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 $X_1, X_2, \ldots, X_n$ 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:

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

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

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

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

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

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

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

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

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

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

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

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 $R\left(\theta, \check{\theta}\right) = (6-\theta)^2$.

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.