HW 3

Maggie Ma

2025-02-10

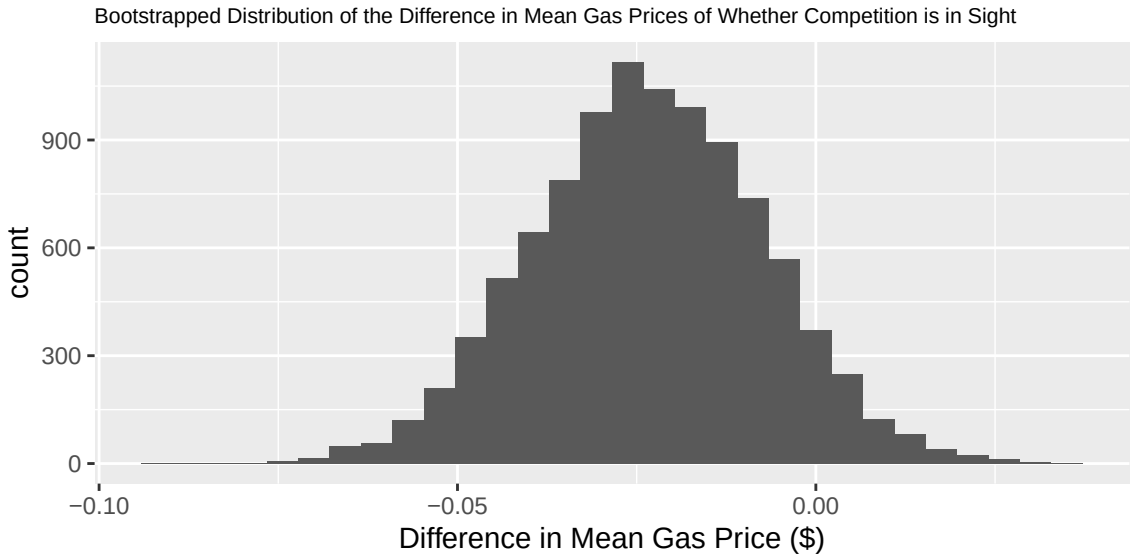
Maggie Ma, mm227339, Github Link:

Problem 1

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

```
## # A tibble: 2 x 2
##   Competitors mean_price
##   <chr>            <dbl>
## 1 N                 1.88
## 2 Y                 1.85
```

Table of Original Sample's Mean Prices



Graph of Bootstrapped Samples' Difference in Means

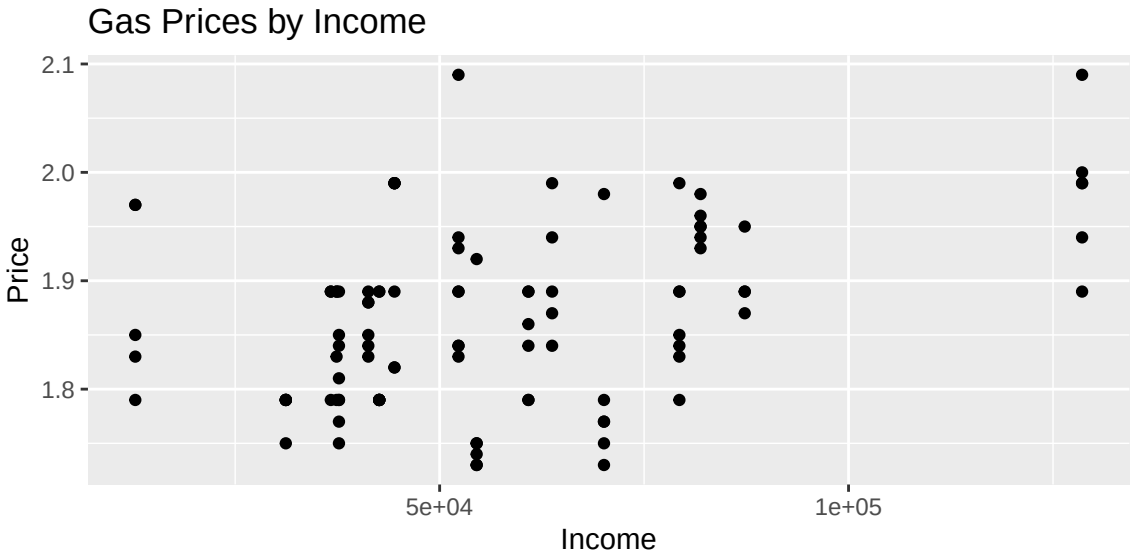
```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05483117 0.007663557  0.95 percentile -0.02348235
```

Confidence Interval

In the original sample, the mean price between the gas stations with and without competition in sight differ by very little. When bootstrapping the difference in the mean gas prices between the gas stations with and without competition in sight, we can see that the mean difference in the gas prices is somewhere between -0.054 and 0.007, with 95% confidence. Since the confidence interval does contain 0, the mean difference is

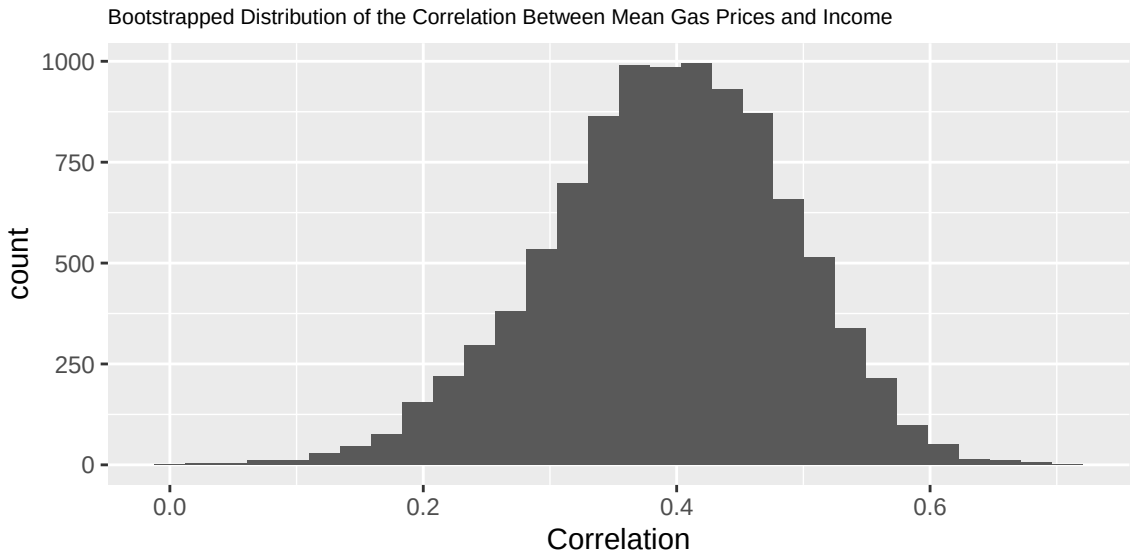
not statistically significant at the 5% level, and there is no substantial evidence that gas stations charge more if they lack direct competition in sight.

Claim: The richer the area, the higher the gas prices.




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

Graph of Original Sample's Prices by Income



Graph of Bootstrapped Sample Correlations

```
##   name     lower     upper level     method  estimate
## 1  cor 0.1960682 0.5651556  0.95 percentile 0.3961546
```

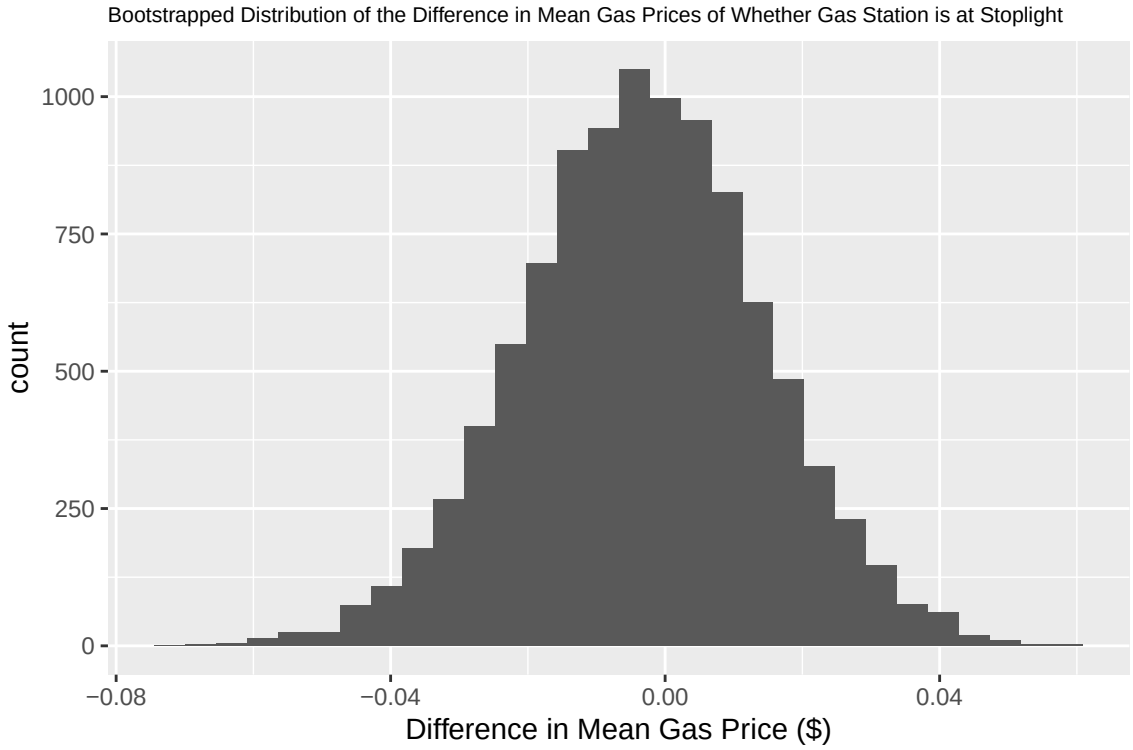
Confidence Interval

In the original sample, the gas prices and the income levels did not have a very strong correlation, with a correlation of 0.396. When bootstrapping the correlation between the gas prices and income, we can see that the correlation between gas prices and income is somewhere between 0.196 and 0.565, with 95% confidence. This range tells us that the correlation is not strong at all, and there is no substantial evidence that the richer the area, the higher the gas prices.

Claim: Gas stations at stoplights charge more.

```
## # A tibble: 2 x 2
##   Stoplight mean_price
##   <chr>          <dbl>
## 1 N               1.87
## 2 Y               1.86
```

Table of Original Sample's Mean Prices



Graph of Bootstrapped Samples' Difference in Means

```
##       name       lower      upper level     method     estimate
## 1 diffmean -0.03838725 0.03118259  0.95 percentile -0.003299916
```

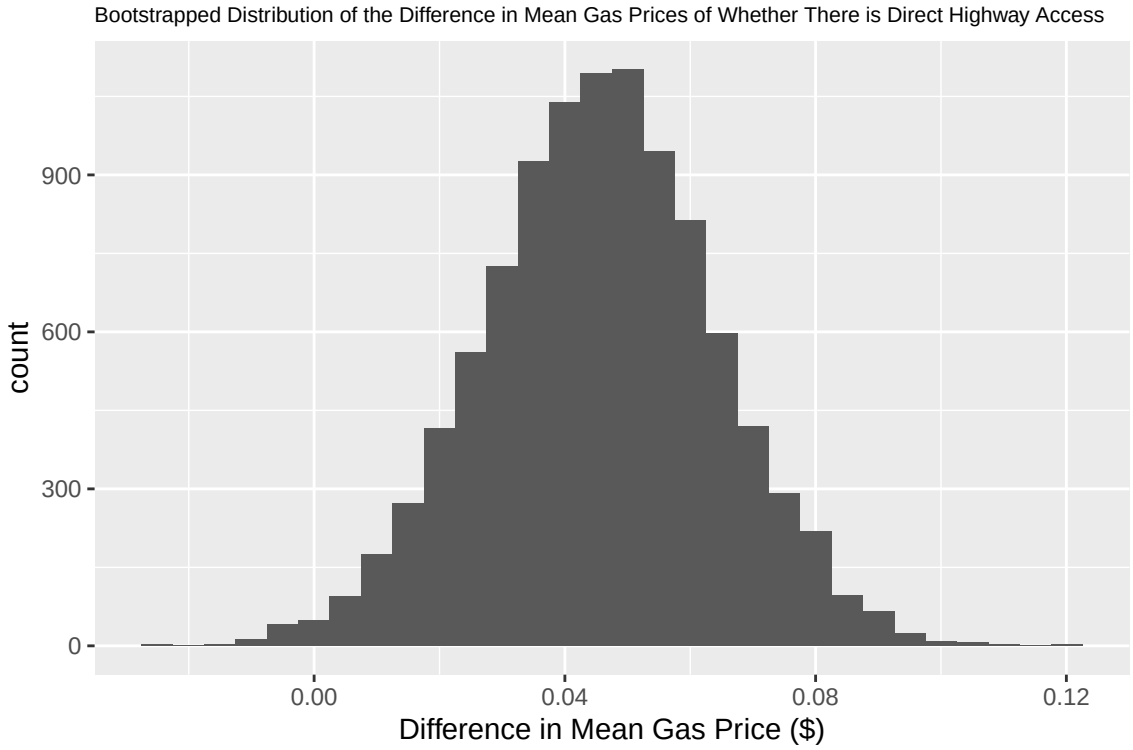
Confidence Interval

In the original sample, the mean price between the gas stations both at and not at a stoplight differ by very little. When bootstrapping the difference in the mean gas prices between the gas stations at/not at stoplights, we can see that the mean difference in the gas prices is somewhere between -0.038 and 0.031, with 95% confidence. Since the confidence interval does contain 0, the mean difference is not statistically significant at the 5% level, and there is no substantial evidence that gas stations at stoplights charge more.

Claim: Gas stations with direct highway access charge more.

```
## # A tibble: 2 x 2
##   Highway mean_price
##   <chr>        <dbl>
## 1 N             1.85
## 2 Y             1.9
```

Table of Original Sample's Mean Prices



Graph of Bootstrapped Samples' Difference in Means

```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.009491074 0.08122717  0.95 percentile 0.0456962
```

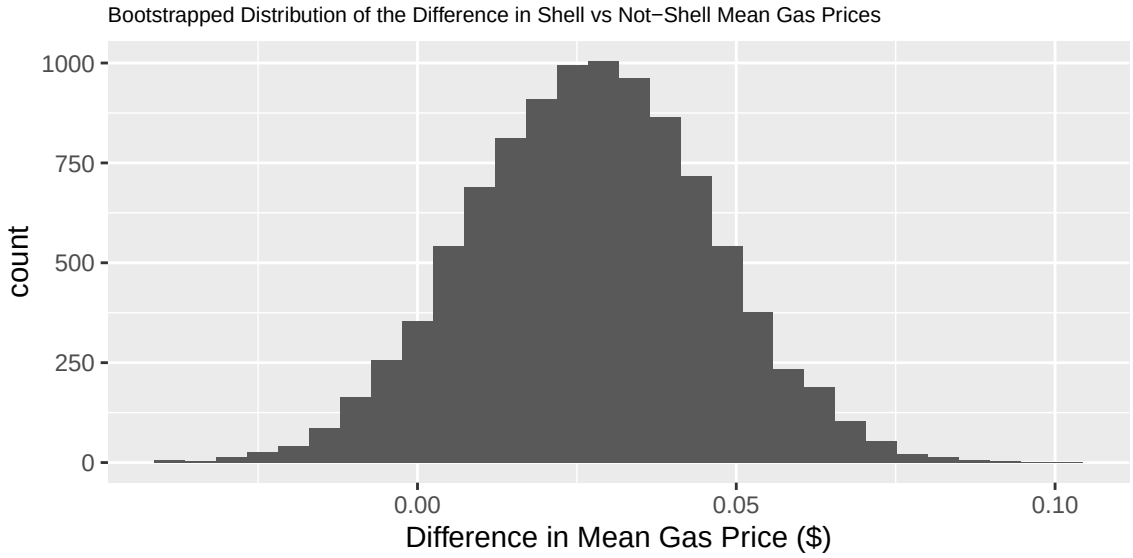
Confidence Interval

In the original sample, the mean price between the gas stations with and without direct highway access differ slightly. When bootstrapping the difference in the mean gas prices between the gas stations with and without direct highway access, we can see that the mean difference in the gas prices is somewhere between 0.0009 and 0.0812, with 95% confidence. Since the confidence interval does not contain 0, the mean difference is statistically significant at the 5% level, and there is evidence that gas stations with direct highway access charge more.

Claim: Shell charges more than all other non-Shell brands.

```
## # A tibble: 2 x 2
##   isShell mean_price
##   <lgl>        <dbl>
## 1 FALSE         1.86
## 2 TRUE          1.88
```

Table of Original Sample's Mean Prices



Graph of Bootstrapped Samples' Difference in Means

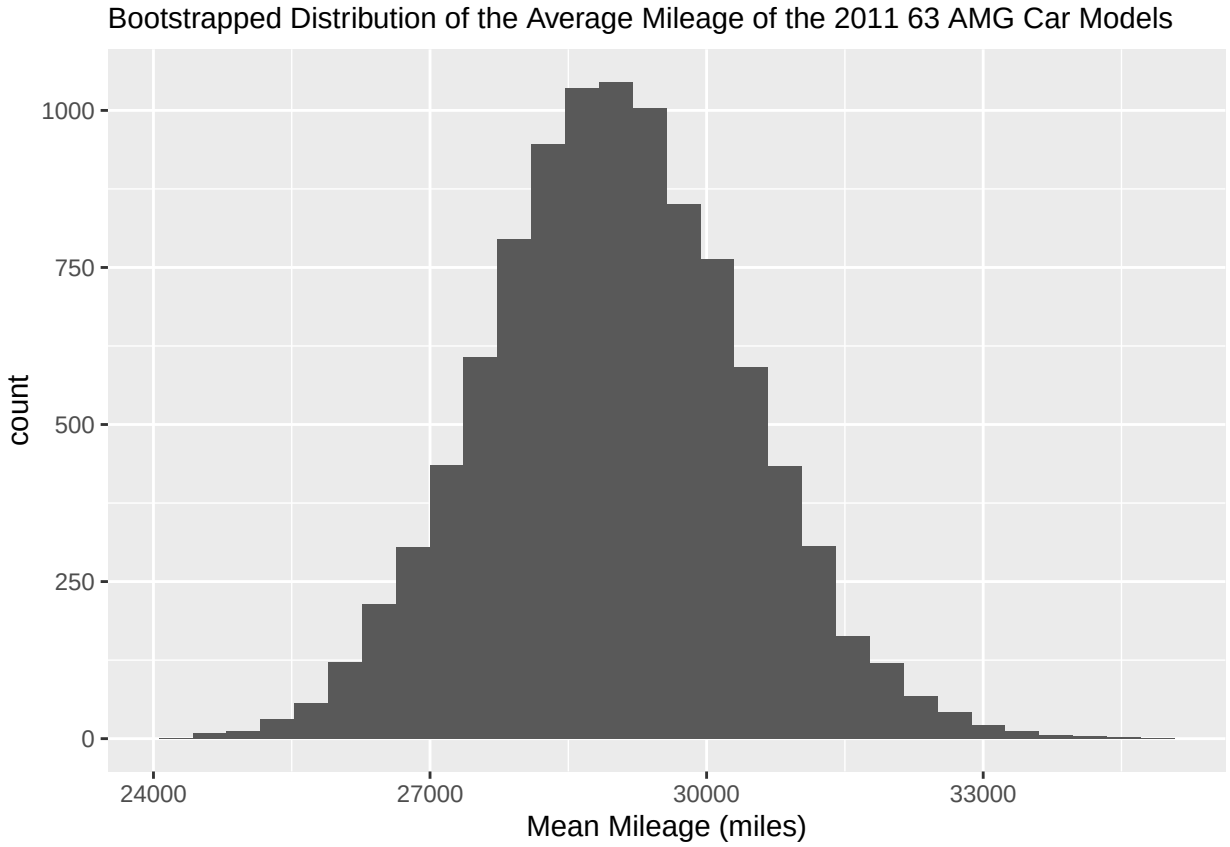
```
##       name       lower      upper level     method   estimate
## 1 diffmean -0.00982087 0.06419789  0.95 percentile 0.02740421
```

Confidence Interval

In the original sample, the mean price for Shell gas stations are indeed slightly higher. However, when bootstrapping the difference in the mean gas prices between the gas stations with and without direct highway access, we can see that the mean difference in the gas prices is somewhere between -0.0009 and 0.0642, with 95% confidence. Since the confidence interval does contain 0, the mean difference is not statistically significant at the 5% level, and there is no substantial evidence that Shell charges more than all other non-Shell brands.

Problem 2

Part A



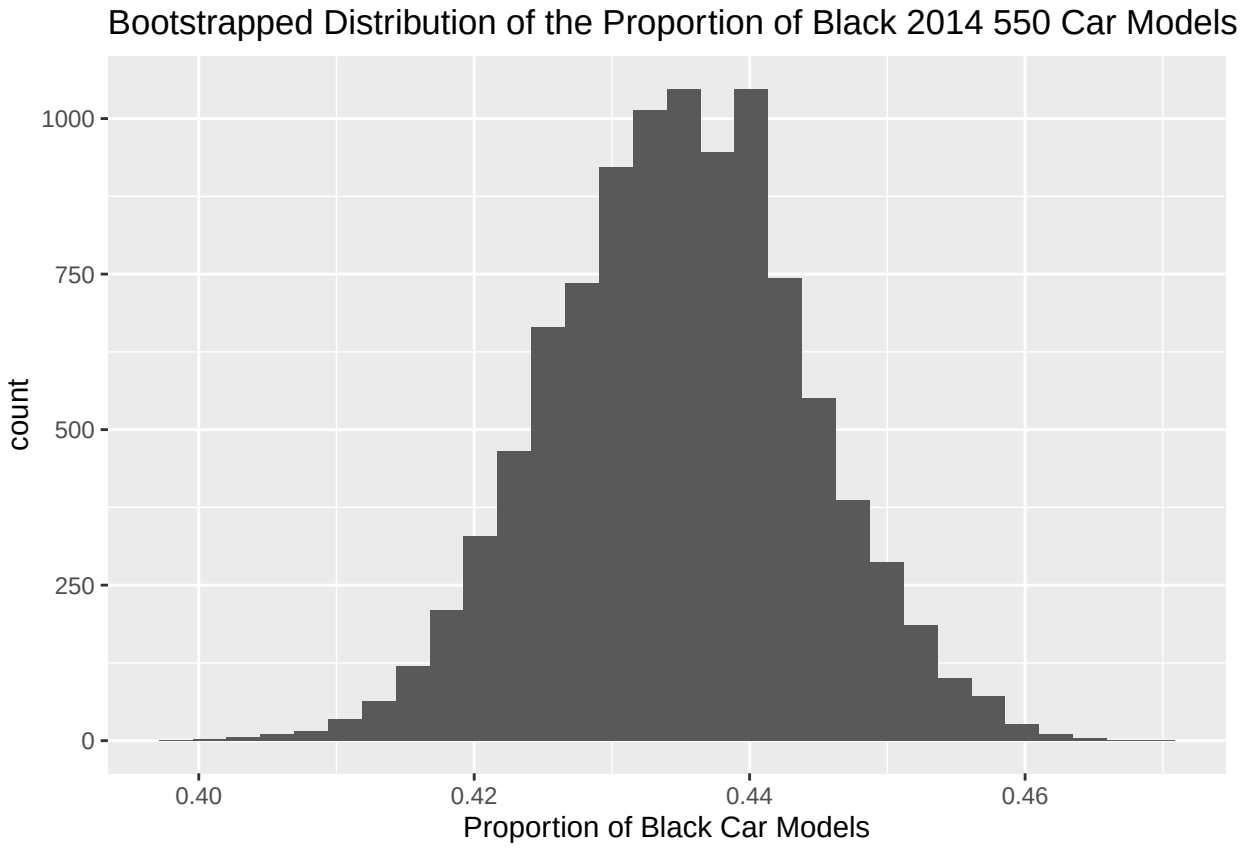
Graph of Bootstrapped Sample Means

```
##   name    lower    upper level     method estimate
## 1 mean 26299.28 31825.47  0.95 percentile 28997.34
```

Confidence Interval

The average mileage of 2011 S-Class 63 AMGs that were hitting the used-car market is somewhere between 26299.28 miles and 31825.47 miles.

Part B



Graph of Bootstrapped Sample Proportions

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

Confidence Interval

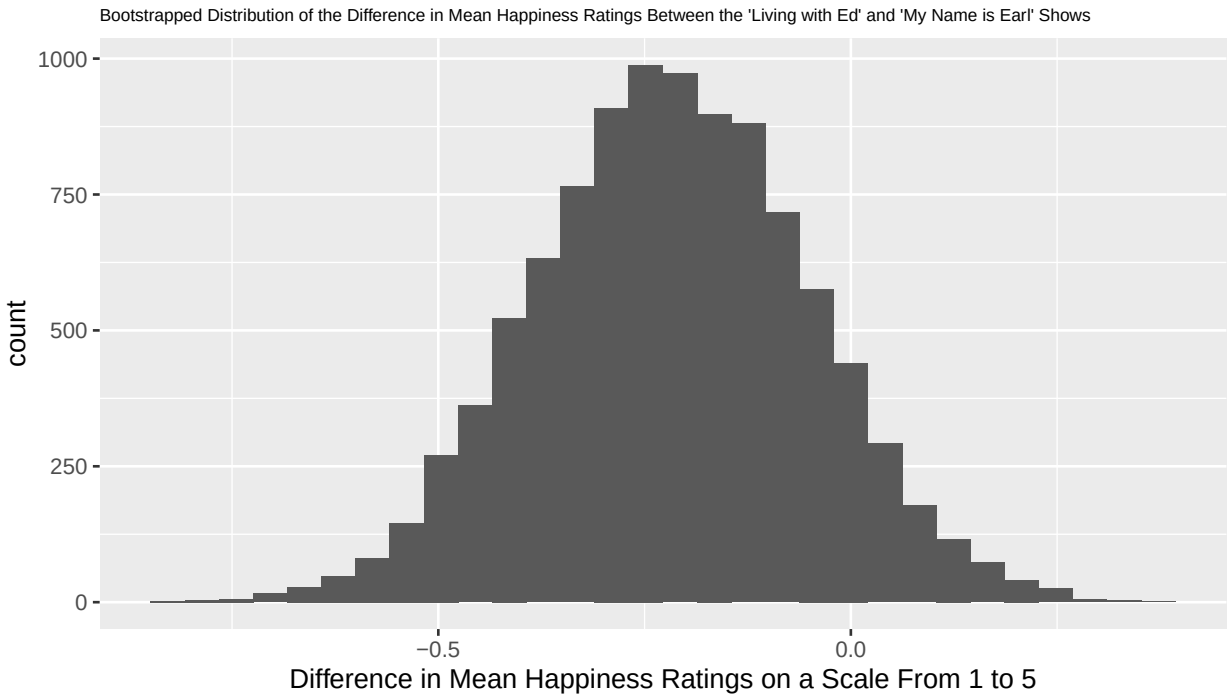
The proportion of 2014 S-Class 550s that were painted black is somewhere between 0.4168 and 0.4531.

Problem 3

Part A

```
## # A tibble: 2 x 2
##   Show            mean_happy_rating
##   <chr>                       <dbl>
## 1 Living with Ed               4.05
## 2 My Name is Earl              3.83
```

Table of Original Sample's Mean Happiness Ratings for Both Shows



Graph of Bootstrapped Samples' Difference in Means

```
##       name      lower     upper level     method   estimate
## 1 diffmean -0.5351187 0.1056923  0.95 percentile -0.2176117
```

Confidence Interval

The question I am answering is: Is there evidence that one show consistently produces a higher mean Q1_Happy response among viewers?

I approached this problem by looking at the mean ratings in the original sample first to get an idea of what I am working with, then bootstrapping the sample and graphing it to determine my confidence interval.

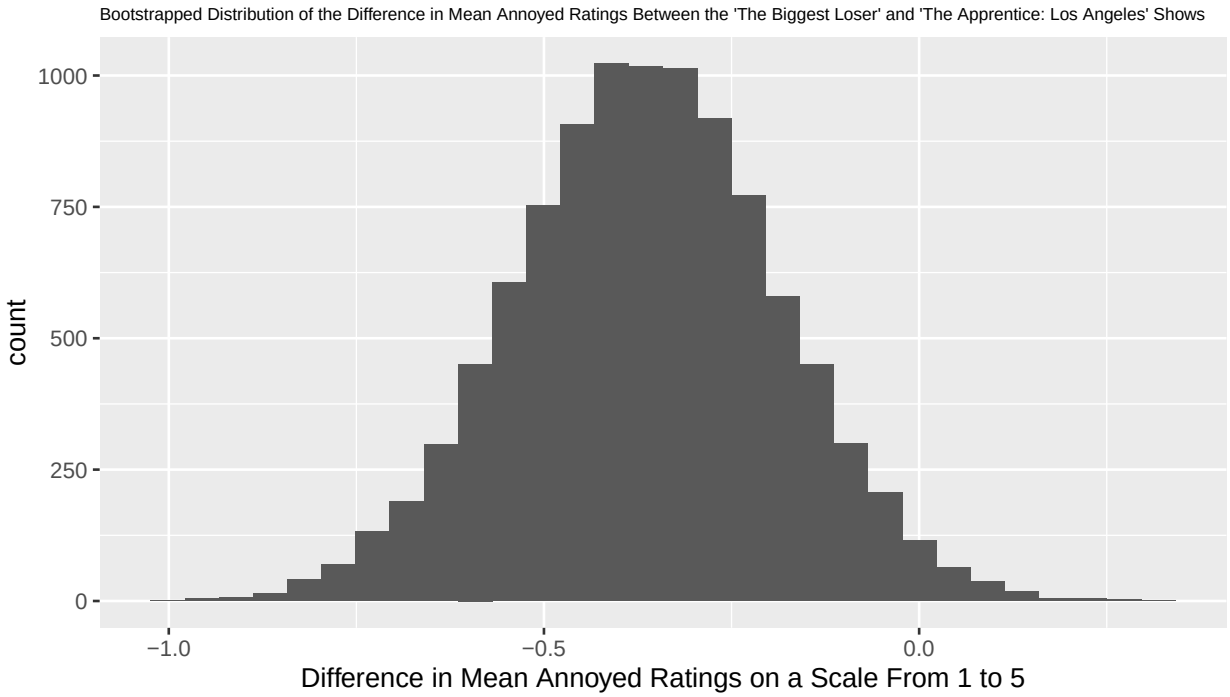
In the original sample, the mean happiness rating of "Living with Ed" is about 0.2 higher than "My Name is Earl". When bootstrapping the difference in the mean happiness ratings between the two shows, we can see that the mean difference in happiness ratings is somewhere between -0.535 and 0.106, with 95% confidence.

Since the confidence interval does contain 0, the mean difference is not statistically significant at the 5% level, which means there is no conclusive evidence that one show consistently produces a higher mean happiness rating among viewers than the other.

Part B

```
## # A tibble: 2 x 2
##   Show                          mean_annoyed_rating
##   <chr>                                       <dbl>
## 1 The Apprentice: Los Angeles                  2.38
## 2 The Biggest Loser                            2.01
```

Table of Original Sample's Mean Annoyed Ratings for Both Shows



Graph of Bootstrapped Samples' Difference in Means

```
##       name     lower       upper level     method estimate
## 1 diffmean -0.710675 -0.02360671  0.95 percentile -0.36329
```

Confidence Interval

The question I am answering is: Is there evidence that one show consistently produces a higher mean Q1_Annoyed response among viewers?

I approached this problem by looking at the mean ratings in the original sample first to get an idea of what I am working with using a table, then bootstrapping the sample and graphing it to determine my confidence interval.

In the original sample, the mean annoyed rating of "The Biggest Loser" is about 0.3 lower than "The Apprentice: Los Angeles". When bootstrapping the difference in the mean annoyed ratings between the two shows, we can see that the mean difference in annoyed ratings is somewhere between -0.710 and -0.024, with 95% confidence.

Since the confidence interval does not contain 0, the mean difference is statistically significant at the 5% level, which means there is evidence that the "The Apprentice: Los Angeles" show consistently produces a higher mean annoyed rating among viewers than the other.

Part C

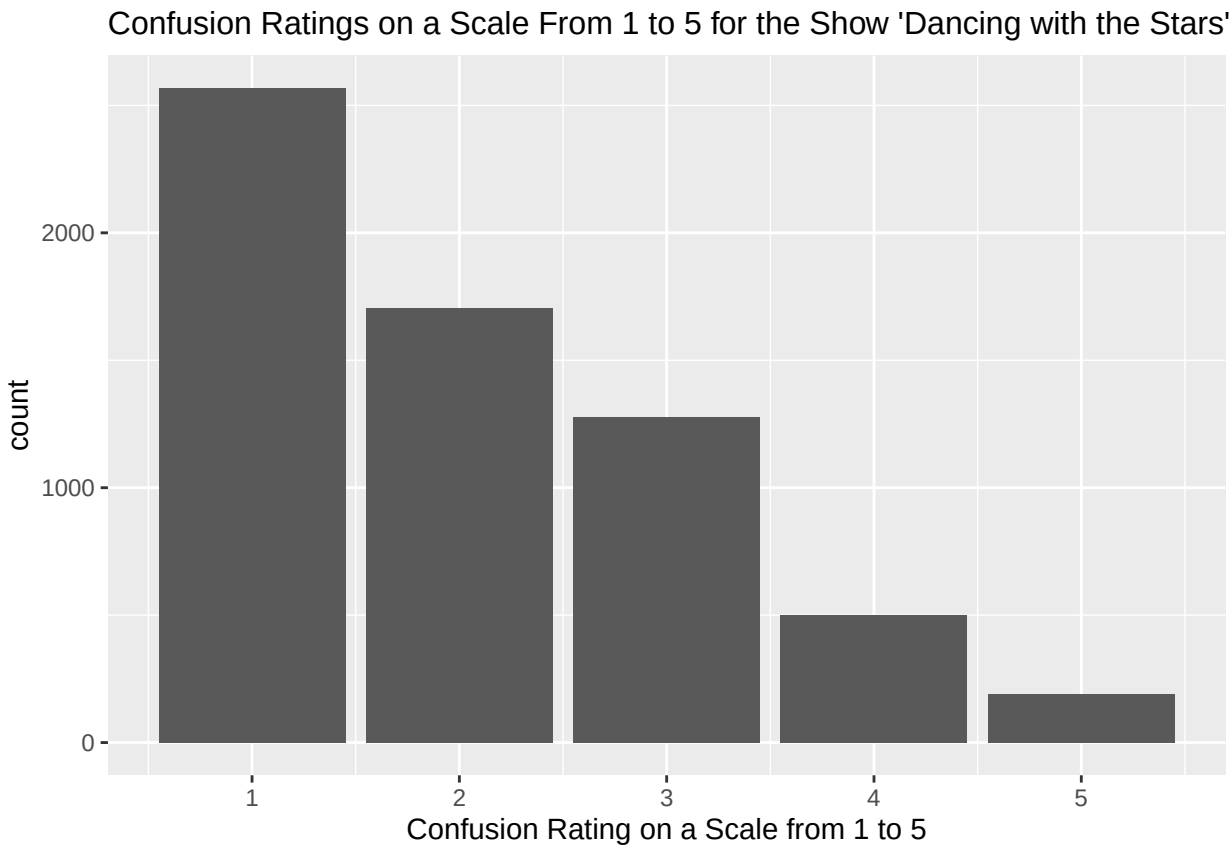
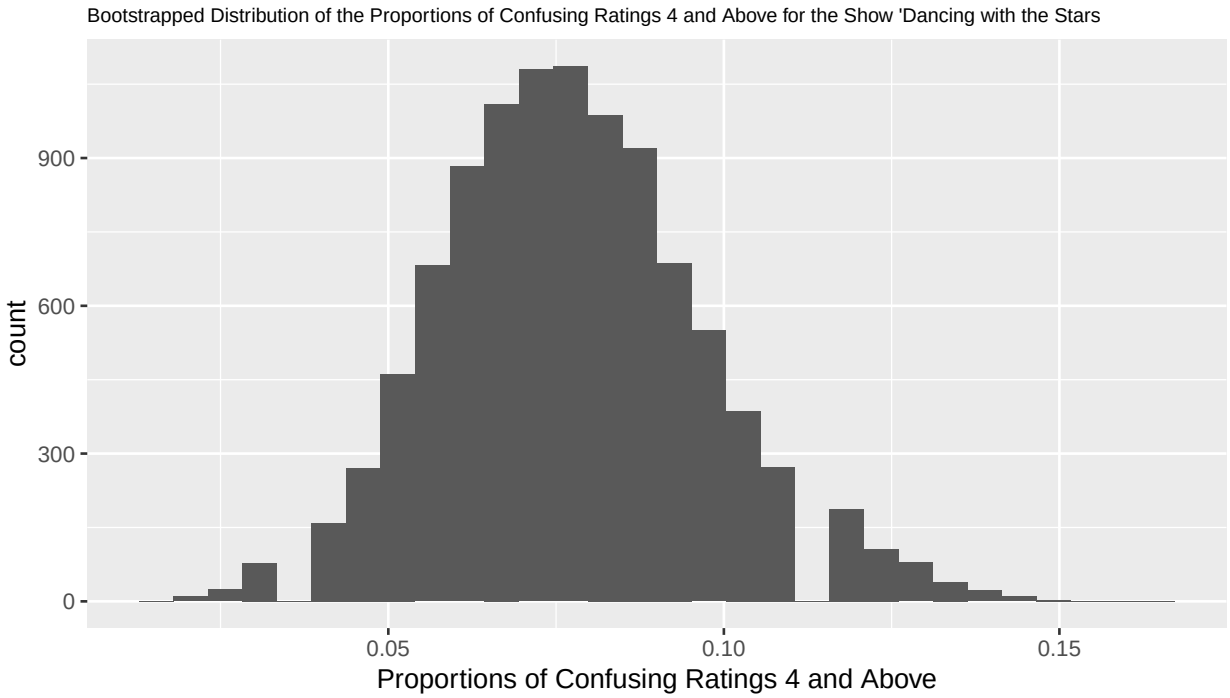


Table of Original Sample's Mean Confused Ratings



Graph of Bootstrapped Samples' Difference in Means

```
##        name      lower    upper level     method   estimate
## 1 prop_TRUE 0.03867403 0.121547  0.95 percentile 0.07734807
```

Confidence Interval

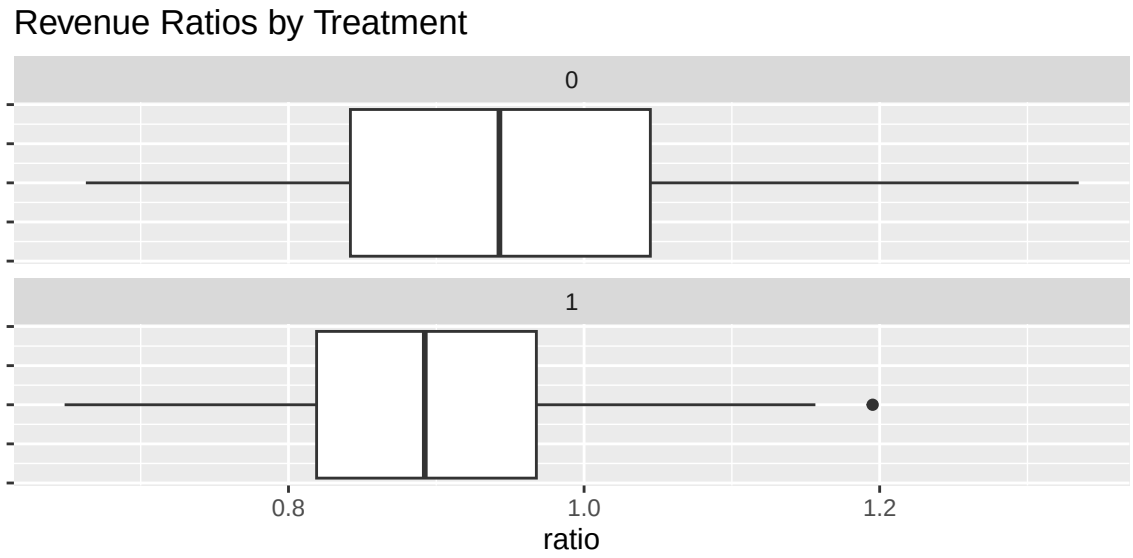
The question I am answering is: What proportion of American TV watchers would we expect to give a response of 4 or greater to the "Q2_Confusing" question?

I approached this problem by looking at the mean confused ratings in the original sample first to get an idea of what I am working with using a bar graph, then bootstrapping the sample and graphing it to determine my confidence interval.

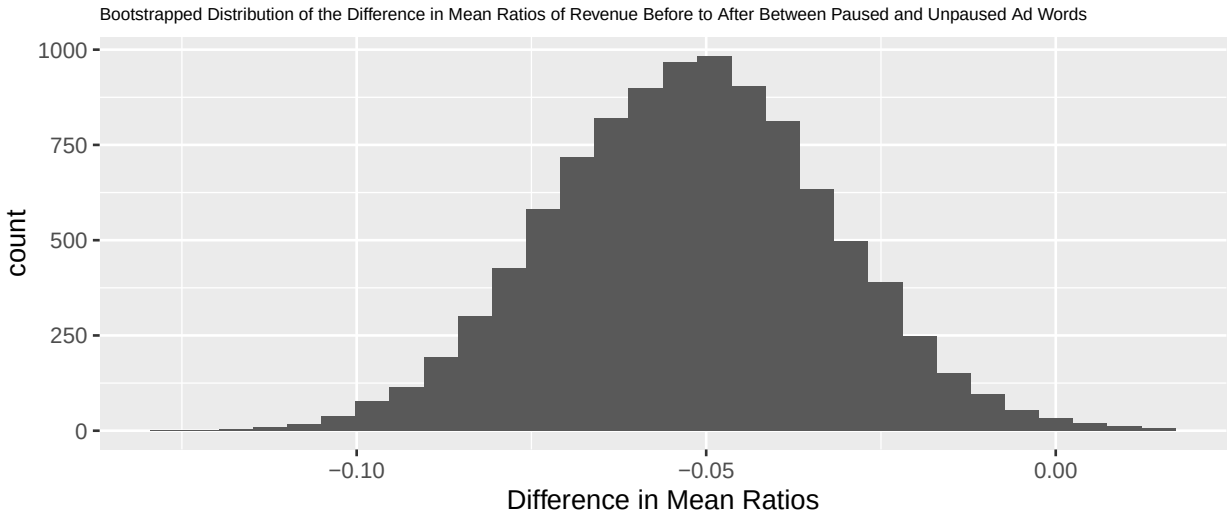
In the original sample, the mean confused ratings of "Dancing with the Stars" is not very high, about 2.045. When bootstrapping the proportion of viewers who rated the show a 4 or 5 on the confusing aspect of the show, we can see that the proportion of confused ratings over 4 is somewhere between 0.038 and 0.122, with 95% confidence.

The confidence interval shows that the proportion of viewers who rate the confusion of the "Dancing with the Stars" show's 4 and above will be between 0.038 and 0.122 95% of the time.

Problem 4



Graph of Original Sample's Mean Ratios (0 = Control Group, 1 = Treatment Group)



Graph of Bootstrapped Samples' Difference in Means

```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.09046619 -0.01315629  0.95 percentile -0.05228145
```

The question I'm trying to answer is whether the revenue ratio is the same in the treatment and control groups, or whether instead the data favors the idea that paid search advertising on Google creates extra revenue for EBay.

I approached this problem by creating a new variable for the revenue ratio, looking at the distribution of the revenue ratios based on their treatment type in the original sample to get an idea of what I am working with using a box plot, then bootstrapping the sample and graphing it to determine my confidence interval.

In the original sample, the median revenue ratio of the control group is 0.05 higher than the treatment group. When bootstrapping the difference in the mean ratios, we can see that the mean difference is somewhere between -0.090 and -0.013, with 95% confidence.

Since the confidence interval does not contain 0, the mean difference is statistically significant at the 5% level, which means there is evidence that the paid search advertising generates more revenue for Ebay.