Homework 2

The purpose of this homework is to explore sampling distributions, to practice using the bootstrap to construct confidence intervals, and to gain more experience programming in R. Please fill in the appropriate code and write answers to all questions in the answer sections, then submit a compiled pdf with your answers through Gradescope by 11pm on Sunday September 17th.

If you need help with the homework, please attend the TA office hours which are listed on Canvas and/or ask questions on Ed Discussions. Also, if you have completed the homework, please help others out by answering questions on Ed Discussions, which will count toward your class participation grade.

Some tips for completing this homework are:

1. Make sure you conceptually understand the problems first before trying to write code for the solutions. For example, if the problem asks you to create a plot, it could be useful to first draw a picture of the plot and think about the steps needed to get to the answer before writing down any code.
2. For any questions asking for a particular value, be sure to print out the value in your R chunks to "show your work" as well as reporting the value in the answer section. In an R chunk, you can print the value in an object by putting the object by itself on a line, or by putting parenthesis around an expression where an assignment is made.
3. Be sure to always label your plots! And spend time to make them look good.
4. In order to see the LaTeX equations, it is useful to look at the knitted pdf document. In general, it could be useful to have a pdf copy of the homework open to more easily read the instructions as you work on the RMarkdown file.

Some useful LaTeX symbols for the problem set are: μ, σ, $\bar{x}$, $\frac{a}{b}$. You can use LaTeX symbols to annotate your plots with the `TeX()` command.

Part 1: Exploring sampling distributions with simulations

As discussed in class:

- A **statistic** is a number computed from a sample of data – like a sample mean, or a sample standard deviation.
- A **sampling distribution** is a probability distribution of a *statistic*. If we repeatedly drew samples from some underlying distribution, and computed the same statistic on each sample, the distribution of these *statistics* is their *sampling distribution*.

The shape of the underlying distribution of data, and the shape of the sampling distribution for a statistic calculated from samples of data, can often be quite different. Below we explore this by looking at the price of condominiums in New Haven.

Problem 1.1: (10 points)

A condominium is a building or complex of buildings that contain a number of individually owned apartments. In this first set of problems, we will examine the assessed prices of *all* condominium apartments in New Haven in order to understand sampling distributions.

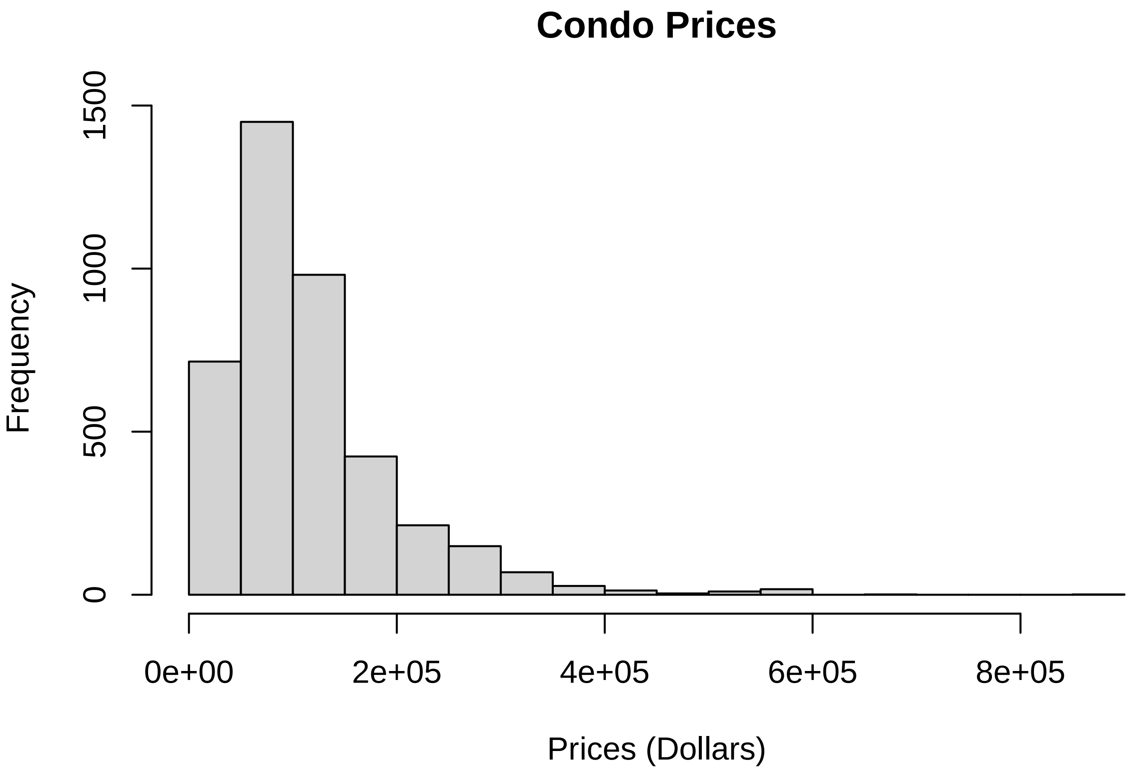
The code below loads the data and creates a variable called `prices` which has the tax assessed prices of all condominium apartments. To start, please plot a histogram of the prices of all the condominium apartments (as always, be sure to label your axes). Also, calculate and print out the mean, median, and standard deviation in this data.

In the answer section, please answer the following questions.

a. Describe the shape of the histogram of the data (e.g., is it left-skewed, right-skewed, symmetric, etc.?).
b. Suppose we are only interested in the condominiums in New Haven (i.e., the data we have is the population). Please write the symbols we should use to denote the mean and standard deviations you calculated above, using the symbols we have discussed in class.

Note about the data: The data was scraped from the website https://gis.vgsi.com/newhavenct/. One outlier data point was removed where the value of the condominium was listed as above $8 million, and was due to listing the whole complex rather than an individual unit).

```r
# load the data and get a vector of condominium apartment prices
load("condos_may_2022.rda")
prices <- condos$assessment








# plot a histogram of the data
hist(prices, xlab = "Prices (Dollars)", main = "Condo Prices")
```



```r
# calculate some parameters of the data
mean(prices)
```

```
## [1] 116040.5
```

```r
median(prices)
```

```
## [1] 94710
```

```r
sd(prices)
```

```
## [1] 82045.6
```

Answers

a. The histogram appears to be right-skewed.

b. If the data we have is the population, then we should use $\bar{x}$ and σ to describe the mean and standard deviation.

Part 1.2: (15 points)

Now let's examine the *sampling distribution* for the mean statistic $\bar{x}$ when taking random samples from our condominium apartment prices. To do this, create a sampling distribution that has 10,000 mean statistics, $\bar{x}$, using $n = 36$ points in each sample that are randomly sampled from the underlying `prices` population data. Hint: you should do this with a `for` loop that generates one mean statistic in each iteration of the loop, and the `sample()` function will be very useful.

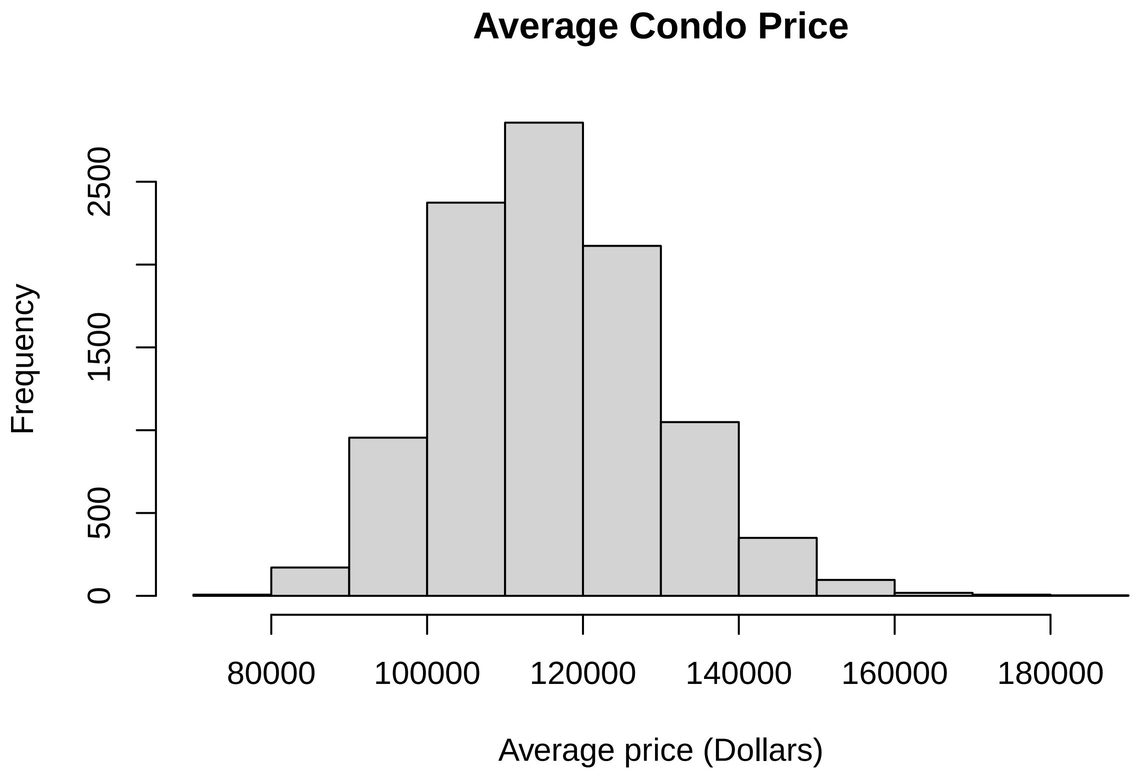
One you have created this sampling distribution, plot it by creating a histogram of these sample statistic values. In the answer section below, answer the following questions:

a. Describe the shape of this distribution and whether this is the shape you would expect.

b. Report the standard error of this sampling distribution.

c. Report whether the mean of the sampling distribution you created here is similar to the mean you calculated in part 1.1.

```r
price_mean <- NULL
for(i in 1:10000){
  price_sample <- sample(prices, 36)
  price_mean[i] <- mean(price_sample)
}

hist(price_mean, xlab = "Average price (Dollars)", main = "Average Condo Price")
```



```
sd(price_mean)
```

```
## [1] 13725.16
```

```
mean(price_mean)
```

```
## [1] 115991.5
```

Answers:

a. The distribution matches a normal distribution, which is what I would expect based on the Central Limit Theorem.

b. The standard error is $13725.16

c. The mean of the sampling distribution ($115991.50) is similar to the mean I calculated in part 1.1 ($116040.50).

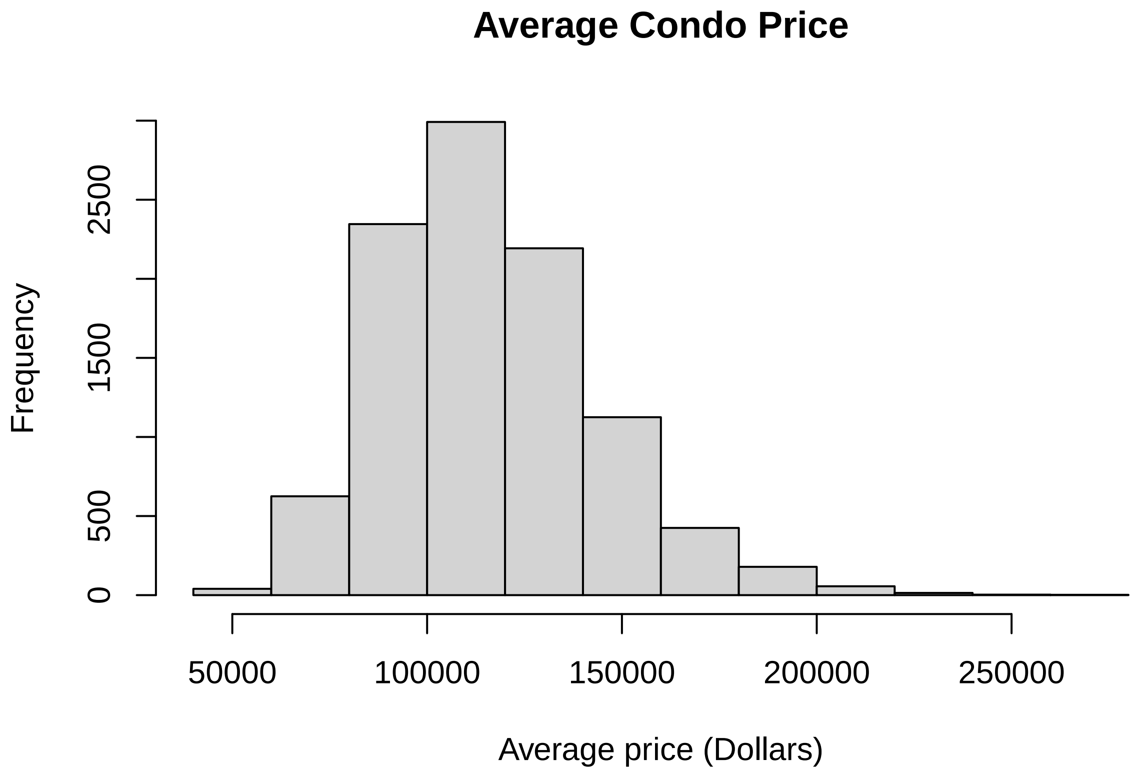
Part 1.3: (5 points) Challenge problem

This is a "challenge problem" that you should try to figure out without getting help from the TAs. Challenge problems might be more difficult than other problems but they won't be worth too many points, so they will not have a large impact on your homework score.

Repeat part 1.2 using sample sizes of $n = 9$ and $n = 144$. Report the standard errors for $n = 9$, 36, and 144, and describe how the relationship between values for the standard error SE change with the different values of n. Also describe why it makes sense the SE would get smaller as n increases.

Finally, describe theoretical results (i.e., a formula) from intro stats that can account for the relationship between between the SE and n. Also describe whether this formula approximately matches the standard errors you computed by computationally sampling from your data.

```
price_mean_2 <- NULL
for(i in 1:10000){
  price_sample <- sample(prices, 9)
  price_mean_2[i] <- mean(price_sample)
}

hist(price_mean_2, xlab = "Average price (Dollars)", main = "Average Condo Price")
```



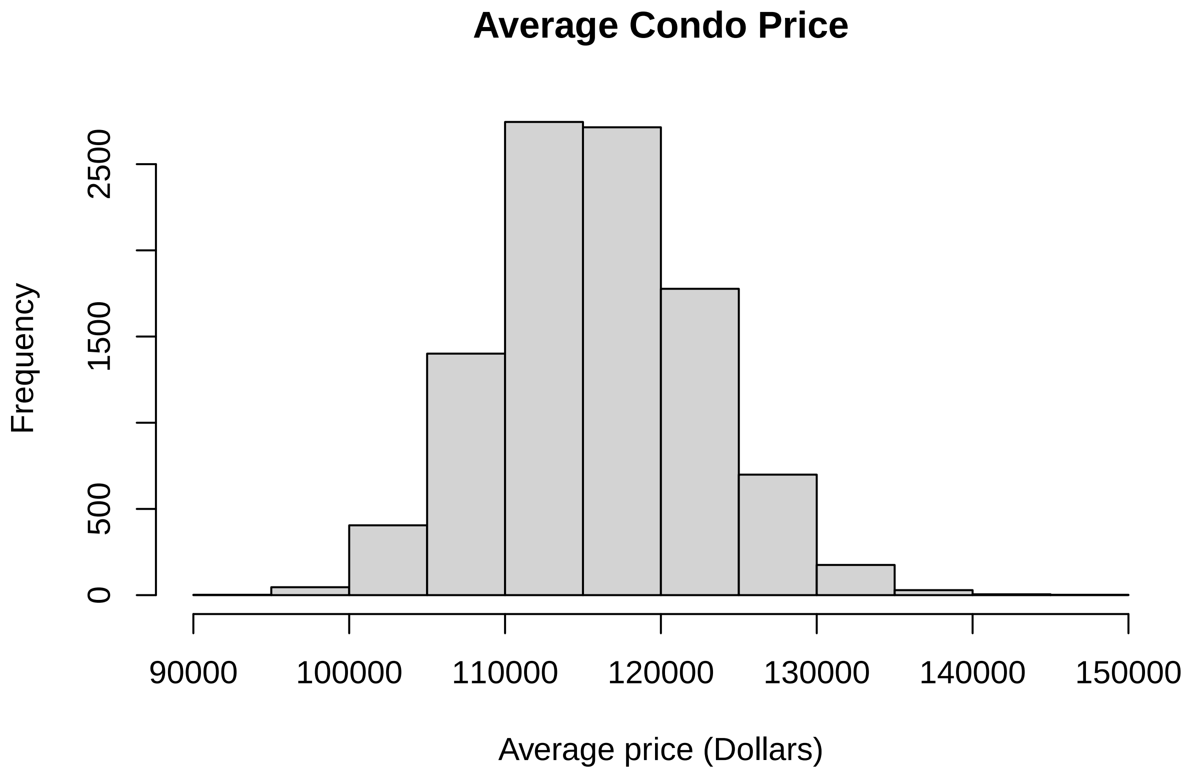
```
sd(price_mean_2)
```

```
## [1] 27507.83
```

```
mean(price_mean_2)
```

```
## [1] 116158
```

```
price_mean_3 <- NULL
for(i in 1:10000){
  price_sample <- sample(prices, 144)
  price_mean_3[i] <- mean(price_sample)
}

hist(price_mean_3, xlab = "Average price (Dollars)", main = "Average Condo Price")
```



```r
sd(price_mean_3)
```

```
## [1] 6690.023
```

```r
mean(price_mean_3)
```

```
## [1] 115968.9
```

Answer

Standard errors:

For $n = 9$: \$27507.83 For $n = 36$: \$13725.16 For $n = 144$: \$6690.023

As n increases, the standard error decreases.

Part 2: Exploring bias correction in the formula for the variance statistic

In intro stats class you learned that the formula for the sample variance statistic is:

$$s^2 = \frac{\Sigma_i^n (x_i - \bar{x})^2}{n-1}$$

Students often ask why the denominator in this formula is n - 1 rather than just n. To answer this, let's create a sampling distribution of the variance statistic using a denominator of n - 1 and compare it to using a denominator of n.

Part 2 (15 points) For this question, use the `var()` function built into R and the `var_n()` function which has been written below. `var()` returns the variance statistic given the data sample, and `var_n()` returns the variance statistic using a denominator of n rather than $n - 1$.

First, create a sampling distribution using `var()` and `var_n()` using data generated from the standard normal distribution for a sample size of n = 10. You can use `rnorm()` to generate random numbers from the standard normal distribution.

Plot histograms of these sampling distributions, and calculate the mean of these sampling distributions. Also use the `abline(v = ...)` function to plot a vertical line in red at the value of the parameter $\sigma^2 = 1$, and a vertical line in blue for the mean (expected value) of the sampling distribution.

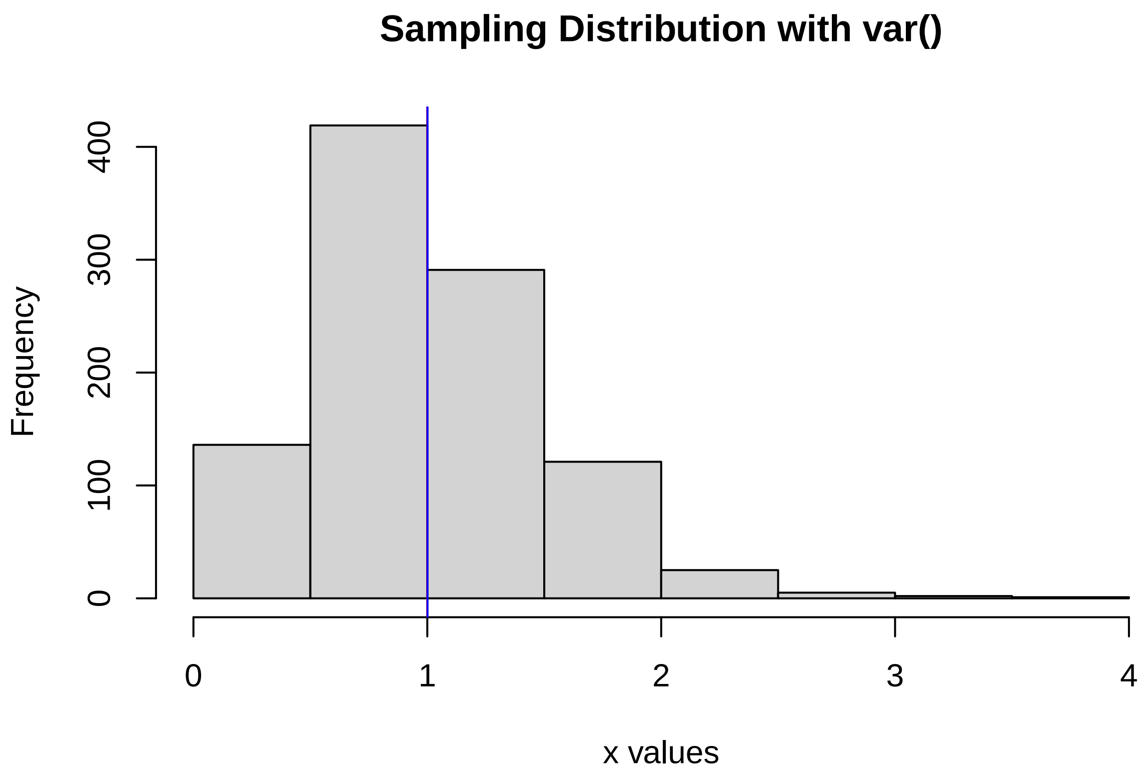
Then report below:

1. The shapes of these distributions.
2. Whether the means of these sampling distribution equal the underlying variance parameter of $\sigma^2 = 1$.

Note: A statistic is called an *estimator* when it's used as an estimate of an underlying parameter – here, we're trying out different estimators for the true variance. An estimator is called *biased* if its mean (expected value) does not equal the population parameter it is trying to estimate. Thus, if the mean value of our sampling distribution does not equal the true variance $\sigma^2 = 1$, which it is trying to estimate, then our statistic (estimator) is biased.

Bonus (0 points): Experiment with using a sample size of n = 100 data points for each statistic that goes into your sampling distribution. Is the bias smaller when the sample size n is larger? Make sure you not only get the code to work, but that you understand the concepts discussed here since there is a reasonable chance these concepts could appear on an exam.

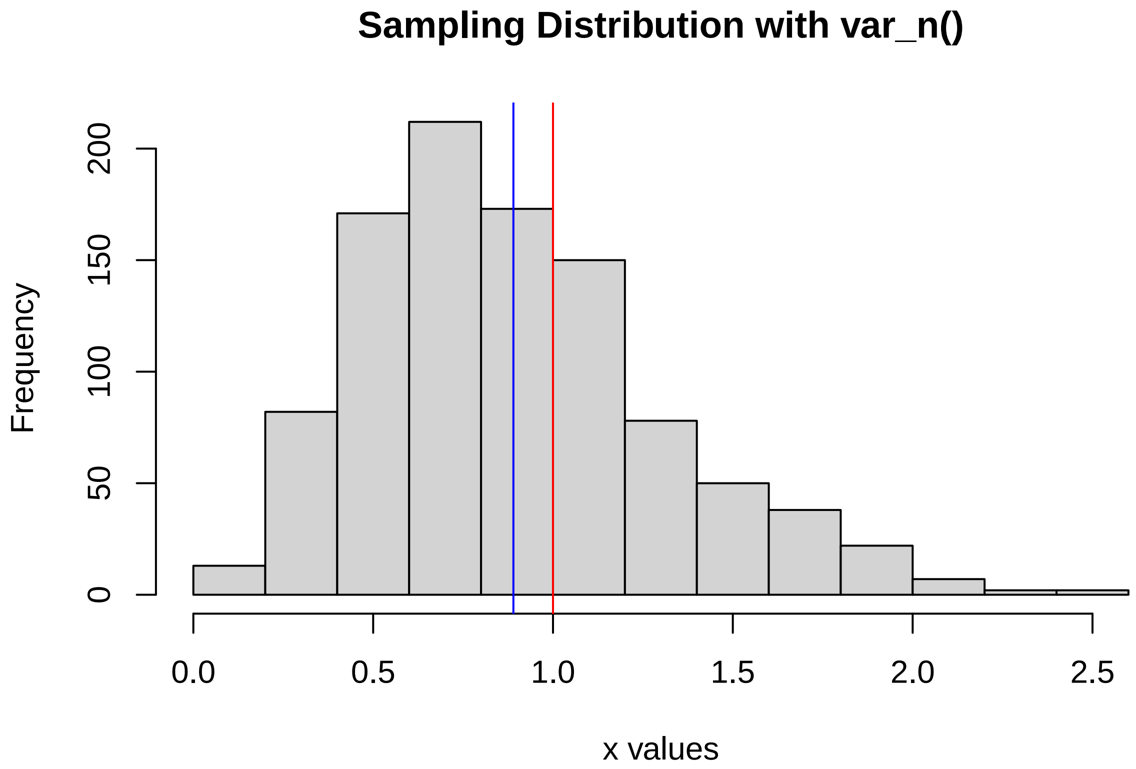
```r
var_n <- function(data_sample){
  var(data_sample) * (length(data_sample) - 1)/length(data_sample)
}




# Create two (approximate) sampling distributions: one using the var() statistic
# and one using the var_n() statistic. These sampling distributions should be
# stored in vectors called sampling_dist_var and sampling_dist_var_n


sampling_dist_var <- NULL
for (i in 1:1000){
  var_sample <- rnorm(10)
  sampling_dist_var[i] <- var(var_sample)
}


sampling_dist_var_n <- NULL
for (i in 1:1000){
  var_sample_n <- rnorm(10)
  sampling_dist_var_n[i] <- var_n(var_sample_n)
}


# Plot a histogram of the sampling distributions for the var() statistic. Add a
# red vertical line at the variance parameter value and a blue vertical line at
# the mean value of this sampling distribution.
hist(sampling_dist_var, xlab = "x values", main = "Sampling Distribution with var()")
abline(v = 1, col = "red")
abline(v = mean(sampling_dist_var), col = "blue")
```



```
mean(sampling_dist_var)
```

```
## [1] 1.000793
```

```
# Plot a histogram of the sampling distributions for the var_n() statistic. Add
# a red vertical line at the variance parameter value and a blue vertical line
# at the mean value of this sampling distribution.
hist(sampling_dist_var_n, xlab = "x values", main = "Sampling Distribution with var_n()")
abline(v = 1, col = "red")
abline(v = mean(sampling_dist_var_n), col = "blue")
```



```r
mean(sampling_dist_var_n)
```

```
## [1] 0.8898828
```

```r
# Print out the mean value of the var() and var_n() sampling distributions.
```

Answers

1. The shapes of these distributions. The distributions have similar shapes in that both are right-skewed, but their means differ.
2. Whether the means of these sampling distribution equal the underlying variance parameter of $\sigma^2 = 1$. The mean of the var() sampling distribution equals the underlying variance parameter, but the mean of the var_n() sampling distribution does not.

Part 3: Calculating confidence intervals using the bootstrap

It is well known that Millennials **LOVE** avocado toast. It is also well known that Millennials prefer to eat organic food when given the option. However, is the additional cost of eating organic avocados worth it? Let's explore this question by using the bootstrap to create confidence intervals for the overall average price of conventional and organic avocados.

The data used in this assignment comes from Kaggle.com and was originally taken from the Hass Avocado Board. Kaggle is a great website to get datasets and to practice your Data Science skills, so I recommend you take a look at the site particularly when you are looking for datasets for your final project.

Part 3.1 (15 points)

The code below loads a data frame called `avocados` that has information about the prices of avocados in the Northeastern United States. We are interested in creating confidence intervals for the overall average price of organic and conventional avocados.

To start the analysis, please complete the following steps:

1. In the answer section below, report what each row (case) in this data set represents. Then, considering that we are trying to infer what the **average price** of organic and conventional avocados are, describe what the underlying population might be that we are making inferences about. Finally, use LaTeX to write the appropriate symbols for the values we are trying to make inferences about using the standard symbols we have discussed in class.

2. Create a vector called `conventional_price` that has the prices of conventional avocados, and a vector called `organic_price` that has the prices of organic avocados. Report how many cases are in each of these vectors. Also report what the average price of the conventional and organic avocados each are in this dataset and use LaTeX to report the appropriate symbol for these average values.

3. Visualize the data by creating one side-by-side box plot and two histograms of the conventional and the organic avocado prices. Be sure to appropriately label your plots and make sure the histogram also has an appropriate number of bins. In the answer section, report whether you believe the overall average price for organic avocados is higher than the overall average price for conventional avocados.

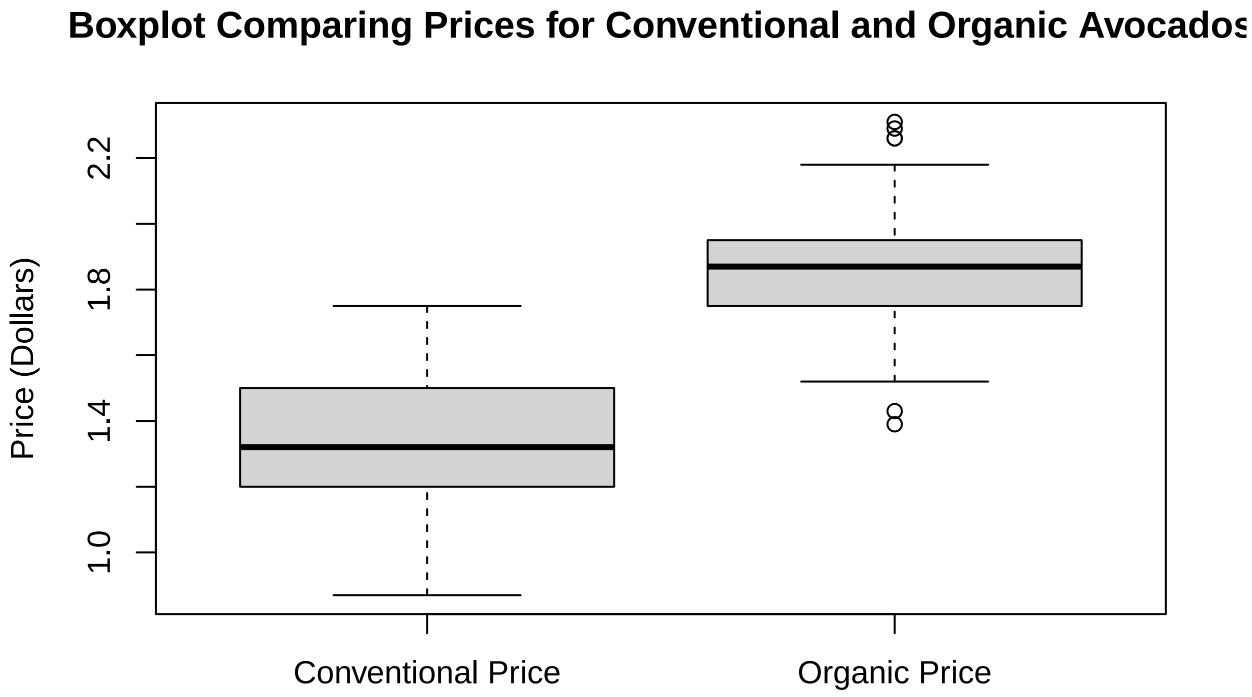
```
# each row represents the weekly price of either conventional or organic avocados from 2015 to 2018

# load the data set
 load("avocados_northeast.rda")

conventional_price <- avocados$WeeklyPrice[avocados$type == "conventional"]

mean(conventional_price)
```

```
## [1] 1.344438
```

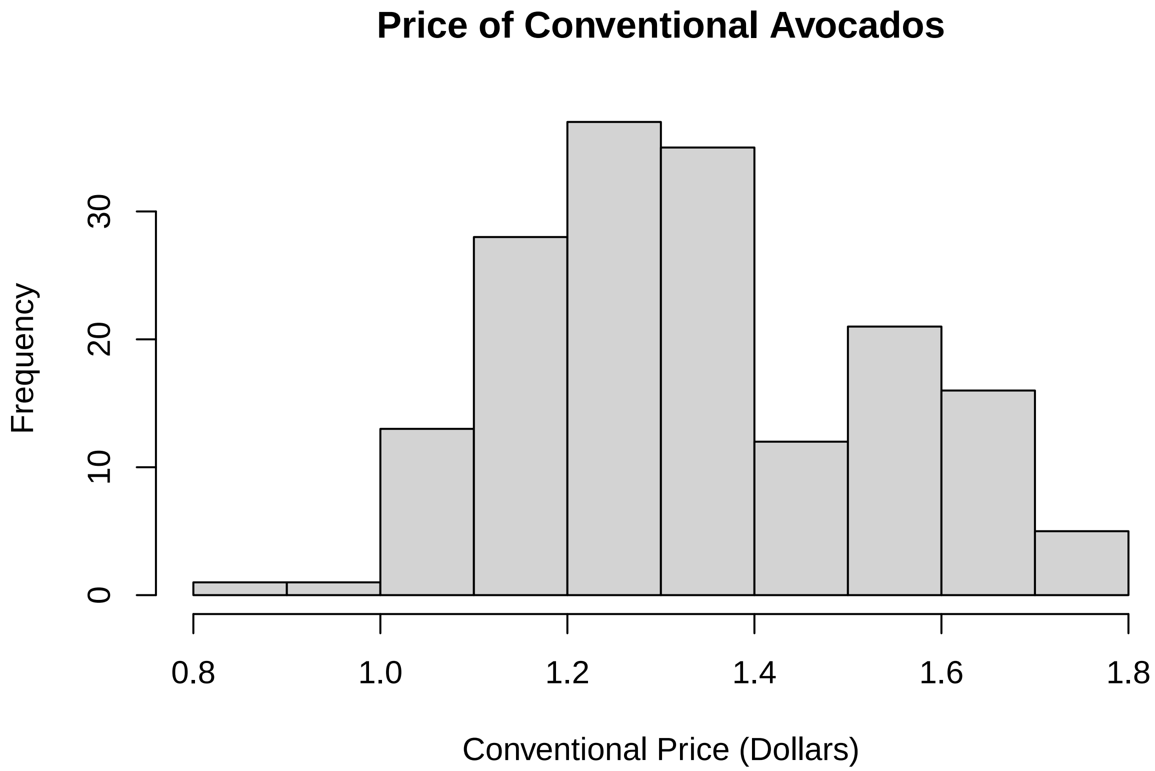
```
organic_price <- avocados$WeeklyPrice[avocados$type == "organic"]

mean(organic_price)
```

```
## [1] 1.859408
```

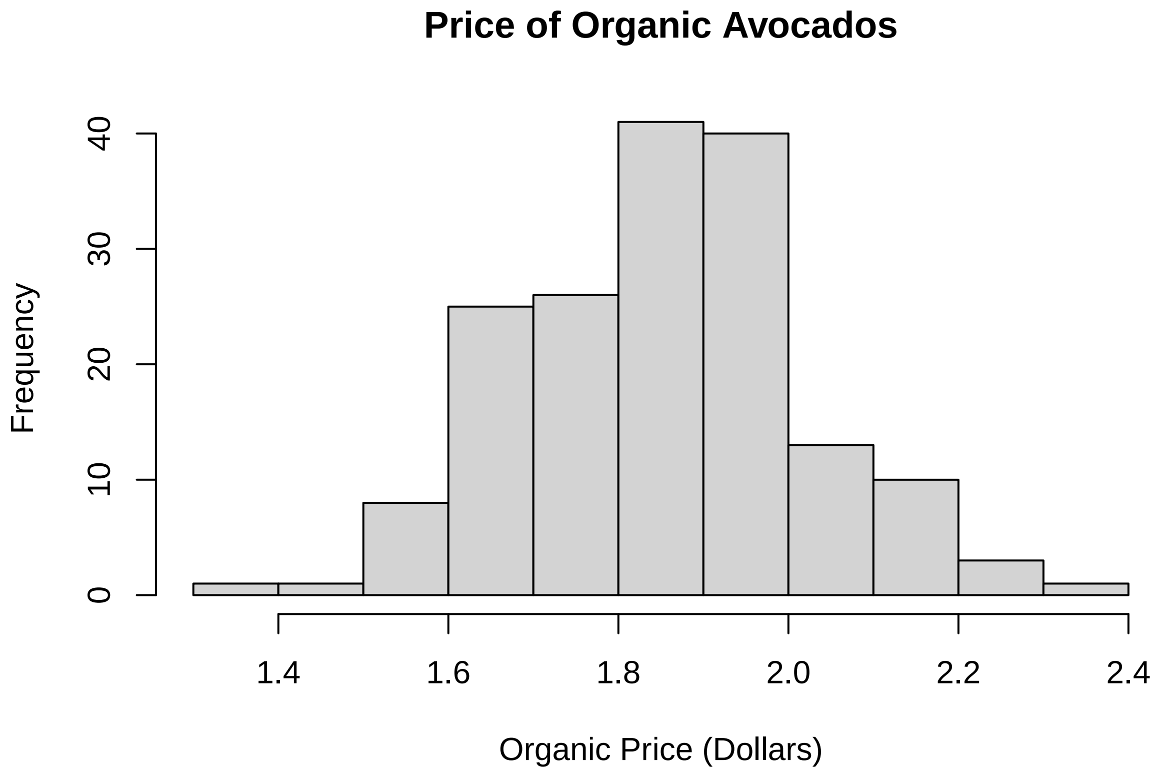
```
boxplot(conventional_price, organic_price, ylab = "Price (Dollars)", names = c("Conventional Price", "O
```



```
hist(conventional_price, xlab = "Conventional Price (Dollars)", main = "Price of Conventional Avocados")
```



```r
hist(organic_price, xlab = "Organic Price (Dollars)", main = "Price of Organic Avocados")
```



Answer:

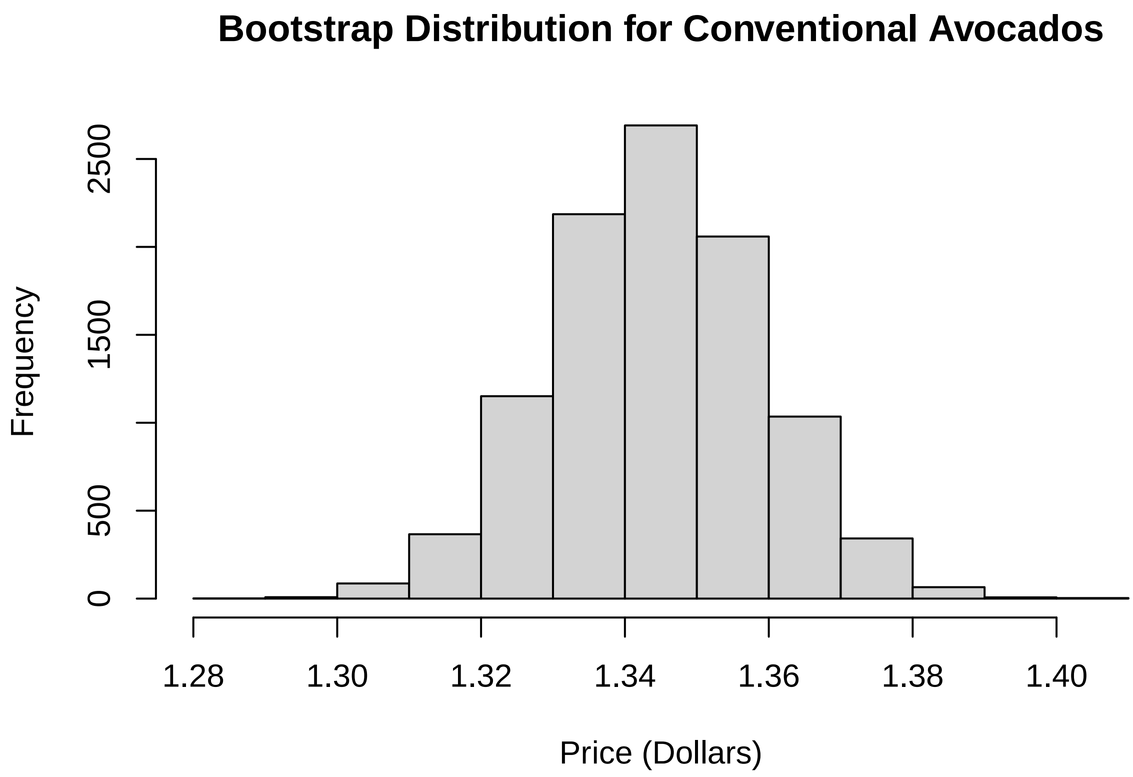
1. Each row represents an order of avocados in the Northeastern US from some time between 2015 and 2018. The underlying population might be every order of avocados ever in the Northeastern US. We are trying to make inferences about $\bar{x}$.
2. There are 169 cases in each vector. The average price of the conventional avocados is $1.34. The average price of the organic avocados is $1.86.
3. I do believe that the overall average price for organic avocados is higher than the overall average price for conventional avocados.

Part 3.2 (15 points)

Now use the bootstrap to create a 95% confidence interval for the **conventional** avocados. Be sure to display the bootstrap distribution you created, and report the bootstrap standard error as well the 95% confidence interval. Based on the confidence interval you created, does it seem likely that the average conventional avocado price is the same as the average organic avocado price?

```
bootstrap_dist_conventional <- NULL
for (i in 1:10000){
  boot_sample_conventional <- sample(conventional_price, replace = TRUE)
  bootstrap_dist_conventional[i] <- mean(boot_sample_conventional)
}
```

```
hist(bootstrap_dist_conventional,
     xlab = "Price (Dollars)",
     main = "Bootstrap Distribution for Conventional Avocados")
```



```
(SE_boot_conventional <- sd(bootstrap_dist_conventional))
```

```
## [1] 0.01466145
```

```
CI_lower_conventional <- mean(conventional_price) - 2 * SE_boot_conventional
CI_upper_conventional <- mean(conventional_price) + 2 * SE_boot_conventional

c(CI_lower_conventional, CI_upper_conventional)
```

```
## [1] 1.315115 1.373761
```

Answer:

Based on the confidence interval, it does not seem likely that the average conventional avocado price is the same as the average organic avocado price.

Part 3.3 (5 points)

In order for the bootstrap confidence interval to truly capture the parameter of interest, a few conditions should be met, including that the data points should be independent draws from the underlying distribution and that the distribution of the bootstrap statistics should be approximately normal. Does it appear that these conditions are met for this data and consequently should we trust our conclusions?

Answers:

It does appear that these conditions are met for this data, and therefore we should trust our conclusions.

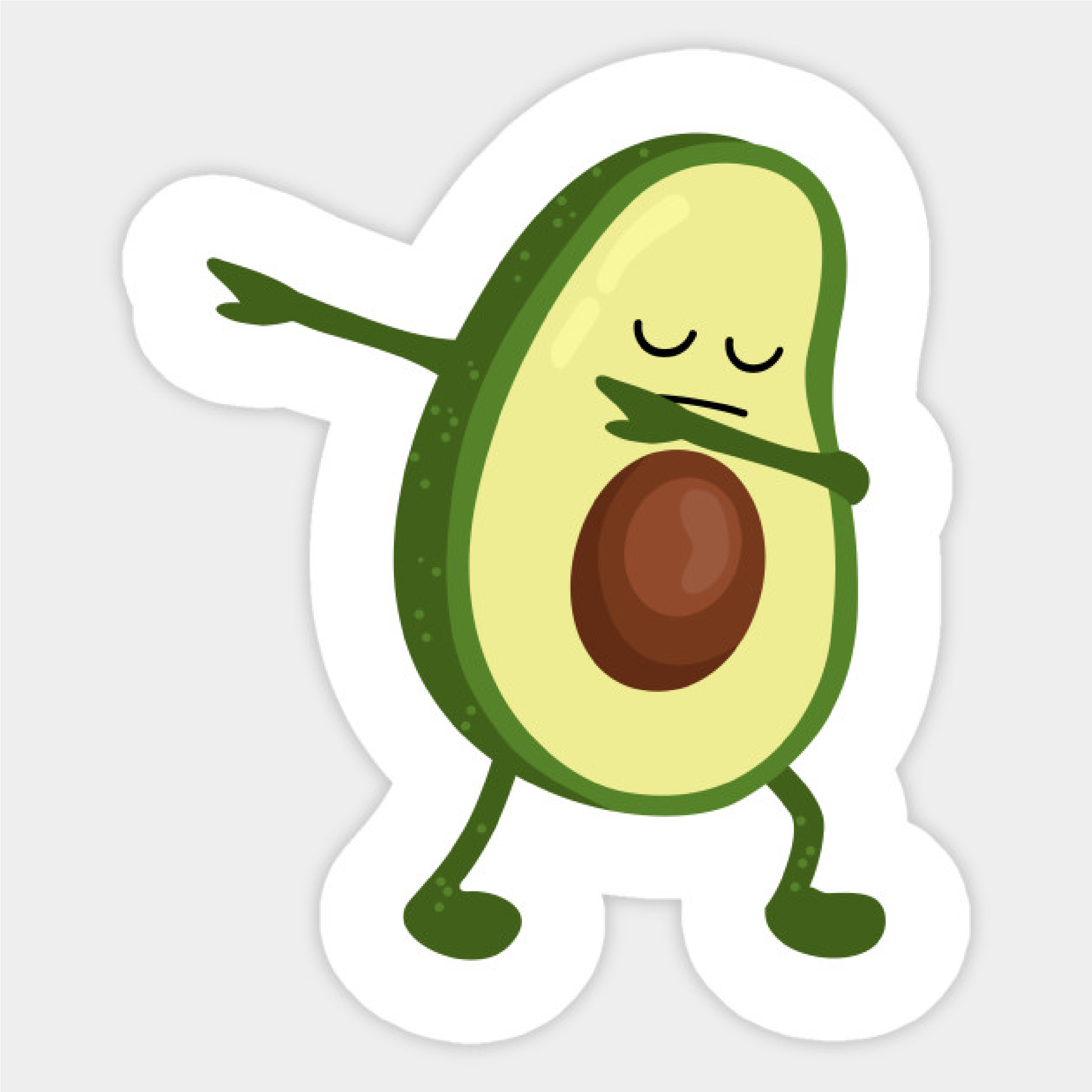


Figure 1: An avocado dancing to the avocado price song

Reflection (3 points)

Please reflect on how the homework went by going to Canvas, going to the Quizzes link, and clicking on Reflection on homework 2