Homework 2

Trent Latz

2025-02-09

Name: Trent Latz **UT EID:** tjl2597 **GitHub Link:** https://github.com/trentjlatz/SDS315-HW3

Problem 1:

Which of these theories seem true, and which are unsupported by data? Take each theory one by one and assess the evidence for the theory in this data set.

Theory A:

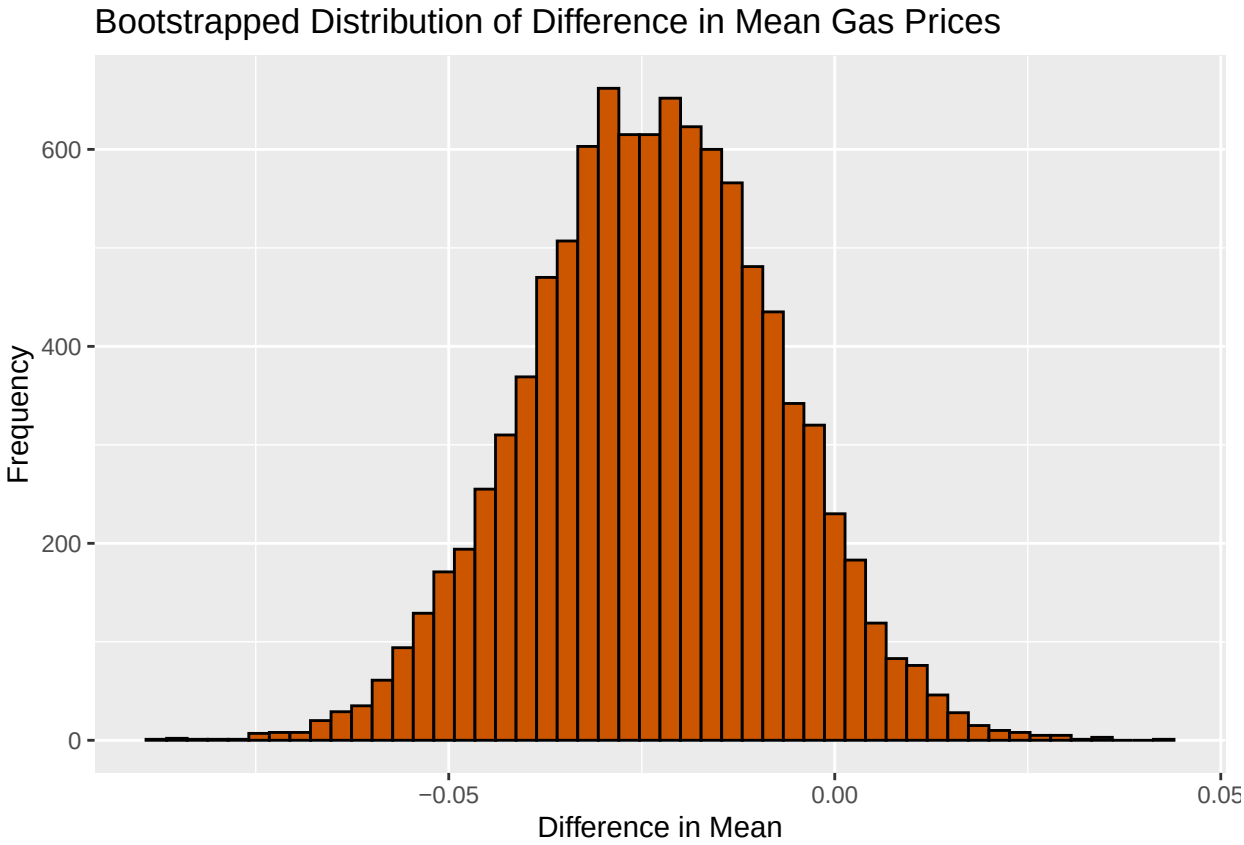
Claim:

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

Evidence:

We will compare average gas price between stations with and without competitors, then bootstrap for accuracy.

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




```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05513981 0.007615941  0.95 percentile -0.02348235
```

Conclusion:

The confidence interval for the difference in mean gas prices between those with and without competition is (-5.41 cents, .87 cents) with 95% confidence. Since the interval includes zero, we cannot reject the null hypothesis. The presence of visible competition does not appear to have a significant effect on the mean gas prices based on this analysis.

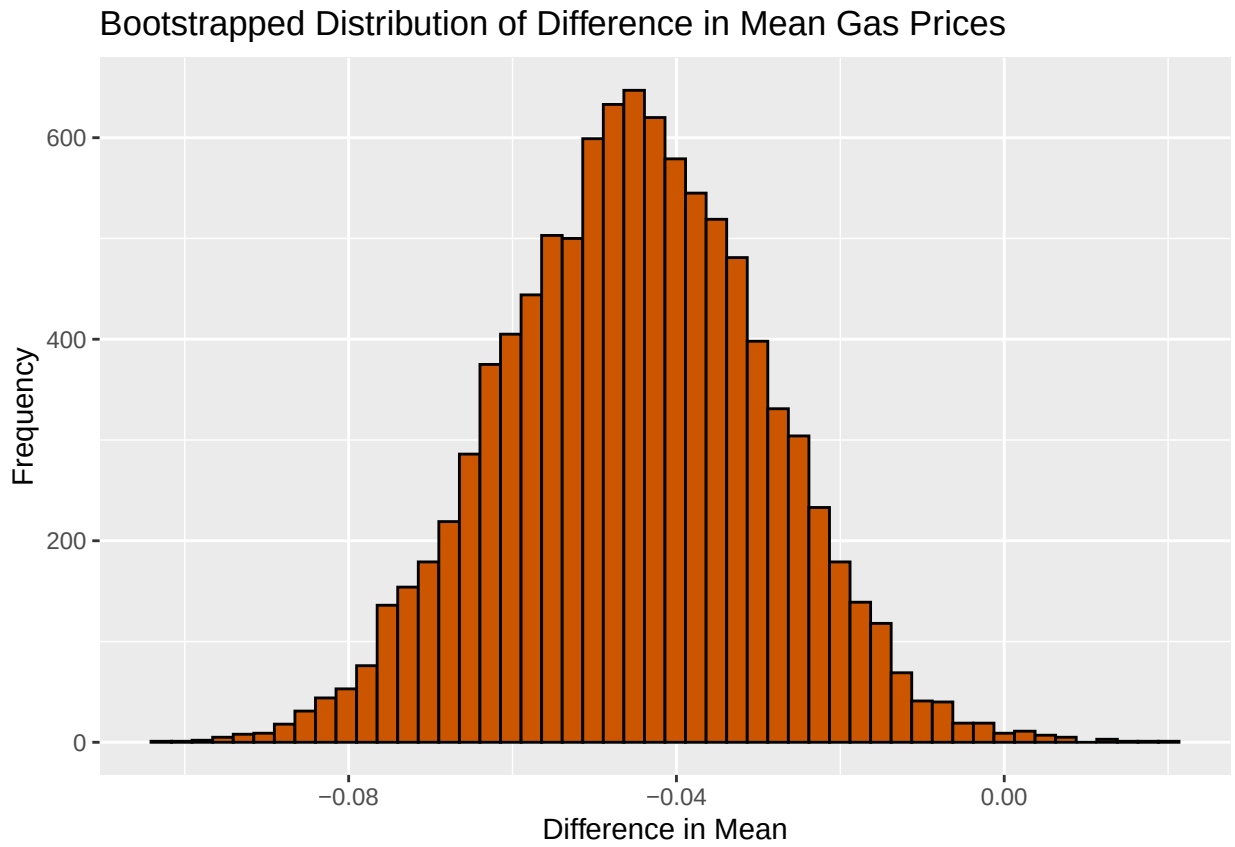
Theory B:

Claim:

The richer the area, the higher the gas prices.

Evidence:

```
##    High     Low
## 1.89175 1.84623
```




```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.07647577 -0.01438238  0.95 percentile -0.04552049
```

Conclusion

The confidence interval for the difference in mean gas prices between those in low versus high income groups (grouped by the mean of incomes) is (-7.65 cents, -1.44 cents) with 95% confidence. Since the interval is negative, we can reject the null hypothesis. Gas prices seem to correlate with income level, with higher prices in higher-income areas based on this analysis. However if grouped by median of incomes this might differ.

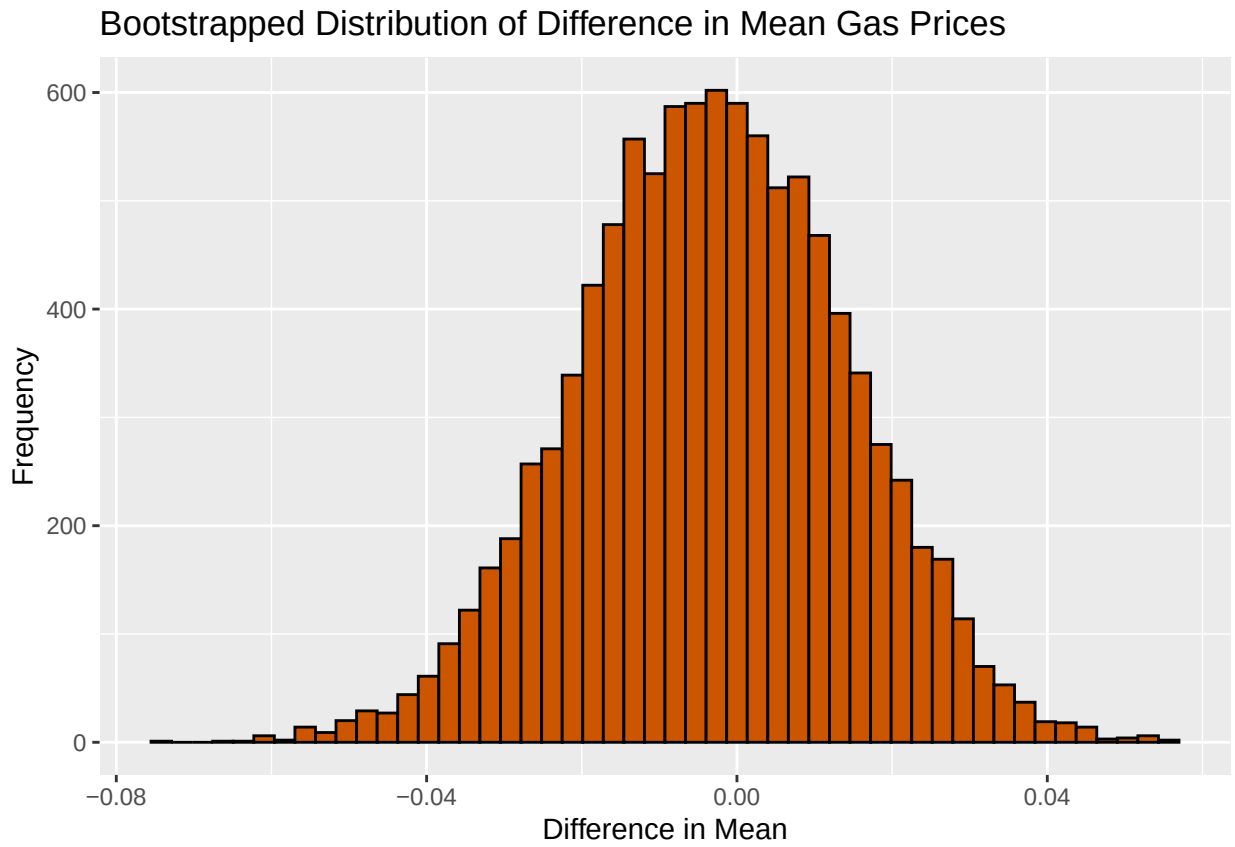
Theory C:

Claim:

Gas stations at stoplights charge more.

Evidence:

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




```
##       name       lower      upper level     method     estimate
## 1 diffmean -0.03735241 0.02979502  0.95 percentile -0.003299916
```

Conclusion:

The confidence interval for the difference in mean gas prices between those with and without a stoplight is (-3.74 cents, 2.98 cents) with 95% confidence. Since the interval includes zero, we cannot reject the null hypothesis. The presence of a stoplight does not appear to have a significant effect on the mean gas prices based on this analysis.

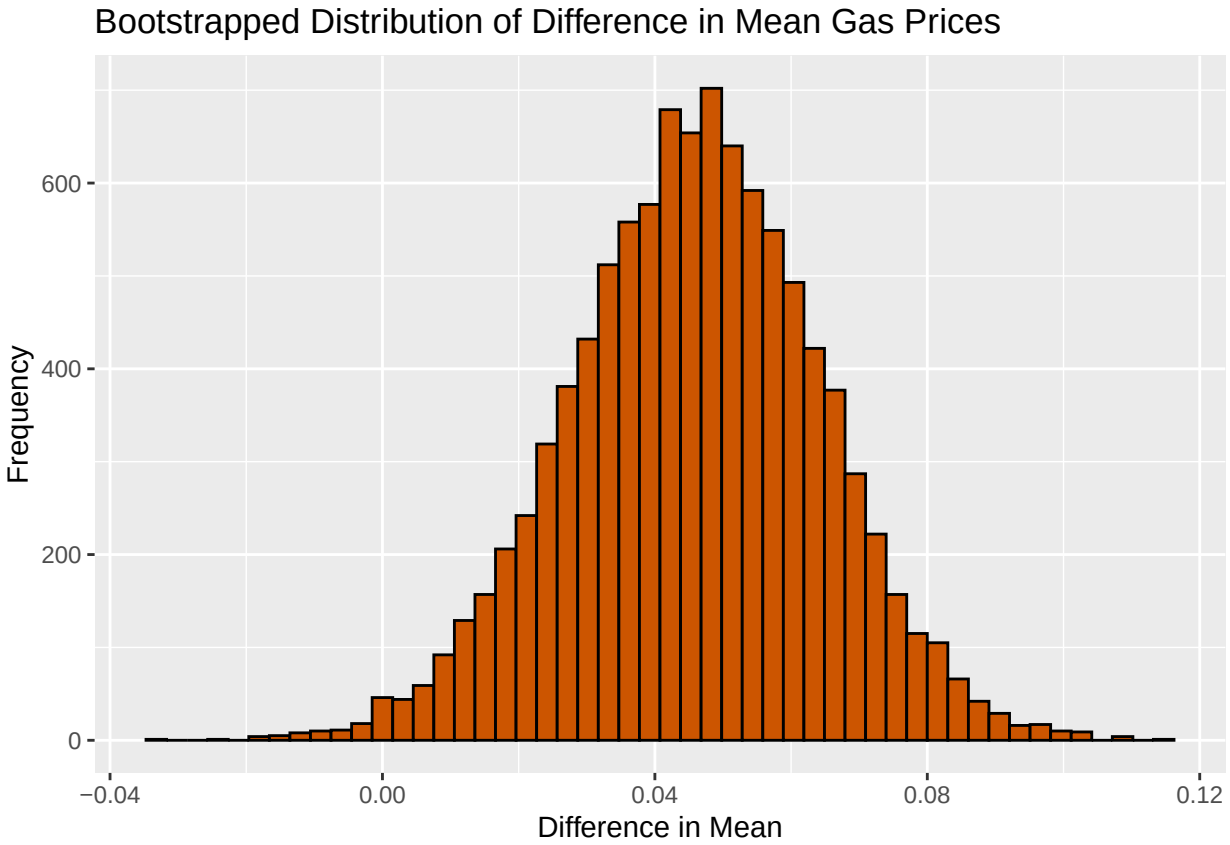
Theory D:

Claim:

Gas stations with direct highway access charge more.

Evidence:

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




```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.009004237 0.08125467  0.95 percentile 0.0456962
```

Conclusion:

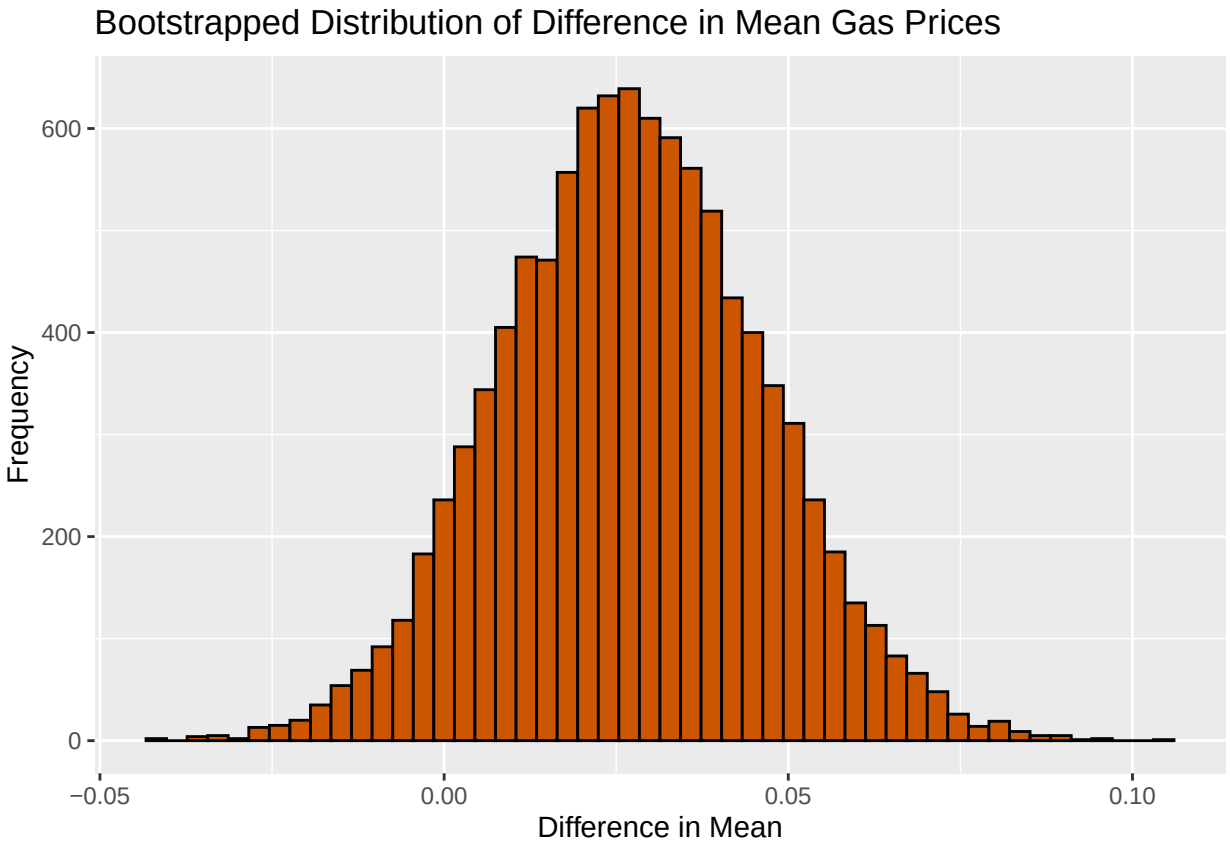
The confidence interval for the difference in mean gas prices between those with direct highway access versus those without is (.90 cents, 8.13 cents) with 95% confidence. Since the interval is positive (does not include zero), we can reject the null hypothesis. This indicates that, on average, gas stations with direct highway access have higher prices than those without.

Theory E:

Claim:

Shell charges more than all other non-Shell brands.

Evidence:




```
##       name        lower      upper level     method   estimate
## 1 diffmean -0.009320642 0.06526881  0.95 percentile 0.02740421
```

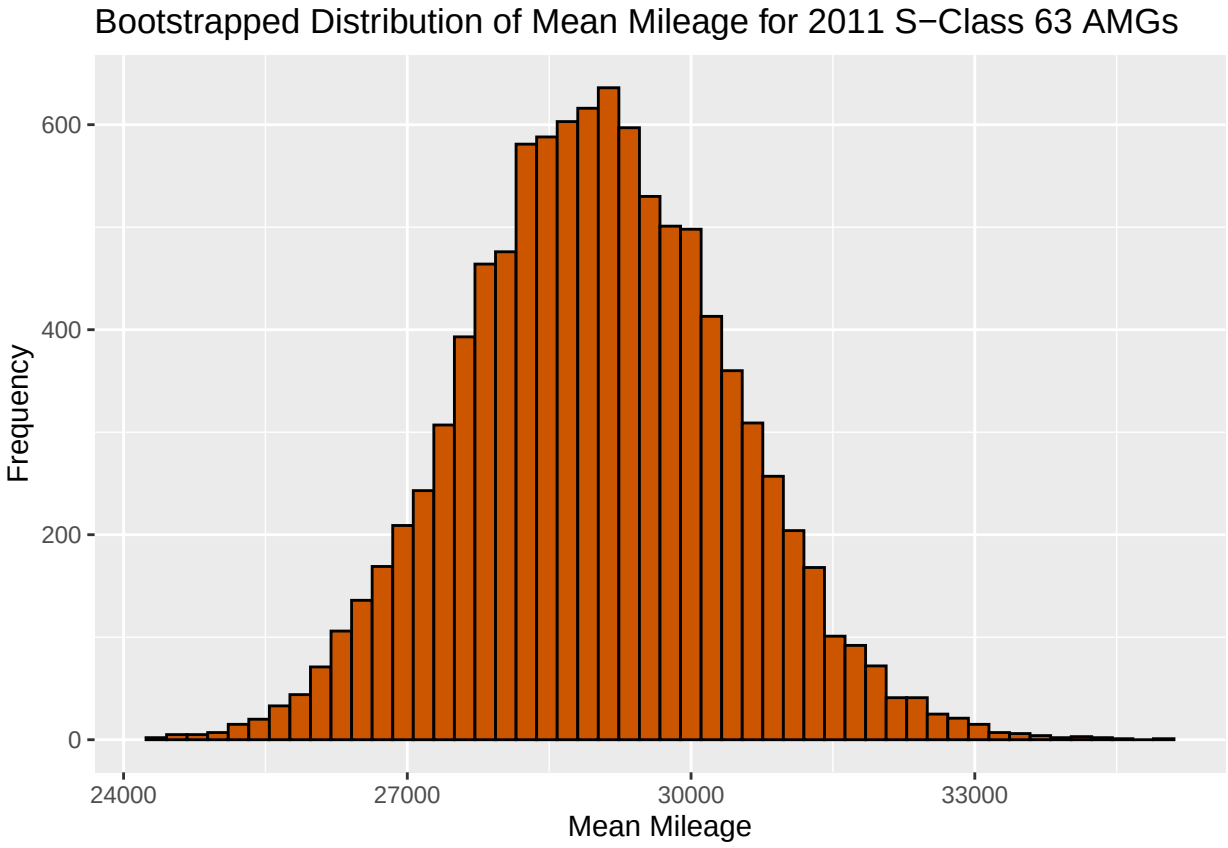
Conclusion:

The confidence interval for the difference in mean gas prices between Shell and any other brand is (-.93 cents, 6.53 cents) with 95% confidence. Since the interval includes zero, we cannot reject the null hypothesis. Whether or not the it is a Shell gas station does not appear to have a significant effect on the mean gas prices based on this analysis.

Problem 2:

Analysis of Used Mercedes S-Class Vehicles

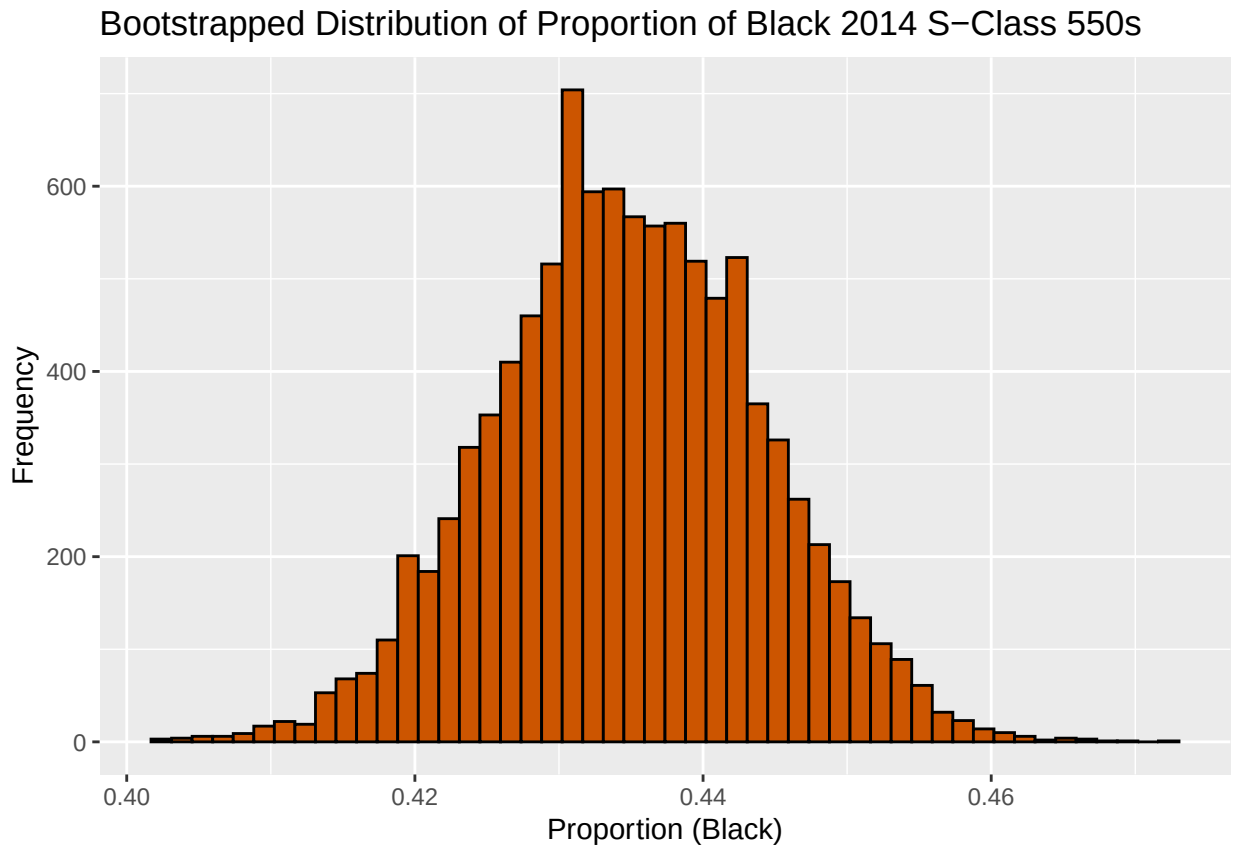
Part A: Average Mileage of 2011 S-Class 63 AMGs




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

Based on the bootstrapped interval, I can say with 95% confidence that used 2011 Mercedes 63 AMGs, on average, range from 26,299.28 miles to 31,825.47 miles.

Part B: Proportion of Black 2011 S-Class 550s




```
##        name     lower     upper level     method  estimate
## 1 prop_TRUE 0.4167532 0.4527518  0.95 percentile 0.4347525
```

Based on the bootstrapped interval, I can say with 95% confidence that the proportion of black used 2014 Mercedes 550s, on average, range from 41.68% to 45.28%.

Problem 3:

Analysis of NBC Pilot Survey

Part A:

Question:

Is there a significant difference in the mean viewer response about happiness between "Living with Ed" and "My Name is Earl"?

Approach:

Filter the data to only include the two shows, then use bootstrapping to get a confidence interval for the difference in mean happiness responses.

Results:

My evidence will be a confidence interval of 95%

```
##       name      lower     upper level     method   estimate
## 1 diffmean -0.3970078 0.1050464  0.95 percentile -0.1490515
```

Conclusion:

Based on the bootstrapped interval, I can say with 95% confidence that the difference in mean happiness scores between the shows is between -.40 and .11. Since the interval includes zero, it suggests that there is no significant difference in the happiness scores between the two shows.

Part B:

Question:

Did "The Biggest Loser" or "The Apprentice: Los Angeles" make people feel more annoyed?

Approach:

Filter the data to only include the two shows, then use bootstrapping to get a confidence interval for the difference in mean annoyed responses.

Results:

My evidence will be a confidence interval of 95%

```
##       name      lower       upper level     method  estimate
## 1 diffmean -0.5234503 -0.01856675  0.95 percentile -0.270997
```

Conclusion:

Based on the bootstrapped interval, I can say with 95% confidence that the difference in mean annoyance scores between the shows is between -.523 and -.019. Since the interval doesn't include zero, it suggests that there is a significant difference in the annoyance scores between the two shows. Since the interval is negative, on average, "The Biggest Loser" makes people less annoyed than "The Apprentice: Los Angeles"

Part C:

Question:

What proportion of American TV watchers would we expect to rate "Dancing with the Stars" as confusing?

Approach:

Filter the data to only include "Dancing with the Stars", then use bootstrapping to get a 95% confidence interval of the proportion of viewers who rated the show as confusing (4 or 5).

Results:

My evidence will be a confidence interval of 95%

```
##     name     lower     upper level     method  estimate
## 1 prop_1 0.3867403 0.5303867  0.95 percentile 0.4585635
```

Conclusion:

Based on the bootstrapped interval, I can say with 95% confidence that between 38.67% and 53.04% of American viewers will find the show confusing. Since the interval doesn't include zero, it suggests that there is a significant portion of the population that will be confused, despite its simple format. This implies that the show may be clear to all viewers as assumed.

Problem 4:

Question:

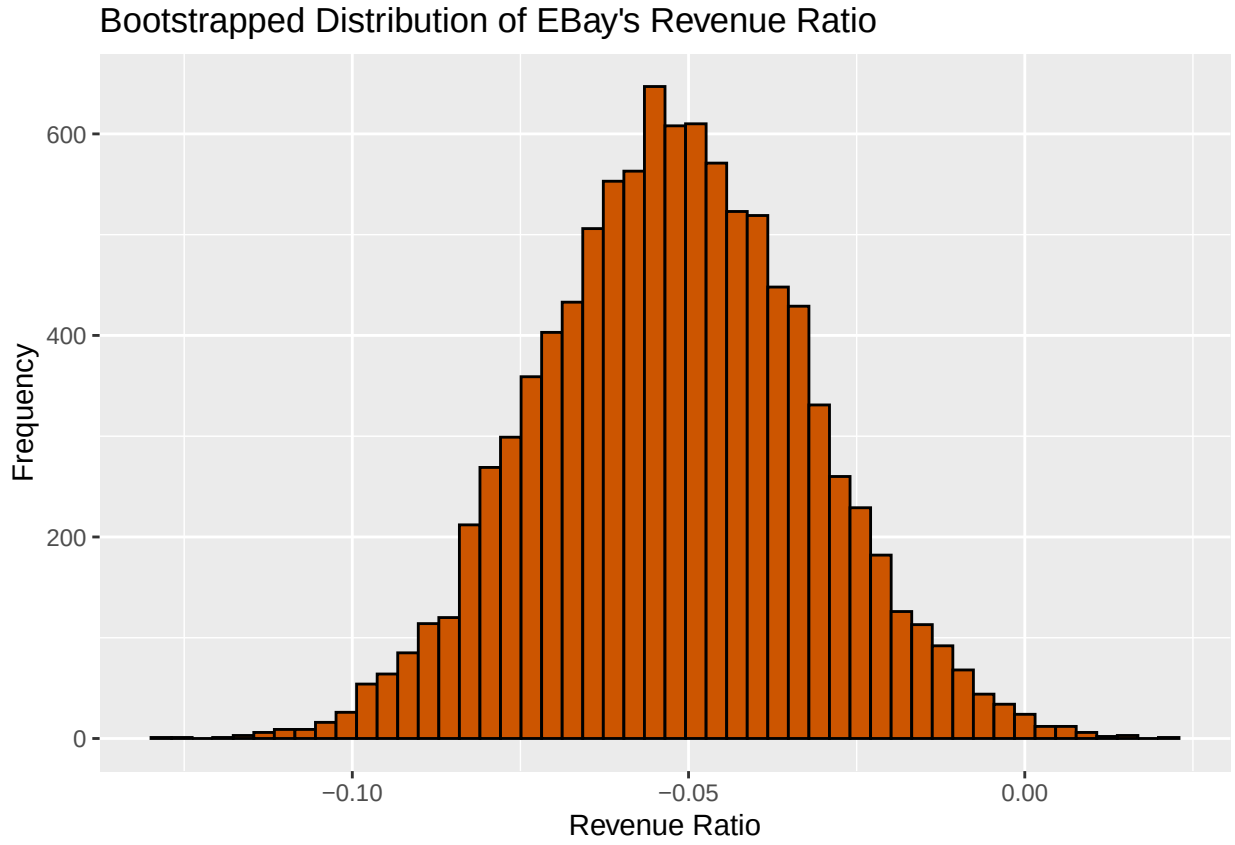
Is the revenue ratio (the ratio of revenue after to before) statistically significantly different between the treatment group (where ads were paused) and the control group (where ads continued) in EBay's experiment? Does EBay's paid search advertising on Google generate additional revenue?

Approach:

Mutate the dataset to get the revenue ratio for each DMA (rev_after/rev_before). Group by whether treatment is true or false (1 or 0). Use bootstrapping to calculate a 95% confidence interval between the two groups.

Results:

My evidence will include a graph and a confidence interval of 95%.




```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.09131885 -0.01251123  0.95 percentile -0.05228145
```

Based on the confidence interval, I can say with 95% confidence that the difference in revenue ratios between DMAs where ads were pause versus where ads were running is between -.0913 and -.0125. Since the interval is entirely negative (doesn't contain zero), pausing paid search ads on Google led to a statistically significant small decrease in revenue. This provides evidence that paid search advertising on Google does contribute additional revenue to EBay.