Problem Set #7

Econ 103

Part I – Problems from the Textbook

Chapter 6: 1, 3, 5, 7
Chapter 7: 1, 3, 5, 9, 13, 17, 18, 19

The answer in the back of the book for 7-19 is wrong. *I will provide full solutions to 7-13 since it's hard, 7-18 since it's even-numbered, and 7-19 since the book is wrong.*

Solution: 7-13
The point is that S, the number of successes in n trials each with probability π of success, is a Binomial(n, π) random variable. We calculated the mean and variance of such a RV in class (see the slides) and we will use this information to find the MSE of $P = S/n$ as well as that of

$$P^* = \frac{nP + 1}{n + 2} = \left(\frac{n}{n+2}\right) P + \left(\frac{1}{n+2}\right)$$

The reason the book gives you the above expression is to give you a hint: namely that once you've solved for the MSE of P you can use this to get the MSE of P^* fairly easily by writing it as above.

$$\begin{aligned} E[P] &= E[S/n] = E[S]/n = n\pi/n = \pi \\ \text{Bias}(P) &= E[P] - \pi = \pi - \pi = 0 \\ Var(P) &= Var(S/n) = Var(S)/n^2 = n\pi(1-\pi)/n^2 = \pi(1-\pi)/n \\ MSE(P) &= \text{Bias}(P)^2 + Var(P) = 0^2 + \pi(1-\pi)/n = \pi(1-\pi)/n \end{aligned}$$

where we have used our rules for manipulating expectation and variance, as well as

the expressions for the mean and variance of a Binomial random variable. Now:

$$\begin{aligned} E[P^*] &= E\left[\left(\frac{n}{n+2}\right)P + \left(\frac{1}{n+2}\right)\right] = \left(\frac{n}{n+2}\right)E[P] + \left(\frac{1}{n+2}\right) \\ &= \left(\frac{n}{n+2}\right)\pi + \left(\frac{1}{n+2}\right) \end{aligned}$$

$$\begin{aligned} \text{Bias}(P^*) &= E[P^*] - \pi = \left(\frac{n}{n+2}\right)\pi + \left(\frac{1}{n+2}\right) - \pi \\ &= \left(\frac{n}{n+2} - 1\right)\pi + \left(\frac{1}{n+2}\right) = \frac{1-2\pi}{n+2} \end{aligned}$$

$$\begin{aligned} Var(P^*) &= Var\left[\left(\frac{n}{n+2}\right)P + \left(\frac{1}{n+2}\right)\right] = \left(\frac{n}{n+2}\right)^2 Var(P) \\ &= \frac{n^2}{(n+2)^2}\frac{\pi(1-\pi)}{n} = \frac{n\pi(1-\pi)}{(n+2)^2} \end{aligned}$$

$$\begin{aligned} MSE(P^*) &= \text{Bias}(P^*)^2 + Var(P^*) = \left(\frac{1-2\pi}{n+2}\right)^2 + \frac{n\pi(1-\pi)}{(n+2)^2} \\ &= \frac{(1-2\pi)^2 + n\pi(1-\pi)}{(n+2)^2} = \frac{1-4\pi+4\pi^2+n\pi-n\pi^2}{(n+2)^2} \\ &= \frac{1+(n-4)\pi-(n-4)\pi^2}{(n+2)^2} = \frac{1+\pi(1-\pi)(n-4)}{(n+2)^2} \end{aligned}$$

If we take limits, we'll see that both P and P^* are consistent, since their mean-squared errors go to zero as $n \to \infty$. For different values of π and n, however, the two estimators will have different MSE. Parts (d) and (e) of this question simply ask you to plug in various values and compare.

Solution: 7-18

The point of this question is non-response bias: the people who respond are not representative of the population as a whole. Note that P and P^* as defined in this question *do not correspond* to question 7-13. Our goal is to estimate the population proportion who will buy a computer. Using the table, we calculate the total number of people who will buy a computer as:

$$0.2 \times 40 + 0.04 \times 5 + 0.1 \times 3 + 0.2 \times 2 = 1.7 \text{ million}$$

which corresponds to a fraction $\pi^* = 1.7/50 = 0.034$. Again using the table, the number of people who will buy a computer *among the sub-population who would*

respond can be calculated as:

$$0.2 \times 7 + 0.04 \times 1 + 0.1 \times 1 + 0.2 \times 1 = 0.48 \text{ million}$$

which corresponds to a fraction $\pi = 0.48/10 = 0.048$. The point is that $\pi \neq \pi^*$. In other words, the proportion of people who would buy a computer *differs* across people who would and would not respond to the phone survey. The estimator P is based on calling 1000 people chosen at random and recording responses for *only those who reply*. The proportion of people who will reply is $10/50 = 1/5$. Thus, P will end up with a sample size of approximately $n = 200$ individuals. These individuals correspond to the sub-population for which the proportion who would buy a computer is $\pi^* = 0.048$. In contrast, the estimator P^* is based on calling $n^* = 100$ people chosen at random and then following up with these people repeatedly until *all of them respond*. Thus, P^* draws from the *full population*, in which a proportion $\pi^* = 0.034$ of people will buy a computer. The *true parameter* is π^* since we want to estimate the *overall* fraction of people who will buy a computer, *not* the fraction of people who would buy a computer among those who are likely to respond to a telephone survey. Hence bias is calculated *relative to* π^*. Variance is calculated *relative to the mean of each sampling distribution*. For P this mean is π while for P^* it is π^*. That is:

$$\begin{aligned}
MSE(P) &= \text{Bias}(P)^2 + Var(P) = (E[P] - \pi^*)^2 + E[(P-\pi)^2] \\
&= (\pi - \pi^*)^2 + E[(P-\pi)^2] \\
&= (\pi - \pi^*)^2 + \pi(1-\pi)/n \\
MSE(P^*) &= \text{Bias}(P^*)^2 + Var(P^*) = (\pi^* - \pi^*)^2 + E[(P^* - \pi^*)^2] \\
&= E[(P^* - \pi^*)^2] = \pi^*(1-\pi^*)/n^*
\end{aligned}$$

The estimator P^* does not have any bias because of the follow-ups to ensure that everyone in the original random sample responds. However, since it is based on a smaller sample, we would expect it to have a higher variance. The question is how this trade-off comes out in the expressions for MSE. To find out, we simply plug in the values $\pi^* = 0.034$, $\pi = 0.048$, $n = 200$ and $n^* = 100$. We find $MSE(P^*) \approx 0.000328$ and $MSE(P) \approx 0.000424$.

Solution: 7-19 *THE ANSWER IN THE BOOK IS WRONG!*

Specifically, they take $n = 100$ rather than $n = 200$ when calculating the variance of P. The reason this is wrong is because 1000 is the number of people *called* not

the number of people who *reply*. The question statement specifically states that P should be calculated relative to those who respond.

All we have to do in this question is take the square root of the answers from the previous question. We find that $RMSE(P^*) \approx 0.018$ and $RMSE(P) \approx 0.021$. These are the root mean squared errors for estimators of the population *proportion*. To answer the question for estimators of *market size*, i.e. the population proportion multiplied by the size of the market (50 million), we simply multiply each of the RMSE values by 50 million yielding values of approximately 900,000 and 1,000,000 respectively.

Part II – Additional Problems

Note: *In questions 1–5,* $X_1, X_2 \sim$ *iid* $N(\mu, \sigma^2)$, $Y = (X_1 - \mu)/\sigma$, $Z = (X_2 - \mu)/\sigma$.

1. (a) What is the distribution of $X_1 + X_2$?

   Solution: $X_1 + X_2 \sim N(2\mu, 2\sigma^2)$

   (b) Use R to calculate $P(X_1 + X_2 > 5)$ if $\mu = 5$ and $\sigma^2 = 50$.

   Solution: In this case, $X_1 + X_2 \sim N(10, 100)$, hence

   $$\begin{aligned} P(X_1 + X_2 > 5) &= 1 - P(X_1 + X_2 \leq 5) \\ &= 1 - P\left(\frac{X_1 + X_2 - 10}{10} \leq \frac{5 - 10}{10}\right) \\ &= 1 - \texttt{pnorm(-0.5)} \\ &\approx 0.6914625 \end{aligned}$$

   Alternatively, we could use `1- pnorm(5, mean = 10, sd = 10)`, which gives the same result.

   (c) Use R to calculate the 10th percentile of the distribution of $X_1 + X_2$.

   Solution: `qnorm(p = 0.1, mean = 10, sd = 10)` gives -2.815516.

2. (a) What is the distribution of Y^2?

Solution: As the sum of squares of one standard normal RV, $Y^2 \sim \chi^2(1)$.

(b) Use R to calculate $P(Y^2 \geq 1)$.

Solution:

$$P(Y^2 \geq 1) = 1 - P(Y^2 \leq 1) = 1 - \texttt{pchisq(1, df = 1)} \approx 0.3173105$$

3. (a) What is the distribution of $Y^2 + Z^2$?

Solution: Since this is the sum of squares of two independent standard normal random variables, $Y^2 + Z^2 \sim \chi^2(2)$.

(b) Use R to calculate the 95th percentile of the distribution of $Y^2 + Z^2$.

Solution: `qchisq(p = 0.95, df = 2)` gives 5.991465

4. (a) What is the distribution of $Z/\sqrt{Y^2}$?

Solution: Since it is the ratio of a standard normal to the square root of an independent χ^2 random variable divided by its degrees of freedom (in this case one), $Z/\sqrt{Y^2} \sim t(1)$.

(b) What value of c satisfies $P(-c \leq Z/\sqrt{Y^2} \leq c) = 0.95$?

Solution: By the symmetry of the t-distribution, it suffices to find the 97.5th percentile (this allocates 2.5% probability to the upper and lower tails). The command `qt(p = 0.975, df = 1)` gives 12.7062, so $c \approx 12.7$. Alternatively, we could have calculated the 2.5th percentile: `qt(p = 0.025, df = 1)` gives -12.7062.

(c) How does the interval in part (b) compare to the corresponding interval for Z?

Solution: Since Z is a standard normal RV, $P(-2 \leq Z \leq 2) \approx 0.95$. We see that the interval for a $t(1)$ RV is *much wider* than the corresponding interval for a standard normal. In other words, extreme outcomes are much more likely under the $t(1)$ distribution.

5. (a) What is the distribution of Y^2/Z^2?

> **Solution:** This is the ratio of two independent χ^2 random variables, each divided by its degrees of freedom (in this case, one). Hence $Y^2/Z^2 \sim F(1,1)$.

(b) Use R to calculate the 95th percentile of the distribution of Y^2/Z^2.

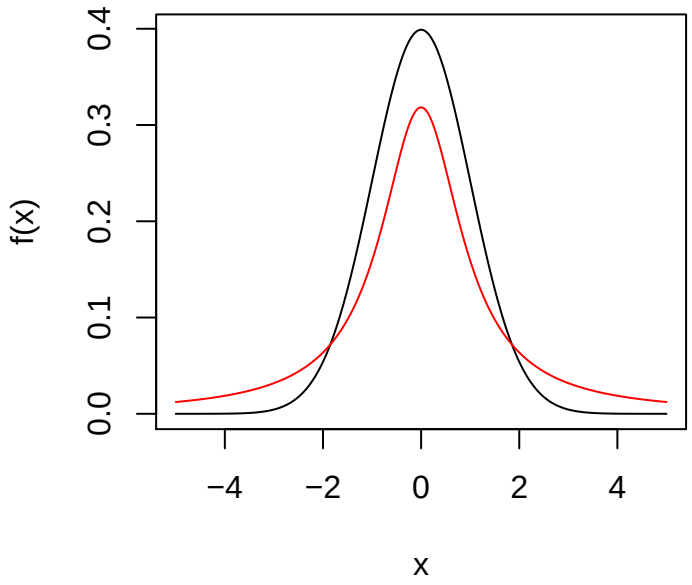
> **Solution:** `qf(p = 0.95, df1 = 1, df2 = 1)` gives 161.4476

6. In this question you will replicate some of the density plots from Lecture 13.

(a) Plot a standard normal pdf on the same graph as a $t(1)$ pdf on the interval $[-5, 5]$. How do the pdfs compare? Explain.

Solution:

```
x <- seq(from = -5, to = 5, by = 0.01)
y1 <- dnorm(x)
y2 <- dt(x, df = 1)
y <- cbind(y1, y2)
matplot(x, y, lty = 1, type = 'l', ylab = 'f(x)')
```

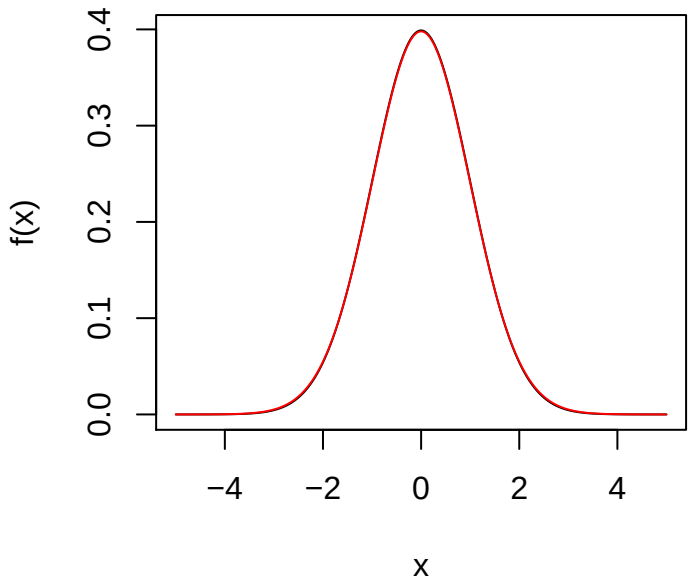


The $t(1)$ has much fatter tails than the normal: it's much more spread out. Although both are centered at zero, the t is much more likely to take on very large postive or negative values.

(b) Plot a standard normal pdf on the same graph as a $t(100)$ pdf on the interval $[-5, 5]$. How do the pdfs compare? Explain.

Solution: Carrying on from the code in the previous part:

```
y3 <- dt(x, df = 100)
y <- cbind(y1, y3)
matplot(x, y, lty = 1, type = 'l', ylab = 'f(x)')
```

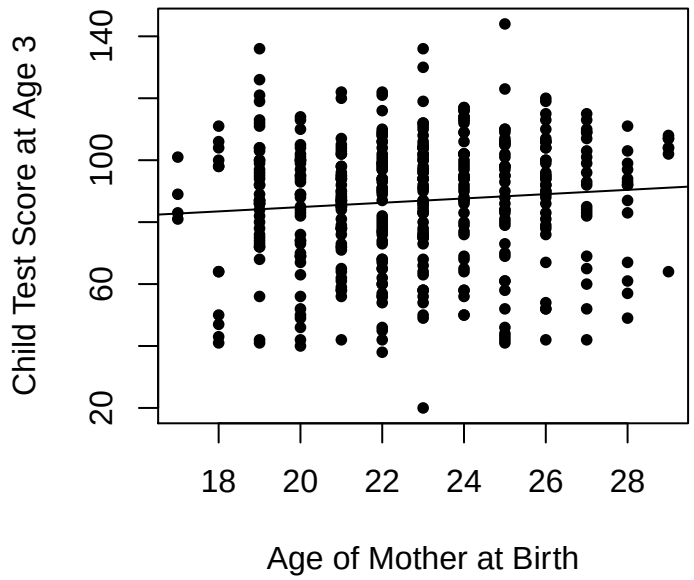


The two plots overlap almost perfectly: it looks like we've only plotted one curve! This is because the t distribution gets closer and closer to the standard normal as its degrees of freedom increase.

(c) Plot a χ^2 pdf with degrees of freedom equal to 4 on the interval $[0, 20]$.

Solution:

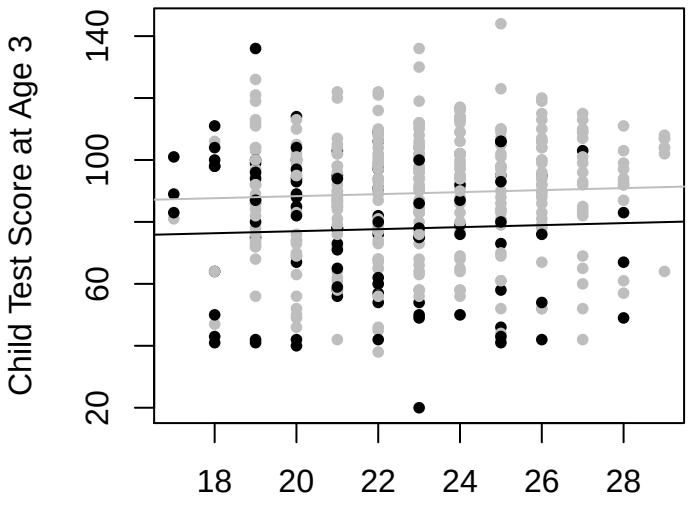
```r
x <- seq(from = 0, to = 20, by = 0.01)
y <- dchisq(x, df = 4)
plot(x, y, type = 'l', ylab = 'f(x)')
```



(d) Plot an F pdf with numerator degrees of freedom equal to 4 and denominator degrees of freedom equal to 40 on the interval $[0, 5]$.

Solution:

```r
x <- seq(from = 0, to = 5, by = 0.01)
y <- df(x, df1 = 4, df2 = 40)
plot(x, y, type = 'l', ylab = 'f(x)')
```



7. In this question you will verify the empirical rule both directly using `pnorm` and by simulation using `rnorm`.

   (a) Draw 100000 iid observations from a standard normal distribution and store your results in a vector called `sims`.

   Solution:

   ```
   sims <- rnorm(100000)
   ```

   (b) What proportion of the observations in `sims` lie in the range $[-1, 1]$?

   Solution:

   ```
   sum((sims >= -1) & (sims <= 1))/length(sims)
   ## [1] 0.6817
   ```

   (c) What proportion of the observations in `sims` lie in the range $[-2, 2]$?

   Solution:

```
sum((sims >= -2) & (sims <= 2))/length(sims)
## [1] 0.9546
```

(d) What proportion of the observations in `sims` lie in the range $[-3, 3]$?

Solution:

```
sum((sims >= -3) & (sims <= 3))/length(sims)
## [1] 0.9969
```

(e) Use `pnorm` to calculate the exact probabilities for a standard normal pdf that correspond to the above simulation experiments. How accurate were your simulations?

Solution:

```
pnorm(1) - pnorm(-1)
## [1] 0.6827
pnorm(2) - pnorm(-2)
## [1] 0.9545
pnorm(3) - pnorm(-3)
## [1] 0.9973
```

8. In this question you will replicate the Monte Carlo Experiment from Lecture 14. In particular, you will use R to study the sampling distribution of the sample mean where we take as our population the heights of all students in Econ 103.

(a) Load the class survey data used in R Tutorial # 2, extract the height column and assign it to a variable called `height`. Use `!is.na` to remove all missing values from `height`.

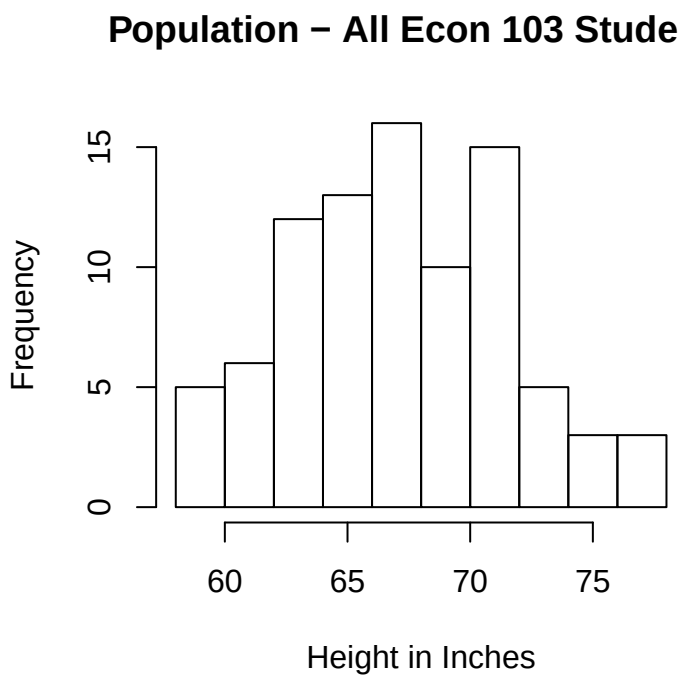
Solution:

```
data.url <- "http://www.ditraglia.com/econ103/old_survey.csv"
survey <- read.csv(data.url)
height <- survey$height
height <- height[!is.na(height)]
```

(b) Make a histogram of height and calculate the mean height for students in the class. For the purposes of this exercise, these correspond to the *population*.

Solution:

```
hist(height, main = 'Population - All Econ 103 Students',
     xlab = 'Height in Inches')
```




```
mean(height)
## [1] 67.55
```

(c) Write a function that takes n as its only input and returns the sample mean of an iid random sample of size n drawn from the vector `height`. Call this function `x.bar.draw`. [Hint: use `sample` with `replace = TRUE`.]

Solution:

```
x.bar.draw <- function(n){

  sim <- sample(height, size = n, replace = TRUE)
  return(mean(sim))

}#END x.bar.draw
```

(d) Test the function you wrote for part 4 by running it with `n = 10000`. What value

do you get? Your answer should be approximately equal the population mean you calculated above. If it isn't, something is wrong with your code.

Solution:

```
x.bar.draw(10000)
## [1] 67.49
```

(e) Using the R function `replicate`, run your function `x.bar.draw` 10000 times with $n = 5$. Store the result in a vector called `x.bar.5`. Do the same for $n = 10, 20$ and 50 and store the results as vectors `x.bar.10`, `x.bar.20` and `x.bar.50`

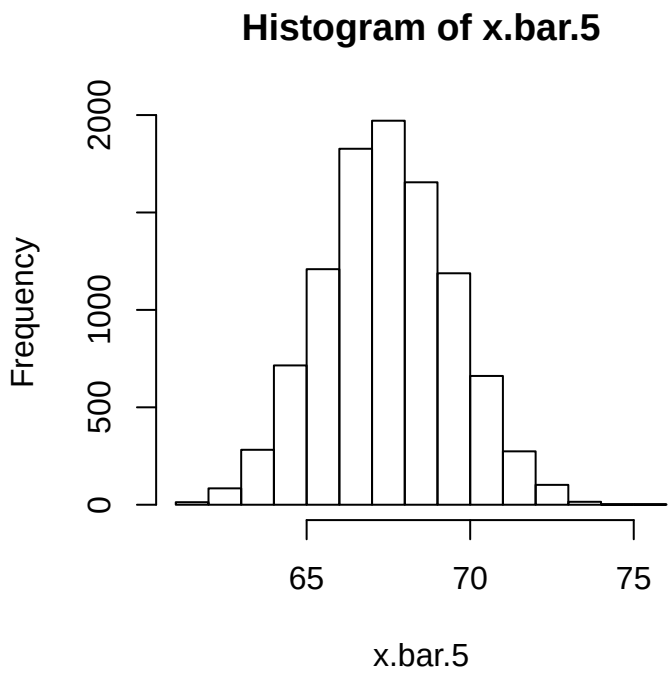
Solution:

```
x.bar.5 <- replicate(10000, x.bar.draw(5))
x.bar.10 <- replicate(10000, x.bar.draw(10))
x.bar.20 <- replicate(10000, x.bar.draw(20))
x.bar.50 <- replicate(10000, x.bar.draw(50))
```

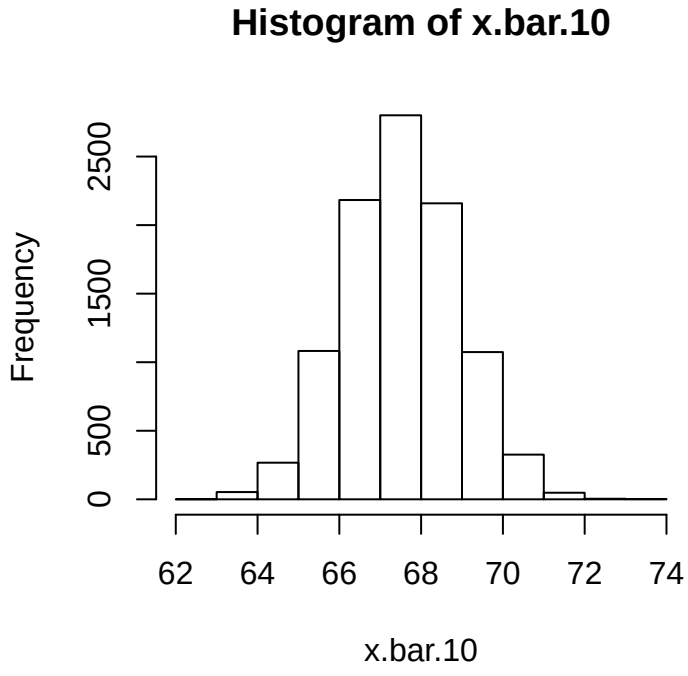
(f) Calculate the mean and variance of `x.bar.5`,`x.bar.10`, `x.bar.20` and `x.bar.50` and plot a histogram of each, being sure to label them.

Solution:

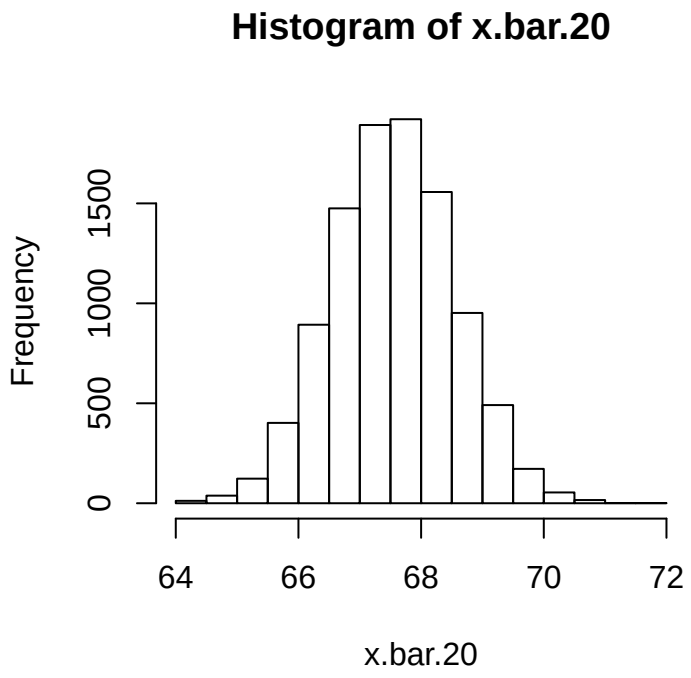
```
mean(x.bar.5)
## [1] 67.57
mean(x.bar.10)
## [1] 67.57
mean(x.bar.20)
## [1] 67.57
mean(x.bar.50)
## [1] 67.55
hist(x.bar.5)
```


Histogram of x.bar.5
Frequency
0
500
1000
2000
65
70
75
x.bar.5

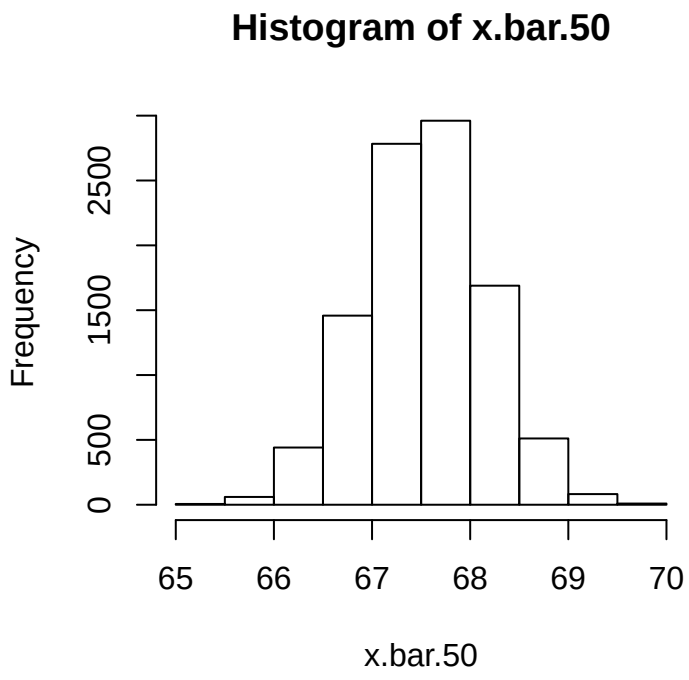
```
hist(x.bar.10)
```


Histogram of x.bar.10
Frequency
0
500
1500
2500
62
64
66
68
70
72
74
x.bar.10

```
hist(x.bar.20)
```




```
hist(x.bar.50)
```




```
var(x.bar.5)
## [1] 3.951
var(x.bar.10)
## [1] 1.933
var(x.bar.20)
## [1] 0.9881
var(x.bar.50)
## [1] 0.3905
```

9. In this question you will replicate the Law of Large Numbers (LLN) visualization from lecture 15, in which we plotted "running" sample means as we kept adding more and more simulations from a $N(\mu = 0, \sigma^2 = 100)$ distribution. Your plot won't look exactly like the one from class since this is a random experiment, but it will show the same qualitative behavior.

   (a) The R command for a "running" or "cumulative" sum is `cumsum`. Look at the help file for this command and test it out on a vector of ten ones and another containing the integers from one to ten to make sure you understand what it does.

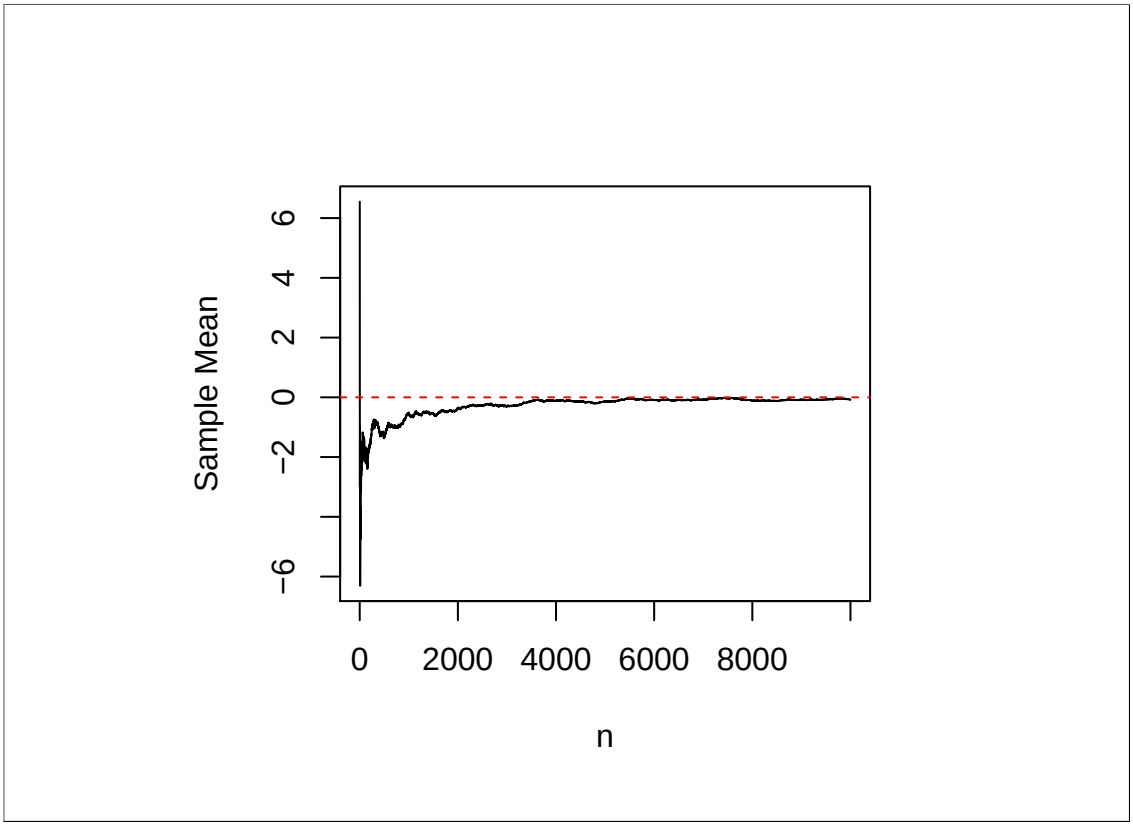
Solution:

```
ones <- rep(1, 10)
ones
##  [1] 1 1 1 1 1 1 1 1 1 1
cumsum(ones)
##  [1]  1  2  3  4  5  6  7  8  9 10
cumsum(1:10)
##  [1]  1  3  6 10 15 21 28 36 45 55
```

(b) Replicate the plot on slide 30/35 of lecture 15. First you'll need to draw 10,000 iid samples from a $N(\mu = 0, \sigma^2 = 100)$ distribution. Then you'll need to calculate the running means. You'll need to figure out how `cumsum` can be used to accomplish this. Finally, plot your results along with a dashed red line at the value to which the sample mean is converging. Make sure to label your axes.

Solution:

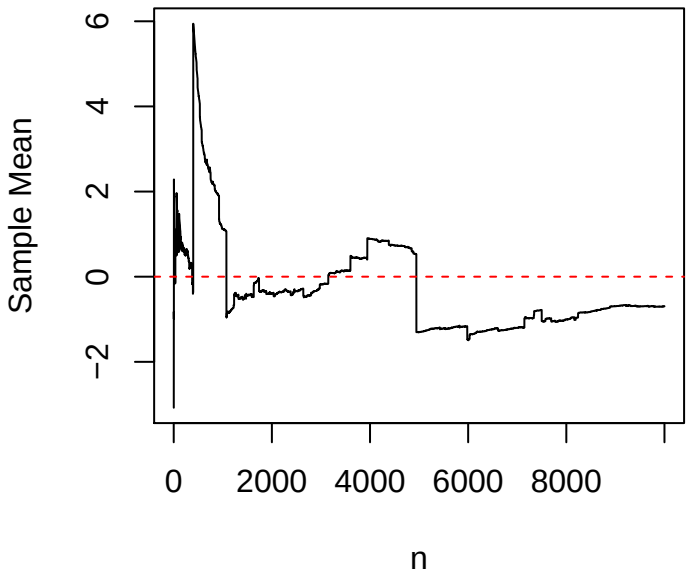
```
n <- 10000
sims <- rnorm(n, mean = 0, sd = 10)
running.mean <- cumsum(sims)/(1:n)
plot(1:n, running.mean, type = 'l', xlab = 'n', ylab = 'Sample Mean')
abline(h = 0, col = "red", lty = 2)
```



(c) Repeat the previous part but, rather than drawing $N(\mu = 0, \sigma^2 = 100)$ simulations, draw from a Student-t distribution with one degree of freedom. How do your results differ? Use what you know about the Student-t distribution to guess why our proof that the sample mean is consistent for the population mean doesn't work here.

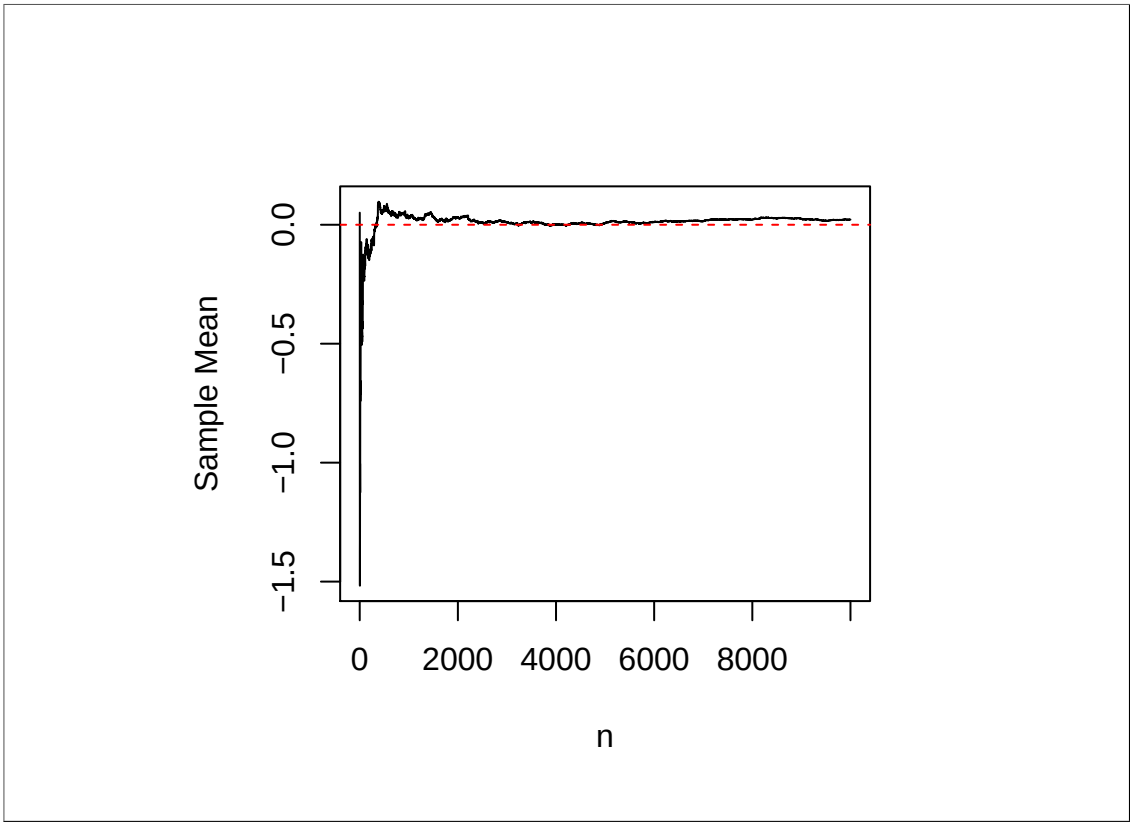
Solution:

```
n <- 10000
sims <- rt(n, df = 1)
running.mean <- cumsum(sims)/(1:n)
plot(1:n, running.mean, type = 'l', xlab = 'n', ylab = 'Sample Mean')
abline(h = 0, col = "red", lty = 2)
```



This plot looks totally different from the previous one: the "running means" never settle down in this case. To prove that the sample mean is consistent for the population mean, we tacitly assumed that both the mean and variance of X_i exist and are finite. (Remember that both quantities are defined as improper integrals, so they could diverge or may be undefined.) It turns out that the Student-t distribution with one degree of freedom has an *infinite variance and a mean that does not exist*. Essentially its mean is $\infty - \infty$ which does *not* equal zero: it's simply undefined. To get the LLN to work for a Student-t, we need a finite mean and variance. Both conditions turn out to hold as long as the degrees of freedom are ≥ 3. For example:

```
n <- 10000
sims <- rt(n, df = 3)
running.mean <- cumsum(sims)/(1:n)
plot(1:n, running.mean, type = 'l', xlab = 'n', ylab = 'Sample Mean')
abline(h = 0, col = "red", lty = 2)
```


Sample Mean
0.0
−0.5
−1.0
−1.5
0
2000
4000
6000
8000
n