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, $\mathrm{E}\left(I_{X_i \le z}\right) = F_X(z)$. Similarly, by the solution to exercise 4a in exercise set 5-2, $\mathrm{Var}\left(I_{X_i \le z}\right) = [1 - F_X(z)]F_X(z)$. 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), $\mathrm{E}\left[\widehat{F_n}(z)\right] = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} I_{X_i \le z}\right) = \frac{1}{n} n \mathrm{E}(I_{X_i \le z}) = F_X(z)$. Similarly, by equations 4.8-4.9 and the independence of the $I_{X_i \le z}$, $\mathrm{Var}\left[\widehat{F_n}(z)\right] = \mathrm{Var}\left([1/n]\sum_{i=1}^{n} I_{X_i \le z}\right) = (1/n^2) n \mathrm{Var}\left(I_{X_i \le z}\right) = (1/n)[1 - F_X(z)]F_X(z)$. $\widehat{F_n}(z)$ 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. (And neither version is robust—the sample standard deviation, because it relies on the sum of squared deviations from the mean—is delicately sensitive to outlying points.)

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) 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

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

2) 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 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 $x_1, x_2, \ldots, x_n$. The empirical distribution function is horizontal everywhere except at points directly above the observations $x_1, x_2, \ldots, x_n$, where there will be vertical line segments of length $1/n$. (If there are sets of observations with equal value among $x_1, x_2, \ldots, x_n$, 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 $\text{E}(\epsilon) = 0$. 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(n = 10, nsim = 500, a = 0, b = 0)
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.