$

b) Start by writing Λ for simple linear regression. We are considering Λ as a random variable, so we use capital Y,

$$\Lambda = 2\lambda(\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \alpha_0 - \beta_0 x_i)^2 - \left(Y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2$$

Under the null hypothesis given in the problem, $\beta_0 = 0$ and $\alpha_0 = \bar{Y}$ (see exercise 1, part b). So the statistic becomes

$$\Lambda = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \left(Y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2.$$

Now notice that the terms in this expression appear in the statement you proved in part (a), the ANOVA identity. Specifically, $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$, and so by rearranging the ANOVA identity,

$$\Lambda = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \left(Y_i - \hat{Y}_i\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2.$$

Now switch $\hat{Y}_i$ back to its definition, $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$, and recall that $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$ to obtain

$$\Lambda = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\hat{\alpha} + \hat{\beta}x_i - \bar{Y}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\bar{Y} - \hat{\beta}\bar{x} + \hat{\beta}x_i - \bar{Y}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\hat{\beta}[x_i - \bar{x}]\right)^2.$$

Pulling $\hat{\beta}^2$ outside the sum gives

$$\Lambda = \frac{\hat{\beta}^2}{\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

We already know that if $\beta = 0$, then $\hat{\beta}\sim\text{Normal}(0, \sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2)$ (see equation 9.13). Thus,

$$\frac{\hat{\beta}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sigma}\sim\text{Normal}(0,1),$$

and the square of this quantity—in other words, Λ—therefore has a $\chi^2(1)$ distribution.

Exercise Set 10-1

1) To begin, let’s set some parameters and simulate data we’ll use in parts (a-c).

```
n <- 20
true.mean <- 2
known.sd <- 1
prior.mean <- 0
prior.sd <- 1

set.seed(8675309)
z <- rnorm(n,true.mean,known.sd)
```

a) The mean of my sample is 2.15, fairly close to the true expectation of 2. Here are functions to compute the posterior mean and variance using equations 10.4 and 10.5:

```
post.conj.norm.norm <- function(z, known.sd, prior.mean,
prior.sd){
  xbar <- mean(z)
  post.expec <- (prior.mean / prior.sd^2 + xbar*length(z) /
        known.sd^2)/(1 /   prior.sd^2 + length(z) / known.sd^2)
  post.var <- 1 / (1 /   prior.sd^2 + length(z) / known.sd^2)
  list("posterior.expectation" = post.expec,
"posterior.variance" = post.var)
}
```

With my simulated data, using it gives

```
> post.conj.norm.norm(z, known.sd, prior.mean, prior.sd)
$posterior.expectation
[1] 2.046966

$posterior.variance
[1] 0.04761905
```

The posterior standard deviation is the square root of the variance, approximately 0.218 here.

b) After installing and loading the `MCMCpack` package, use

```
mn.mod <- MCnormalnormal(z, sigma2 = 1, mu0 = prior.mean, tau20
= prior.sd^2, mc = 10000)
```

Calling `summary(mn.mod)` reveals results extremely similar to part (a).

c) Using the `reject.samp.norm()` function from problem 2(d) gives, with my simulated data,

```
> rsamps <- reject.samp.norm(z, known.sd, prior.mean, prior.sd)
> mean(rsamps)
[1] 2.04764
> sd(rsamps)
[1] 0.2194684
```

These results are extremely similar to those obtained in parts (a) and (b). The small differences are due to stochasticity inherent in MCMC and rejection sampling, which are Monte Carlo procedures.

2) a) The ratio is equal to the likelihood, $f_D(d|\theta)f_\theta(\theta)/f_\theta(\theta) = f_D(d|\theta) = L(\theta)$.

b) If c times the unscaled posterior is equal to the prior, then

$$c\mathcal{L}(\theta)f_\theta(\theta) = f_\theta(\theta),$$

which implies $c = 1/L(\theta)$. Similarly, if c times the unscaled posterior is less than the prior, then $c < 1/L(\theta)$. If we set $c = 1/\max\big(L(\theta)\big)$, then $cL(\theta)f_\theta(\theta) = f_\theta(\theta)$ for the value(s) of θ that maximize the likelihood, and $cL(\theta)f_\theta(\theta) < f_\theta(\theta)$ for all other values of θ.

c) The only step that changes from the algorithm given in the text is the second one. Instead of computing m as the likelihood, we can compute m as the likelihood divided by the maximum possible value of the likelihood, which guarantees that m is between 0 and 1 but also that it can, in least in principle, be as large as 1. This usually means that a much larger proportion of the samples can be accepted, which increases the efficiency of the algorithm.

d) The below two functions are one way (of many) to do it. Rather than multiplying the densities associated with each datum in the input vector `z`, we take the logs of the densities, sum them, and then exponentiate the result. This is useful because `R` (and other programs) can lose precision when forced to perform arithmetic with extremely small numbers—using the logs is a trick to get around the numerical instability that can result.

```
#Get 1 sample under rejection sampling from normal with known sd.
#z is a vector of data.
get.1.samp.norm <- function(z, known.sd = 1, prior.mn = 0, prior.sd =
1){
  accepted <- FALSE
  max.like <- exp(sum(log(dnorm(z, mean = mean(z), sd = known.sd))))
  while(accepted == FALSE){
    cand <- rnorm(1, prior.mn, prior.sd)
    like <- exp(sum(log(dnorm(z, mean = cand, sd = known.sd))))
    crit <- like / max.like
    xunif <- runif(1,0,1)
    if(xunif <= crit){accepted <- TRUE}
  }
  cand
}

#Wrapper for get.1.samp.norm() that gets rejection sample from
posterior of desired size.
reject.samp.norm <- function(z, known.sd = 1, prior.mn = 0, prior.sd =
1, nsamps = 10000){
  samps <- numeric(nsamps)
  for(i in seq_along(samps)){
    samps[i] <- get.1.samp.norm(z, known.sd, prior.mn, prior.sd)
  }
  samps
}
```

Exercise Set 10-2

1) a) Obtain the least-squares estimates of the intercept and slope—3 and ½, respectively—with

```
y <- anscombe$y1
x <- anscombe$x1
reg.ml <- lm(y~x)
summary(reg.ml)
```

b) Code to fit the model with all 9 possible priors is:

```
reg11 <- MCMCregress(y ~ x, b0 = c(0,0), B0 = 0.0001)
reg12 <- MCMCregress(y ~ x, b0 = c(0,0), B0 = 1)
reg13 <- MCMCregress(y ~ x, b0 = c(0,0), B0 = 100)
reg21 <- MCMCregress(y ~ x, b0 = c(3,0), B0 = 0.0001)
reg22 <- MCMCregress(y ~ x, b0 = c(3,0), B0 = 1)
reg23 <- MCMCregress(y ~ x, b0 = c(3,0), B0 = 100)
reg31 <- MCMCregress(y ~ x, b0 = c(10,-5), B0 = 0.0001)
reg32 <- MCMCregress(y ~ x, b0 = c(10,-5), B0 = 1)
reg33 <- MCMCregress(y ~ x, b0 = c(10,-5), B0 = 100)
```

When prior precision is low—meaning that prior variance is high—then the prior means are not especially important; these three choices lead to similar conclusions. If the prior precision is higher, then the prior means matter much more, and estimates are generally close to the prior means.

2) a) Squared-error loss implies that the expected loss given a particular choice of θ_0 is $\int_{-\infty}^{\infty}(\theta - \theta_0)^2 f_\theta(\theta)d\theta$. Expanding the square term inside the integral gives $\int_{-\infty}^{\infty}\left(\theta^2 - 2\theta\theta_0 + {\theta_0}^2\right)f_\theta(\theta)d\theta$. Splitting this into three integrals by distributing the $f_\theta(\theta)d\theta$ term and then applying the definition of expectation and the fact that density functions integrate to 1 gives

$$\int_{-\infty}^{\infty} \theta^2 f_{\theta|D}(\theta)d\theta - 2\theta_0 \int_{-\infty}^{\infty} \theta f_{\theta|D}(\theta)d\theta + {\theta_0}^2 \int_{-\infty}^{\infty} f_{\theta|D}(\theta)d\theta = \mathrm{E}(\theta^2|D) - 2\theta_0\mathrm{E}(\theta|D) + {\theta_0}^2$$

To minimize the loss, $\mathrm{E}(\theta^2|D) - 2\theta_0\mathrm{E}(\theta|D) + {\theta_0}^2$, we take the derivative with respect to θ_0, giving

$$-2\mathrm{E}(\theta|D) + 2\theta_0.$$

Setting this derivative equal to 0 gives

$$\theta_0 = \mathrm{E}(\theta|D),$$

which is the value of θ_0 that minimizes the expectation of the loss function, and thus the Bayes estimator.

b) Under 0-1 loss, if we choose θ_i, then the loss is 0 if $\theta_i = \theta$ and the loss is 1 if $\theta_i \neq \theta$. The expected loss if we choose θ_i, given the data, is thus

$$1 * \mathrm{P}(\theta_i \neq \theta | D = d) + 0 * \mathrm{P}(\theta_i = \theta | D = d) = 1 - \mathrm{P}(\theta_i = \theta | D = d).$$

Minimizing the expected loss is equivalent to maximizing the posterior probability $\mathrm{P}(\theta_i = \theta | D = d)$. By definition, the θ_i with the highest posterior probability is the posterior mode, and so the posterior mode minimizes the expected loss.

c) Under absolute-error loss, the expected loss is $\int_{-\infty}^{\infty} |\theta - \theta_0| f_{\theta|D}(\theta) d\theta$. To minimize the expected loss, we want to set

$$\frac{\partial}{\partial \theta_0} \int_{-\infty}^{\infty} |\theta - \theta_0| f_{\theta|D}(\theta) d\theta = 0.$$

It is hard to make progress in this form because the function $|\theta - \theta_0|$ has an undifferentiable point at $\theta = \theta_0$. We will respond by breaking the integral into two pieces,

$$\frac{\partial}{\partial \theta_0} \left[\int_{-\infty}^{\theta_0} (\theta_0 - \theta) f_{\theta|D}(\theta) d\theta + \int_{\theta_0}^{\infty} (\theta - \theta_0) f_{\theta|D}(\theta) d\theta \right].$$

The Leibniz integral rule holds that we can change the order of integration and differentiation, using

$$\frac{\partial}{\partial t} \left(\int_{a(t)}^{b(t)} g(x,t) dx \right) = \int_{a(t)}^{b(t)} \frac{\partial g(x,t)}{\partial t} dx + g(b(t),t) b'(t) - g(a(t),t) a'(t).$$

Replacing t with θ_0, $b(t)$ with $b(\theta_0) = \theta_0$, $a(t)$ with $a(\theta_0) = -\infty$, x with θ, and $g(x,t)$ with $g(\theta, \theta_0) = (\theta_0 - \theta) f_{\theta|D}(\theta)$ gives

$$\begin{aligned} &\frac{\partial}{\partial \theta_0} \int_{-\infty}^{\theta_0} (\theta_0 - \theta) f_{\theta|D}(\theta) d\theta \\ &\qquad = \int_{-\infty}^{\theta_0} f_{\theta|D}(\theta) d\theta + (\theta_0 - \theta_0) f_{\theta|D}(\theta_0) * 1 - \lim_{\theta \to -\infty} (\theta_0 - \theta) f_{\theta|D}(\theta) * 0 \\ &\qquad = \int_{-\infty}^{\theta_0} f_{\theta|D}(\theta) d\theta = \mathrm{P}(\theta < \theta_0 | D). \end{aligned}$$

Similarly, replacing t with θ_0, $b(t)$ with $b(\theta_0) = \infty$, $a(t)$ with $a(\theta_0) = \theta_0$, x with θ, and $g(x,t)$ with $g(\theta, \theta_0) = (\theta - \theta_0) f_{\theta|D}(\theta)$ gives

$$\frac{\partial}{\partial\theta_0}\int_{\theta_0}^{\infty}(\theta-\theta_0)f_{\theta|D}(\theta)d\theta$$

$$= -\int_{\theta_0}^{\infty} f_{\theta|D}(\theta)d\theta + \lim_{\theta\to\infty}(\theta-\theta_0)f_{\theta|D}(\theta)*0-(\theta_0-\theta_0)f_{\theta|D}(\theta_0)*1$$

$$= -\int_{\theta_0}^{\infty} f_{\theta|D}(\theta)d\theta = \ -\mathrm{P}(\theta>\theta_0|D)$$

Summing these two derivatives, the derivative of the expected absolute error loss with respect to θ_0 is

$$\frac{\partial}{\partial\theta_0}\int_{-\infty}^{\infty}|\theta-\theta_0|f_{\theta}(\theta)d\theta = \ \mathrm{P}(\theta<\theta_0|D)- \ \mathrm{P}(\theta>\theta_0|D).$$

Setting the derivative equal to zero gives

$$\mathrm{P}(\theta<\theta_0|D) = \ \mathrm{P}(\theta>\theta_0|D),$$

which, by definition of the median, is satisfied when θ_0 is the median of the posterior distribution.

Exercise Set 10-3

1) Assuming you have already fit the models in the previous exercise set, problem 1, you can get the quantile interval by looking at the 2.5th and 97.5th quantiles of the slope parameter in the summary output. For these models, the quantile and highest-posterior-density intervals are similar. When the prior precision is very low (i.e. its variance is very high), the credible interval largely agrees with the frequentist confidence intervals you have already calculated. If the precision is higher, then the credible interval is pulled toward the prior mean.

Exercise Set 10-4

1) a) Here is the code:

```
reg0 <- MCMCregress(y~1, b0 = 0, B0 = 1/100, c0 = 0.001, d0 =
0.001, marginal.likelihood = "Laplace")
reg1 <- MCMCregress(y~x, b0 = 0, B0 = 1/100, c0 = 0.001, d0 =
0.001, marginal.likelihood = "Laplace")
summary(reg0)
summary(reg1)
summary(BayesFactor(reg1, reg0))
```

Some relevant outputs are:
Intercept under H_1: posterior mean of 2.97, 95% credible interval [0.48, 5.52]

Slope under H_1: posterior mean of 0.50, 95% credible interval [0.23, 0.77]
Bayes factor B_{10}: 3.26. By Kass & Raftery's scale, this is positive evidence for H_1 over H_0.

b) Here is the code:

```
reg01 <- MCMCregress(y~1, b0 = 0, B0 = 1/16, c0 = 0.001, d0 =
0.001, marginal.likelihood = "Laplace")
reg11 <- MCMCregress(y~x, b0 = 0, B0 = 1/16, c0 = 0.001, d0 =
0.001, marginal.likelihood = "Laplace")
summary(reg01)
summary(reg11)
summary(BayesFactor(reg11, reg01))
```

Some relevant outputs are:
Intercept under H_1: posterior mean of 2.76, 95% credible interval [0.33, 5.13]
Slope under H_1: posterior mean of 0.52, 95% credible interval [0.27, 0.78]
Bayes factor B_{10}: 26.9. By Kass & Raftery's scale, this is strong evidence for H_1 over H_0.

c) Here is the code:

```
reg02 <- MCMCregress(y~1, b0 = 0, B0 = 1/10000, c0 = 0.001, d0 =
0.001, marginal.likelihood = "Laplace")
reg12 <- MCMCregress(y~x, b0 = 0, B0 = 1/10000, c0 = 0.001, d0 =
0.001, marginal.likelihood = "Laplace")
summary(reg02)
summary(reg12)
summary(BayesFactor(reg12, reg02))
```

Some relevant outputs are:
Intercept under H_1: posterior mean of 3.01, 95% credible interval [0.51, 5.61]
Slope under H_1: posterior mean of 0.50, 95% credible interval [0.22, 0.76]
Bayes factor B_{10}: 0.26. By Kass & Raftery's scale, this is positive evidence for H_0 over H_1—notice that we have switched from having the data support H_1 to having them support H_0.

d) In this case, changing the prior precision/variance had only a small effect on the point estimates and credible intervals obtained. In contrast, the Bayes factors changed in consequential ways. With intermediate prior variance, we had relatively weak but "positive" support for the model with the slope included. After decreasing the prior variance, that support became much stronger. But increasing the prior variance causes it to reverse, and the intercept-only model becomes supported over the model with the slope included. One has to be careful about prior specification when working with Bayes factors.

Exercise Set Postlude-1

1) See the set of plots for the first dataset in the quartet with

```
lm.fit <- lm(y1 ~ x1, data = anscombe)
```

```
plot(lm.fit)
```

and press enter to cycle through the plots. For the other datasets in the quartet, change the variable names accordingly.

Exercise Set Postlude-2

1) Once the package is installed and loaded, fit the four models and run the diagnostics with

```
lm.fit1 <- lm(y1 ~ x1, data = anscombe)
gvlma(lm.fit1)
lm.fit2 <- lm(y2 ~ x2, data = anscombe)
gvlma(lm.fit2)
lm.fit3 <- lm(y3 ~ x3, data = anscombe)
gvlma(lm.fit3)
lm.fit4 <- lm(y4 ~ x4, data = anscombe)
gvlma(lm.fit4)
```

As expected from the plot, tests using the first dataset reveal no clear reasons for concern. But remember that the samples size in this example is very small, which means that power to detect deviations is low.

The second model shows that the “link function” test—which is intended to detect departures from linearity—returns a low *p* value. This makes sense: the data clearly fit a curve and not a line.

The third dataset reveals trouble with the normality assumption. This isn’t as informative as looking at the plot, which would likely lead us to investigate whether the single point falling off the line is some sort of error. But at least the test has returned an alarm when it should.

The fourth result is more disquieting—the tests detect no problems with the assumptions, even though the plot suggests something is wrong.

In short, the tests are useful in conjunction with the plots, but they do not replace them.

Exercise Set Postlude-3

1) a) Code in text.
b) Here’s some code:

```
resid.am.hp <- lm(am ~ hp, data = mtcars)$residuals #1
resid.mpg.hp <- lm(mpg ~ hp, data = mtcars)$residuals #2
resid.mod.hp <- lm(resid.mpg.hp ~ resid.am.hp) #3
summary(resid.mod.hp)
```

The estimated slope for the residuals of transmission type here is the same as the estimated slope for transmission type in part (a). (It is *not* the same as the estimated slope as you would get from a simple linear regression of transmission type and miles per gallon. The reason for this difference is the association between transmission type and horsepower, which matters in the multiple regression. The standard errors and test statistics are differ from the multiple regression case, for a similar reason.)

As to what's going on, it's possible to state a concise proof that this will happen using matrix algebra or a much more verbose one using approaches similar to the ones you used in chapter 3. Conceptually, though, think of multiple regression coefficient estimates as assessing the association between an independent variable and the dependent variable when the other independent variables are held constant. In this case, by residualizing both weight and miles per gallon on horsepower, we are obtaining versions of these two variables whose association with horsepower has been "removed."

c)
```
resid.hp.am <- lm(hp ~ am, data = mtcars)$residuals #1
resid.mpg.am <- lm(mpg ~ am, data = mtcars)$residuals #2
resid.mod.am <- lm(resid.mpg.am ~ resid.hp.am) #3
summary(resid.mod.am)
```

d)
```
mean(mtcars$mpg) - mean(mtcars$am)*5.277 - mean(mtcars$hp)*(-0.05888)
```

2) In (a) and (b), the t-statistics and associated p values from the two pairs of analyses are equal. In part (c), the F statistics and associated p values are equal.

Exercise Set Postlude-4

1) `glm()` is for "generalized linear model." Once the car package is loaded, you can use

```
probit.fit <- glm(volunteer ~ extraversion + neuroticism + sex,
data = Cowles, family = binomial("probit"))
summary(probit.fit)
```

to get the estimates. To fit the logistic model instead, you would use `family = binomial("logit")`

2) a) Yes, they are consistent. Consistent with this, setting the n to 10,000 or 100,000 gives estimates very close to the coefficients specified in the simulation.
b) In the presence of heteroscedasticity in the model for the latent variables, the estimators for the model coefficients are inconsistent—they converge on the wrong numbers. The degree to which they're wrong increases with the severity of the heteroscedasticity.

Exercise Set Postlude-5

1) For concreteness, we will compute the correlation of the first and second observation from country 1, Y_{11} and Y_{12}, but the argument applies to any pair of observations from the same

country. First, using the hint, adding non-random terms to a pair of random variables does not affect their correlation, so

$$\mathrm{Cor}(Y_{11}, Y_{12}) = \mathrm{Cor}(\alpha + \beta x_{11} + \mu_1 + \epsilon_{11}, \alpha + \beta x_{12} + \mu_1 + \epsilon_{12}) = \mathrm{Cor}(\mu_1 + \epsilon_{11}, \mu_1 + \epsilon_{12}).$$

Next, by the definition of correlation,

$$\mathrm{Cor}(\mu_1 + \epsilon_{11}, \mu_1 + \epsilon_{12}) = \frac{\mathrm{E}([\mu_1 + \epsilon_{11}][\mu_1 + \epsilon_{12}]) - \mathrm{E}(\mu_1 + \epsilon_{11})\mathrm{E}(\mu_1 + \epsilon_{12})}{\sqrt{\mathrm{Var}(\mu_1 + \epsilon_{11})}\sqrt{\mathrm{Var}(\mu_1 + \epsilon_{12})}}.$$

By the linearity of expectation and the fact that $\mathrm{E}(\mu_i) = \mathrm{E}(\epsilon_{ij}) = 0$ for all i and j, $\mathrm{E}(\mu_1 + \epsilon_{11})\mathrm{E}(\mu_1 + \epsilon_{12}) = 0$. Further, because the μ_i and ϵ_{ij} are independent, $\mathrm{Var}(\mu_1 + \epsilon_{11}) = \mathrm{Var}(\mu_1 + \epsilon_{12}) = \tau^2 + \sigma^2$. Applying these insights and expanding the expectation in the numerator gives

$$\frac{\mathrm{E}([\mu_1 + \epsilon_{11}][\mu_1 + \epsilon_{12}])}{\tau^2 + \sigma^2} = \frac{\mathrm{E}(\mu_1^2) + \mathrm{E}(\mu_1\epsilon_{11}) + \mathrm{E}(\mu_1\epsilon_{12}) + \mathrm{E}(\epsilon_{11}\epsilon_{12})}{\tau^2 + \sigma^2}.$$

Because the ϵ_{ij} are independent of each other, they have 0 covariance, which implies $\mathrm{E}(\epsilon_{11}\epsilon_{12}) = \mathrm{E}(\epsilon_{11})\mathrm{E}(\epsilon_{12}) = 0$. The same goes for the second two terms in the numerator, because the ϵ_{ij} are independent of the μ_i. Finally, rearranging the definition of variance gives $\mathrm{E}(\mu_i^2) = \mathrm{Var}(\mu_i) + [\mathrm{E}(\mu_i)]^2$. Because the expectation of the random effects is zero, $\mathrm{E}(\mu_i^2) = \mathrm{Var}(\mu_i) = \tau^2$. Applying these results gives the desired outcome,

$$\mathrm{Cor}(Y_{11}, Y_{12}) = \frac{\tau^2}{\tau^2 + \sigma^2}.$$

2) Here is code to carry out the simulations:

```
#Set parameters
alpha <- 3
beta <- 1/2
eps.sd <- sqrt(1/2)
re.sd <- 1
yrs <- 10
n.sims <- 10000

#Initialize variables
ints <- numeric(n.sims)
slopes <- numeric(n.sims)

#Simulate datasets and save least-squares estimates
for(i in 1:n.sims){
  x <- rep(anscombe$x1, yrs)
  rand.ints <- rnorm(length(anscombe$x1), 0, re.sd)
  y <- alpha + beta*x + rep(rand.ints, yrs) + rnorm(length(x),
0, eps.sd)
  mod.fit <- lm(y ~ x)
  ints[i] <- mod.fit$coefficients[1]
  slopes[i] <- mod.fit$coefficients[2]
}

#Plot and summarize estimates
hist(ints)
summary(ints)
sd(ints)

hist(slopes)
summary(slopes)
sd(slopes)
```

The least-squares estimates are unbiased. With these parameters, the standard deviation of the intercept estimates (which estimates the standard error of the estimator) is about 0.93, and the standard deviation of the slope estimates is about 0.10. These standard deviations are roughly in agreement with the mixed-model standard error estimates reported in the main text. They are much larger than the standard error estimates from simple linear regression. Ignoring dependence among the observations causes us to overestimate the amount of information we have, leading to standard error estimates that are too small.

Exercise Set A-1

1)
a) ii) $f'(x) = 2x - 2$; iii) $x = 1$.
b) ii) $f'(x) = -6x + 12$; iii) $x = 2$.
c) ii) $f'(x) = 3x^2 - 6x = 3x(x - 2)$; iii) $x = 0$ and $x = 2$.

2) Notice that the function in (1a) is minimized at $x = 1$, the function in (1b) is maximized at $x = 2$, and the function in (1c) is locally maximized at $x = 0$ and locally minimized at $x = 2$. The second derivatives are a) $f''(x) = 2$, b) $f''(x) = -6$, and c) $f''(x) = 6x - 6$. Plugging in the appropriate *x*-coordinates suggests that if $f'(c) = 0$ and $f''(c)$ is positive, then $f(x)$ is locally minimized when $x = c$. Similarly, if $f'(c) = 0$ and $f''(c)$ is negative, then $f(x)$ is locally maximized when $x = c$. These conjectures turn out to be true in general. One way to think of this is that when the slope is zero but increasing (as indicated by the positive second derivative), the function is minimized.

3) a) By the definition of the derivative,

$$g'(x) = \lim_{\Delta x \to 0} \frac{g(x + \Delta x) - g(x)}{\Delta x}$$

Because $g(x) = af(x)$,

$$g'(x) = \lim_{\Delta x \to 0} \frac{af(x+\Delta x) - af(x)}{\Delta x} = a \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = af'(x).$$

The first step comes from the definition of $g(x)$, the second step comes from the distributive property, and the third step comes from the definition of $f'(x)$.

b) Because $h(x) = f(x) + g(x)$,

$$h'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) + g(x+\Delta x) - f(x) - g(x)}{\Delta x} =$$
$$\lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} + \lim_{\Delta x \to 0} \frac{g(x+\Delta x) - g(x)}{\Delta x} = f'(x) + g'(x).$$

4) To find the derivative of the function $f(x) = ax^n$, where n is a positive integer, we take an approach similar to the one we took to find the derivative of $f(x) = x^2$. The steps are:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{a(x+\Delta x)^n - ax^n}{\Delta x}$$

Expanding using the binomial theorem gives

$$= \lim_{\Delta x \to 0} \frac{a(x^n + nx^{n-1}\Delta x + \binom{n}{2}x^{n-2}(\Delta x)^2 + \binom{n}{3}x^{n-3}(\Delta x)^3 + \cdots + \Delta x^n) - ax^n}{\Delta x}.$$

Distributing the *a* across the expression in parentheses and rewriting all but the first two terms in parentheses as a sum gives

$$= \lim_{\Delta x \to 0} \frac{ax^n + anx^{n-1}\Delta x + a(\Delta x)^2 \sum_{k=2}^{n} \binom{n}{k} x^{n-k} (\Delta x)^{k-2} - ax^n}{\Delta x}.$$

The positive and negative ax^n terms in the numerator cancel to give

$$= \lim_{\Delta x \to 0} \frac{anx^{n-1}\Delta x + a(\Delta x)^2 \sum_{k=2}^{n} \binom{n}{k} x^{n-k} (\Delta x)^{k-2}}{\Delta x}.$$

We can divide out the Δx in the denominator:

$$= \lim_{\Delta x \to 0} anx^{n-1} + a\Delta x \sum_{k=2}^{n} \binom{n}{k} x^{n-k} (\Delta x)^{k-2}.$$

The first term does not have Δx in it, so it is unaffected by the limit. In contrast, the second term is multiplied by Δx, so it goes to zero as Δx does, giving the result

$$f'(x) = anx^{n-1}.$$

Exercise Set A-2

1) a) 1 b) 4 c) 8 d) 0 e) 8

2)

b	The definite integral of $f(x) = 2x$ from 0 to b (i.e., $\int_0^b 2xdx$)
1	1
2	4
3	9
4	16
5	25

Exercise Set A-3

1) To solve this problem, consider that the derivative of ax^n with respect to x is nax^{n-1} and that differentiation and integration are inverse processes, meaning that $\int f'(x)dx = f(x) + C$. Jointly, these two facts imply that $\int nax^{n-1}dx = ax^n + C$. That is, to integrate x raised to an exponent, we raise the power of the exponent by one (so $n - 1$ becomes n) and divide by the new value of the exponent (this gets rid of the n in front). Then we add C to get the indefinite integral. The same thing applies when we start with n in the exponent instead of $n - 1$: we add one to the exponent and divide by the new value of the exponent. Thus, $\int ax^n dx = (ax^{n+1})/(n + 1) + C$. My calculus teacher taught us to say "up and under" as a way to remember this. Remember that for differentiation of polynomials, the phrase to remember is "out in front and down by one." When we integrate, we are doing the opposite.

This rule does not apply when $n = -1$. When $n = -1$, the "up and under" rule would force us to divide by zero, which is not allowed. We will need the integral of ax^{-1} in chapter 9.