SDS 315 Homework 3

Elizabeth 'Betsy' Sherhart UT EID: eas5778
Click here for link to GitHub repository

February 13, 2025

Problem 1

Theory A

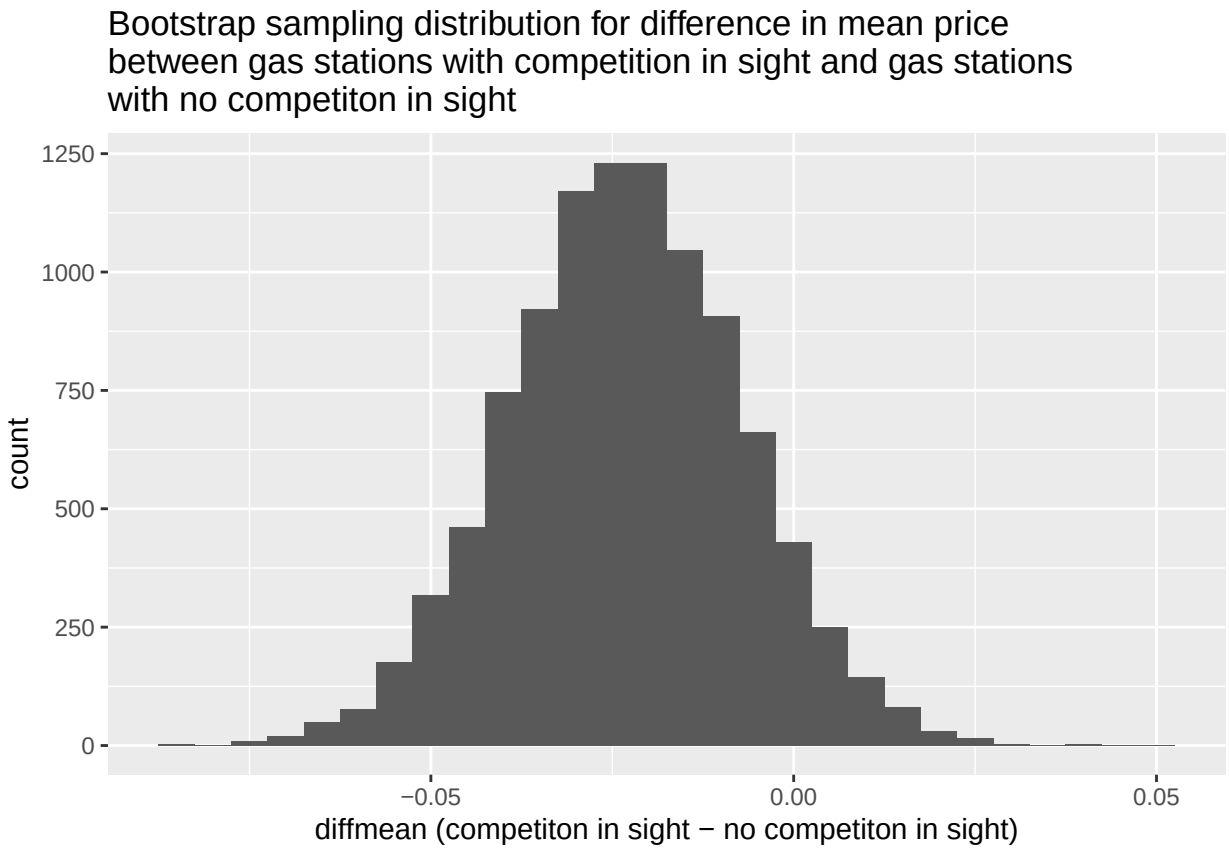
Claim:

Gas stations charge more if they lack direct competition in sight.

Evidence:

```
##        N        Y
## 1.875882 1.852400
```

```
##    diffmean
## -0.02348235
```



```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05468139 0.008361431  0.95 percentile -0.02348235
```

The difference in price between gas stations with competitors in sight and gas stations without competitors in sight is somewhere between -0.05 and 0.01 dollars, with 95% confidence.

Conclusion:

The theory that gas stations that lack direct competition in site charge more is supported by the data because when bootstrapped the difference of mean price has a considerably larger negative bound than positive bound, and is centered on a negative decimal when the difference is the mean price of gas stations with competition subtracted by the mean price of gas stations with no competition. This means that the mean price of gas stations with no competition is consistently higher than the mean price of gas stations with competition.

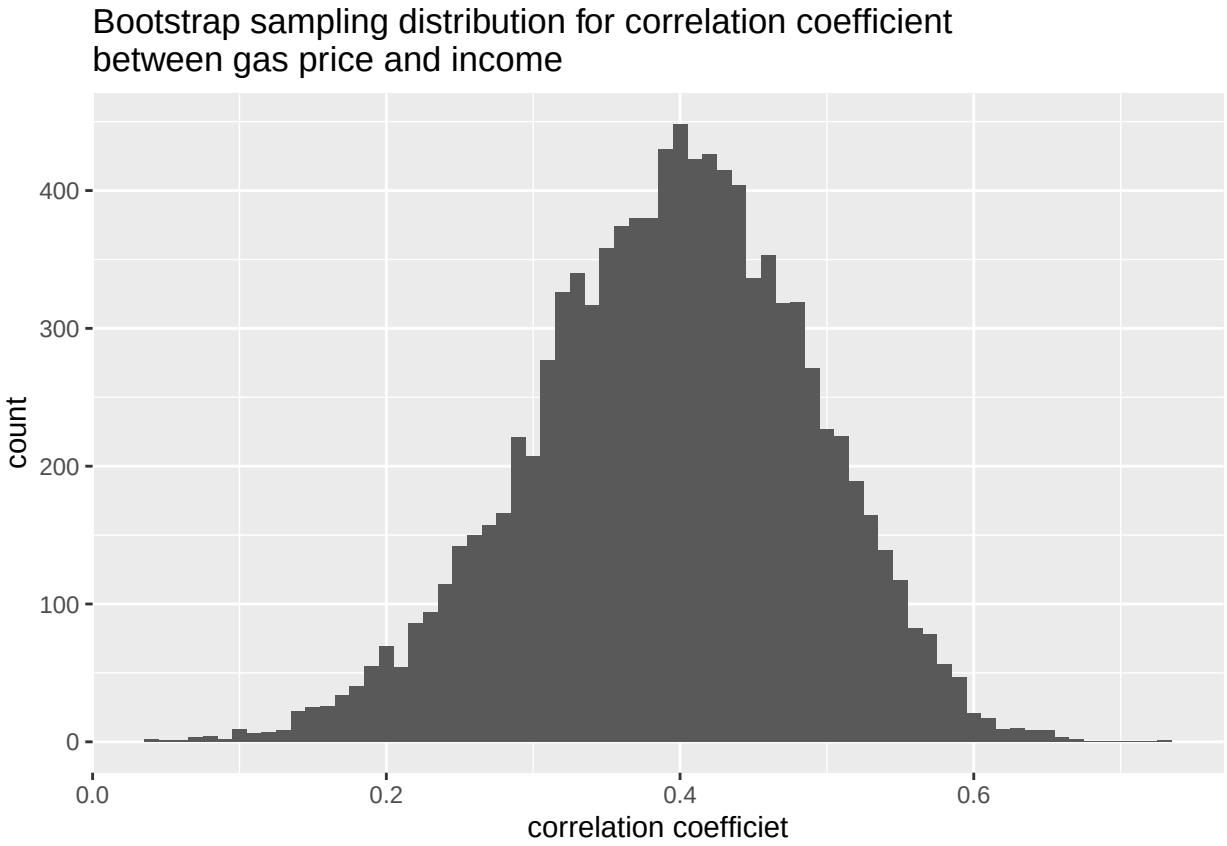
Theory B

Claim:

The richer the area, the higher the gas prices.

Evidence:

```
## [1] 0.3961546
```




```
##   name     lower     upper level     method  estimate
## 1  cor 0.1958088 0.5658386  0.95 percentile 0.3961546
```

The correlation coefficient between gas price and income is somewhere between 0.2 and 0.57, with 95% confidence.

Conclusion:

The theory that the richer the area, the higher the gas prices is supported by the data because when bootstrapped the price and income have a correlation coefficient between 0.2 and 0.57. Since the 95% range only has positive and non-zero decimals the relationship between the price of gas and income of the area or positively or directly related indicating that as income increases so does gas price.

Theory C

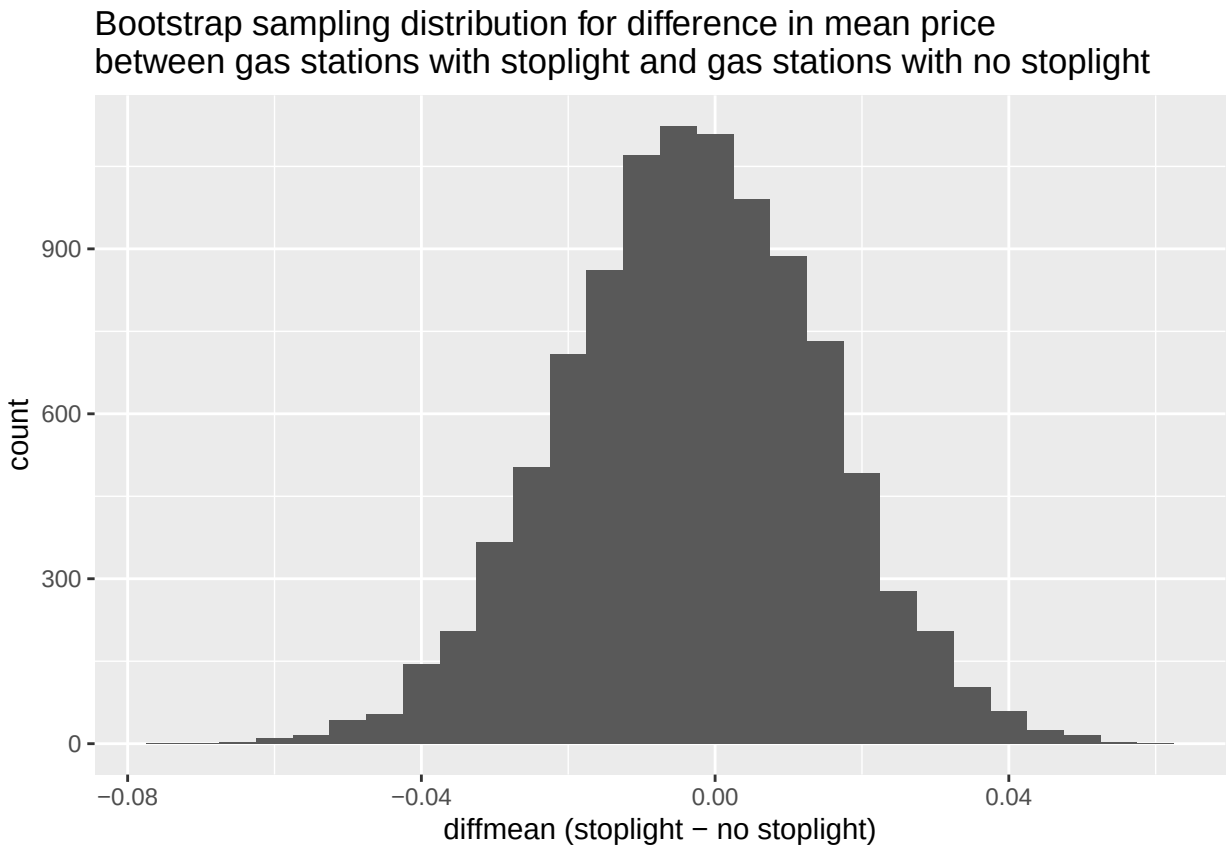
Claim:

Gas stations at stoplights charge more.

Evidence:

```
##        N        Y
## 1.866316 1.863016
```

```
##     diffmean
## -0.003299916
```




```
##       name      lower      upper level     method     estimate
## 1 diffmean -0.0379561 0.03127101  0.95 percentile -0.003299916
```

The theory is unsupported by the data because the distribution for difference is mean price between gas stations with stoplights and gas stations without stoplights is centered around zero meaning overall despite variability the stoplight variable has no affect on the price of gas.

Conclusion:

The theory that gas stations at stoplights charge more is not supported by the data because when bootstrapped the difference of mean price has about the same negative and positive bound, and is centered at zero when the difference is the mean price of gas stations with a stoplight subtracted by the mean price of gas stations with no stoplight. This means that the mean price of gas stations is not affected by whether the gas station has a stoplight.

Theory D

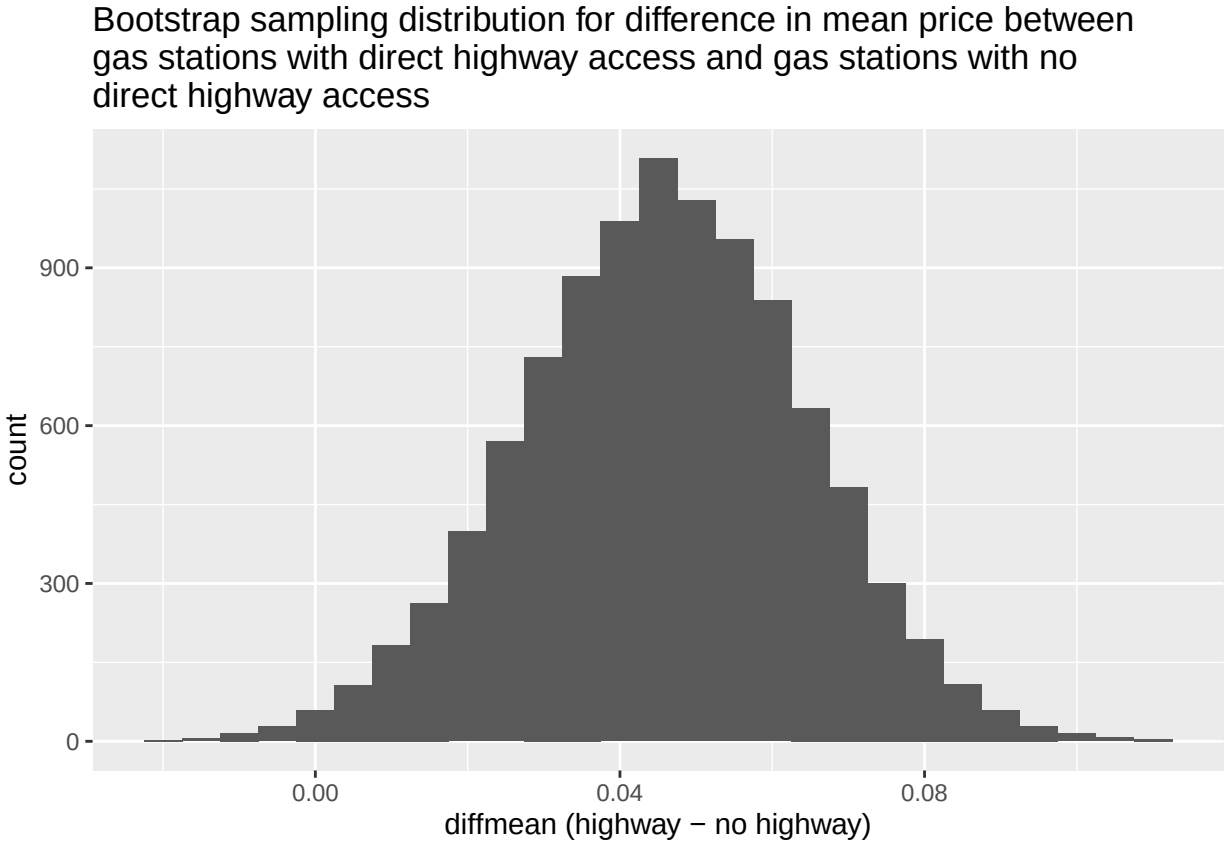
Claim:

Gas stations with direct highway access charge more.

Evidence:

```
##        N        Y
## 1.854304 1.900000
```

```
##  diffmean
## 0.0456962
```




```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.008436699 0.08161745  0.95 percentile 0.0456962
```

The difference in price between gas stations with direct highway access and gas stations without direct highway access is somewhere between 0.008 and 0.082, with 95% confidence.

Conclusion:

The theory that gas stations with direct high access charge more is supported by the data because when bootstrapped the difference of mean price has a considerably larger positive bound than negative bound, and is centered on a positive decimal when the difference is the mean price of gas stations with highway subtracted by the mean price of gas stations with no highway. This means that the mean price of gas stations with direct highway access is consistently higher than the mean price of gas stations without direct highway access.

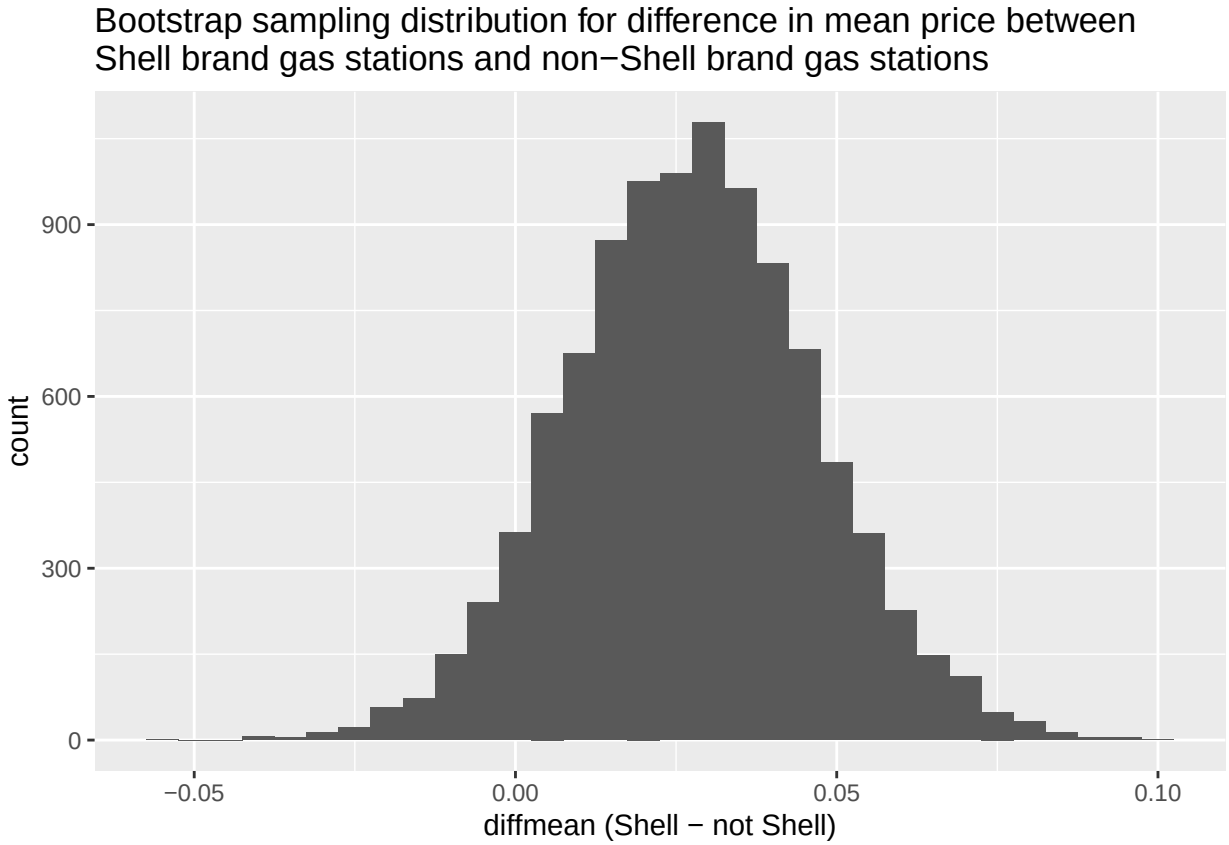
Theory E

Claim:

Shell charges more than all other non-Shell brands.

```
##    FALSE     TRUE
## 1.856389 1.883793
```

```
##   diffmean
## 0.02740421
```




```
##       name        lower      upper level     method   estimate
## 1 diffmean -0.009488232 0.06628881  0.95 percentile 0.02740421
```

The difference in price between Shell brand gas stations and non-Shell brand gas is somewhere between -0.009 and 0.066, with 95% confidence.

Conclusion:

The theory that Shell gas stations charge more than other non-shell brands is supported by the data because when bootstrapped the difference of mean price has a considerably larger positive bound than negative bound,

and is centered on a positive decimal when the difference is the mean price of Shell gas stations subtracted by the mean price of non-Shell gas stations. This means that the mean price of Shell gas stations is consistently higher than the mean price of non-Shell gas stations.

Problem 2

Part A

```
##   name    lower    upper level     method estimate
## 1 mean 26235.41 31802.56  0.95 percentile 28997.34
```

The average mileage of 2011 S-Class 63 AMGs based on this data is somewhere between 26235.4 and 31802.6 miles, with 95% confidence.

Part B

```
##        name     lower    upper level     method  estimate
## 1 prop_TRUE 0.4167532 0.453098  0.95 percentile 0.4347525
```

The proportion of all 2014 S-Class 550s that were painted black based on this data is somewhere between 0.417 and 0.453, with 95% confidence.

Problem 3

Part A

Question:

Which show made people happier: Living with Ed or My name is Earl?

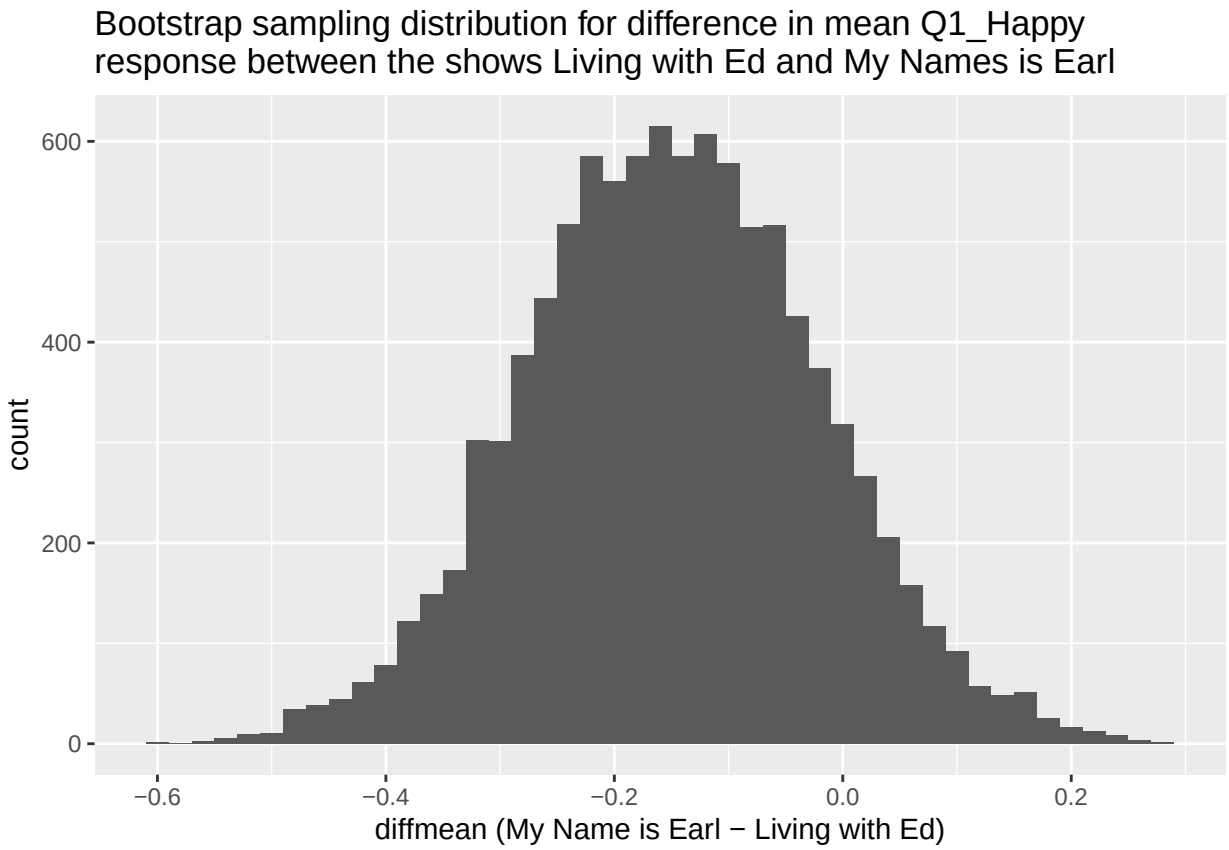
Approach:

I first had to filter the data to only include Living with Ed or My name is Earl. I then bootstrapped meaning resampled the filtered data set taking the difference in means of the Q1_Happy responses 10,000 times. Lastly, I found the 95% confidence interval.

Results:

```
##  Living with Ed My Name is Earl
##        3.926829        3.777778
```

```
##   diffmean
## -0.1490515
```




```
##       name      lower     upper level     method   estimate
## 1 diffmean -0.3981137 0.1024322  0.95 percentile -0.1490515
```

The difference of mean Q1_Happy responses between My Name is Earl and Living with Ed is somewhere between -0.4 and 0.1, with 95% confidence.

Conclusion:

Living with Ed made more people happier than My Name is Earl. The viewers were asked to rate the strength of their agreement on a 1-5 scale (where 5 means "strongly agree") with the statement "This show made me feel happy." This means that the show with the greater mean response made the viewers more happy, as they agreed with the statement more. So since the difference of mean Q1_Happy when bootstrapped has a considerably larger negative bound than positive bound, and is centered on a negative decimal when the difference is the mean Q1_Happy for My Name is Earl subtracted by mean Q1_Happy for Living with Ed. This means that the mean Q1_Happy for Living with Ed is consistently higher than the mean Q1_Happy for My Name is Earl, meaning Living with Ed makes people happier.

Part B

Question:

Which reality/contest show made people feel more annoyed: The Biggest Loser or The Apprentice: Los Angeles?

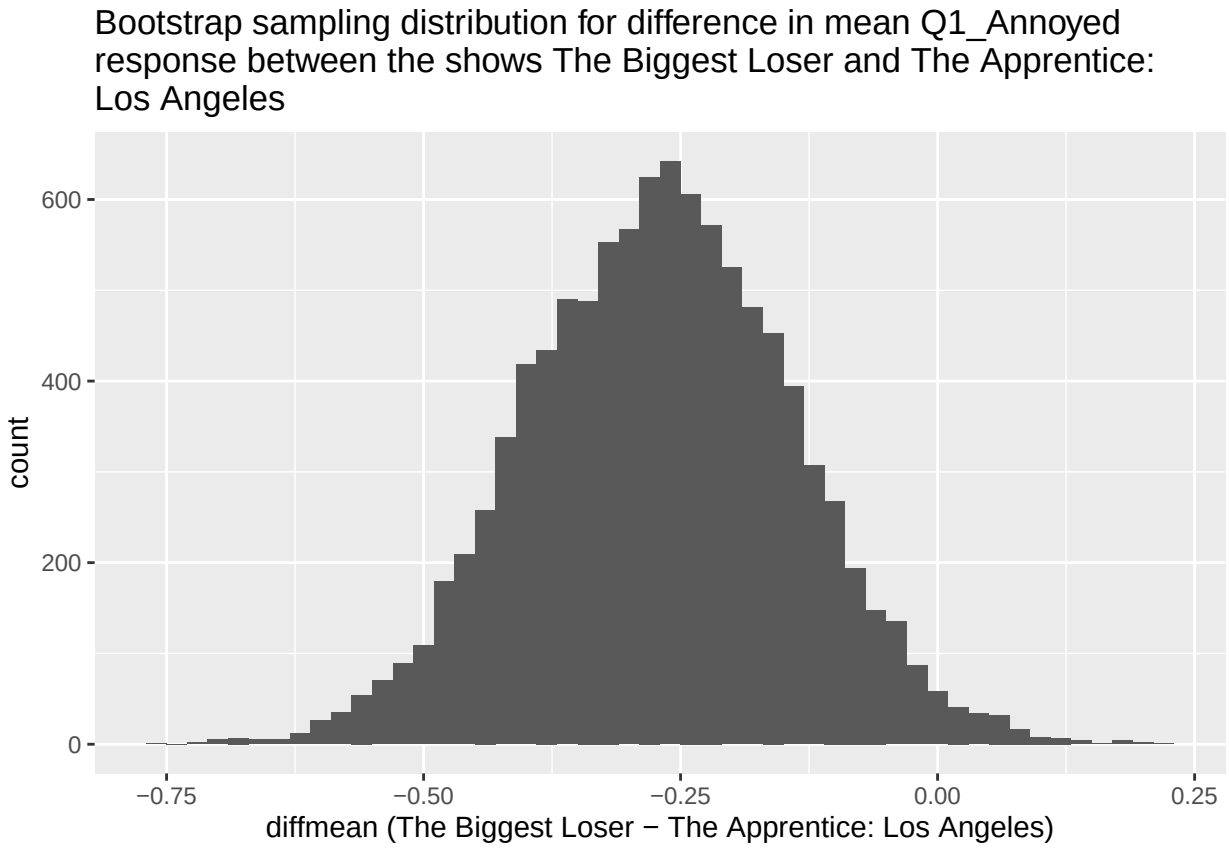
Approach:

I first had to filter the data to only include The Biggest Loser and The Apprentice: Los Angeles. I then bootstrapped meaning resampled the filtered data set taking the difference in means of the Q1_Annoyed responses 10,000 times. Lastly, I found the 95% confidence interval.

Results:

```
## The Apprentice: Los Angeles           The Biggest Loser
##                    2.307229                    2.036232
```

```
##  diffmean
## -0.270997
```




```
##       name      lower       upper level     method  estimate
## 1 diffmean -0.5231823 -0.02170233  0.95 percentile -0.270997
```

The difference of mean Q1_Annoyed responses between The Biggest Loser and The Apprentice: Los Angeles is somewhere between -0.52 and -0.02, with 95% confidence.

Conclusion:

The Apprentice: Los Angeles made more people annoyed than The Biggest Loser. The viewers were asked to rate the strength of their agreement on a 1-5 scale (where 5 means "strongly agree") with the statement "This show made me feel annoyed." This means that the show with the greater mean response made the viewers more annoyed, as they agreed with the statement more. So since the difference of mean Q1_Annoyed when bootstrapped has all negative bounds in 95% range, and is centered on a negative decimal when the difference is the mean Q1_Annoyed for The Biggest Loser subtracted by mean Q1_Annoyed for The Apprentice: Los Angeles. This means that the mean Q1_Annoyed for The Apprentice: Los Angeles is consistently higher than the mean Q1_Annoyed for The Biggest Loser, meaning The Apprentice: Los Angeles makes people more annoyed.

Part C

Question:

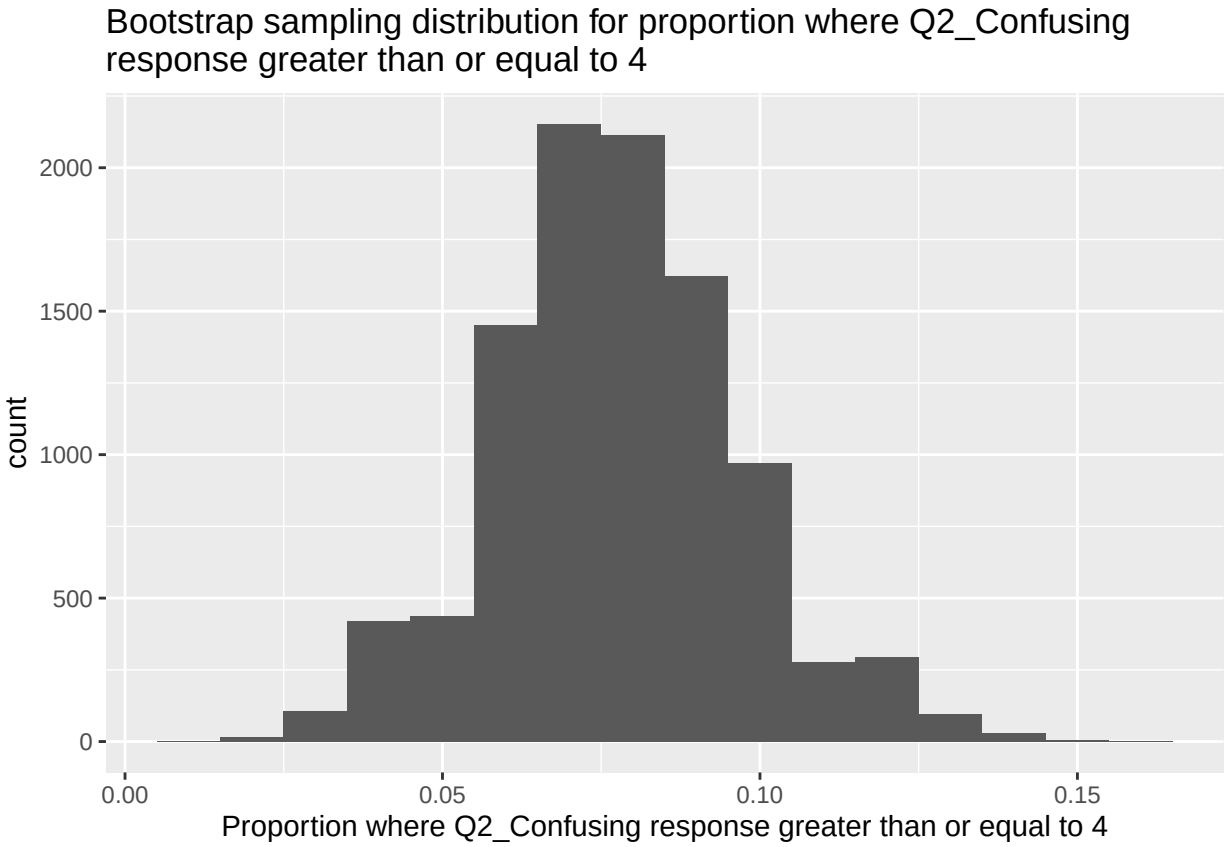
What proportion of American TV watchers would we expect to give a response of 4 or greater to the "Q2_Confusing" question?

Approach:

I first had to filter the data to only Dancing with the Stars. Next, I needed to make a new variable to either be true if Q2_Confusing was greater than or equal to 4 or false if Q2_Confusing was less than 4. I then bootstrapped meaning resampled the filtered data set taking the proportion of true responses of that new variable 10,000 times. Lastly, I found the 95% confidence interval.

Results:

```
##  prop_TRUE
## 0.07734807
```




```
##        name      lower     upper level     method   estimate
## 1 prop_TRUE 0.03867403 0.1160221  0.95 percentile 0.07734807
```

The proportion of viewers with a Q2_Confusing response greater than or equal to 4 is somewhere between 0.039 and 0.116, with 95% confidence.

Conclusion:

The viewers were asked to rate the strength of their agreement on a 1-5 scale (where 5 means "strongly agree") with the statement "I found this show confusing." Based on this sample of respondents, we would

expect to get a response of 4 or greater to the Q2_Confusing question from somewhere between 0.039 and 0.116, with 95% confidence.

Problem 4

Question:

Does the extra traffic brought to our site from paid search results—above and beyond what we'd see if we "went organic"—justify the cost of the ads themselves?

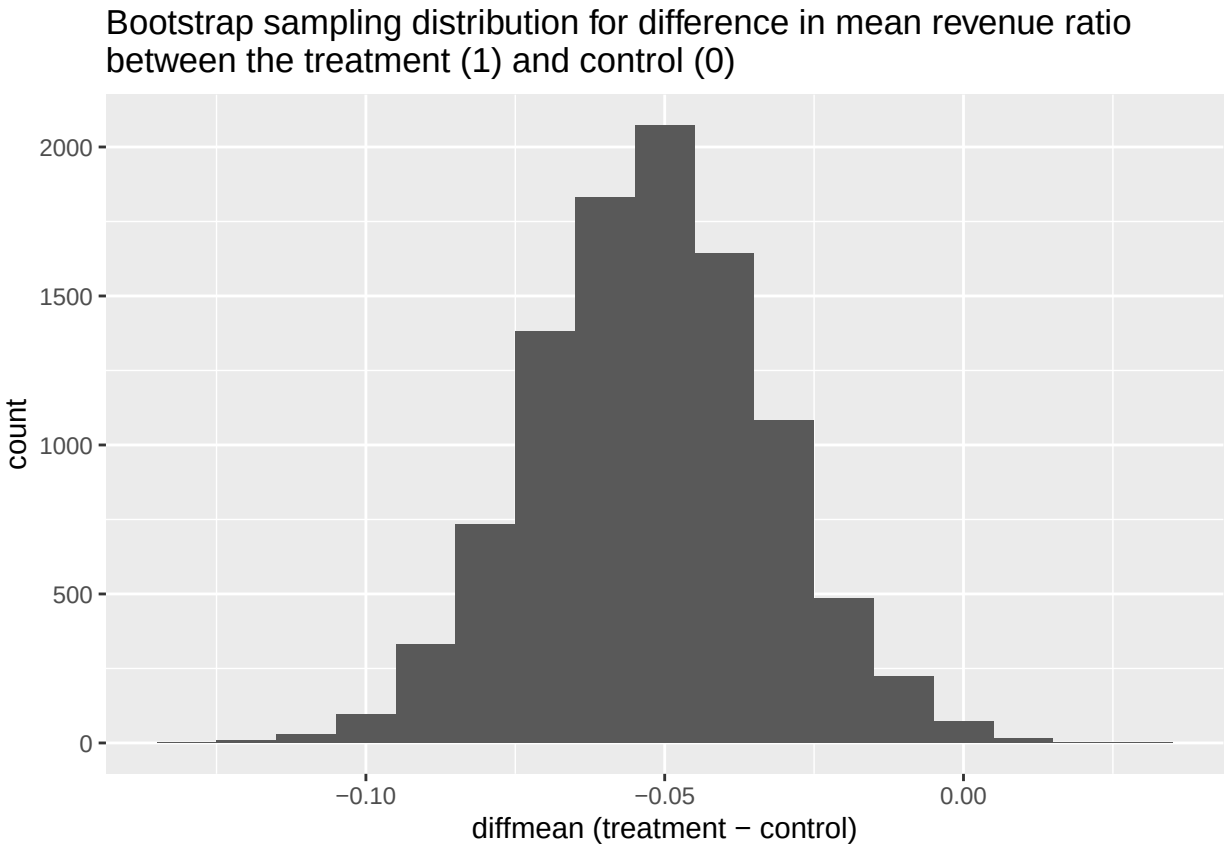
Approach:

I first needed to make a new variable for the revenue ratio which was revenue before the experiment started (30 days before May 22) divided by revenue after the experiment started (30 days beginning on May 22). I then bootstrapped meaning resampled the the data set taking the difference of the mean revenue ratio between the treatment and control group 10,000 times, by Monte Carlo simulations. Lastly, I found the 95% confidence interval.

Results

```
##         0         1
## 0.9488775 0.8965961
```

```
##    diffmean
## -0.05228145
```




```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.09052436 -0.01306996  0.95 percentile -0.05228145
```

The difference of mean revenue ratio between the treatment group and control group is somewhere between -0.091 and -0.013, with 95% confidence.

Conclusion:

Paying for the adds is supported by the data because when bootstrapped the difference of mean revenue ratio has all negative bounds in 95% range, and is centered on a negative decimal when the difference is the mean revenue ratio of the treatment group subtracted by the mean revenue ratio of the control grouo. This means that the mean revenue ratio of the control group is consistently higher than revenue ratio of the treatment group. The treatment group, is where advertising on Google AdWords for the whole DMA was paused for a month starting on May 22, and the control group, is where advertising on Google AdWords continued as before. So in conclusion, going organic decreased the revenue ratio by on average around 5%.