Hw_3

Pranita

2025-02-13

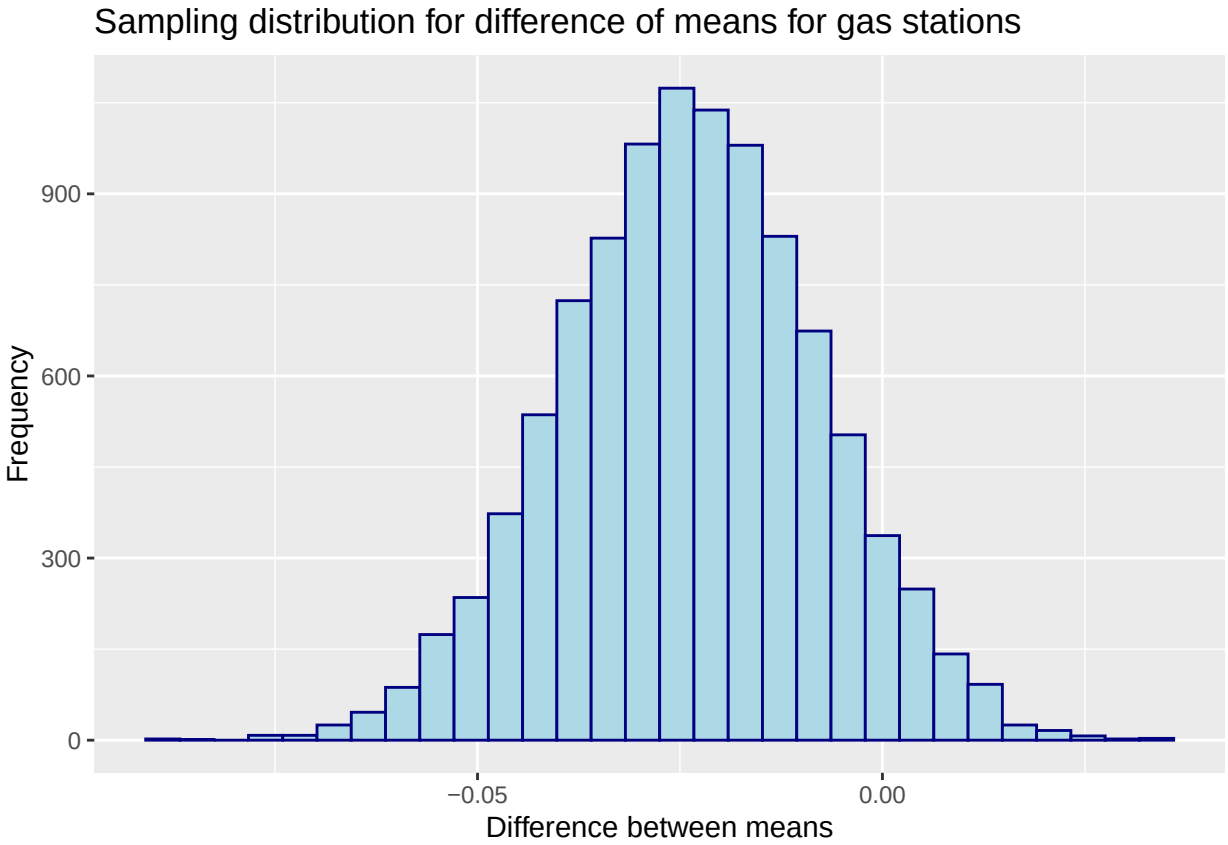
Name: Pranita Chaudhury

UT EID: pc28377

Github: https://github.com/PranitaChau/Hw_3.git

Problem 1

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

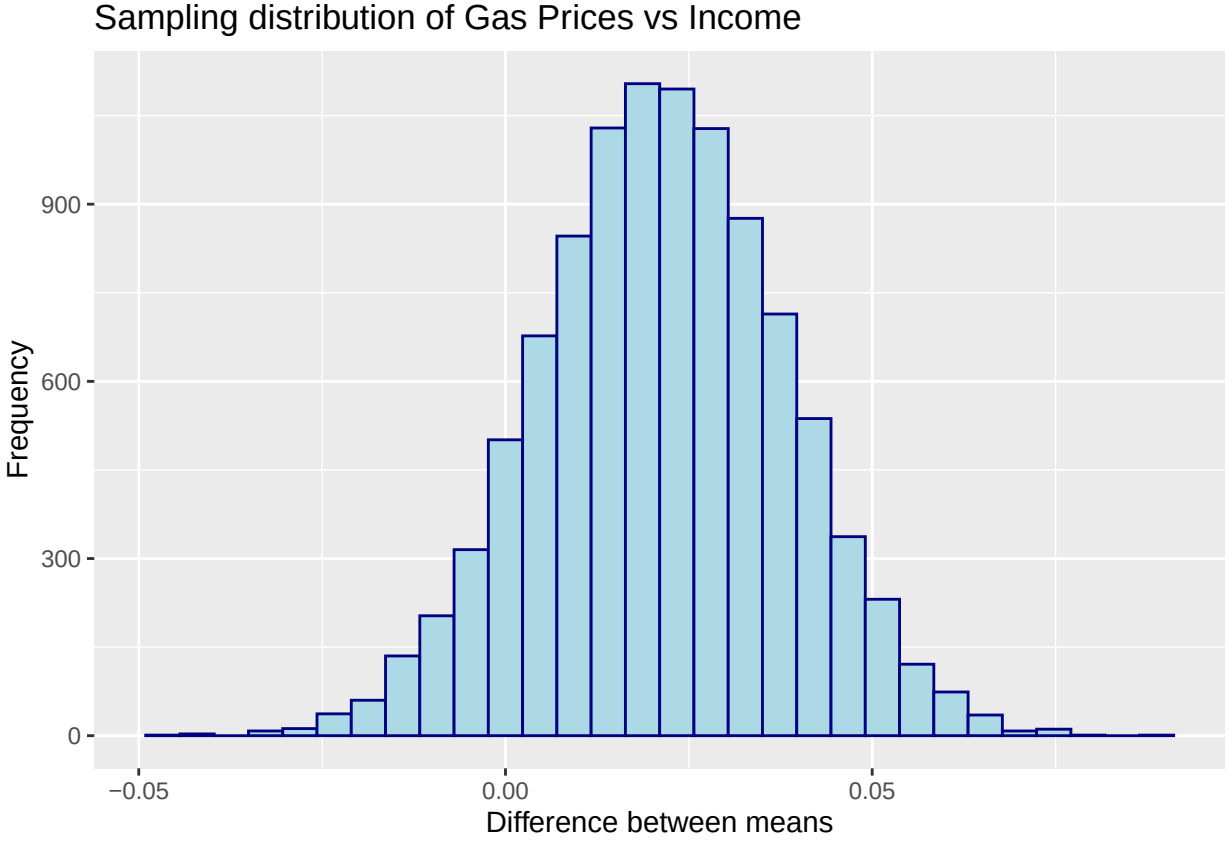



```
##       name       lower    upper level     method    estimate
## 1 diffmean -0.05504481 0.007348  0.95 percentile -0.02348235
```

The claim is gas stations that lack direct competition in sight charge higher prices for gas compared to those that are located with competitors in sight. We can collect evidence against this claim by comparing the mean price for gas stations with and without competitors in sight. The difference in prices between gas stations with and without direct compitition in sight is somewhere between -0.054 to 0.008, with 95% confidence.

However, since 0 is included in this interval, it should be noted that there is no statistical significance in prices for gas stations that do and do not have competitors nearby. Therefore the claim that gas stations with competitors in sight cannot be supported with our data using a 95% confidence interval.

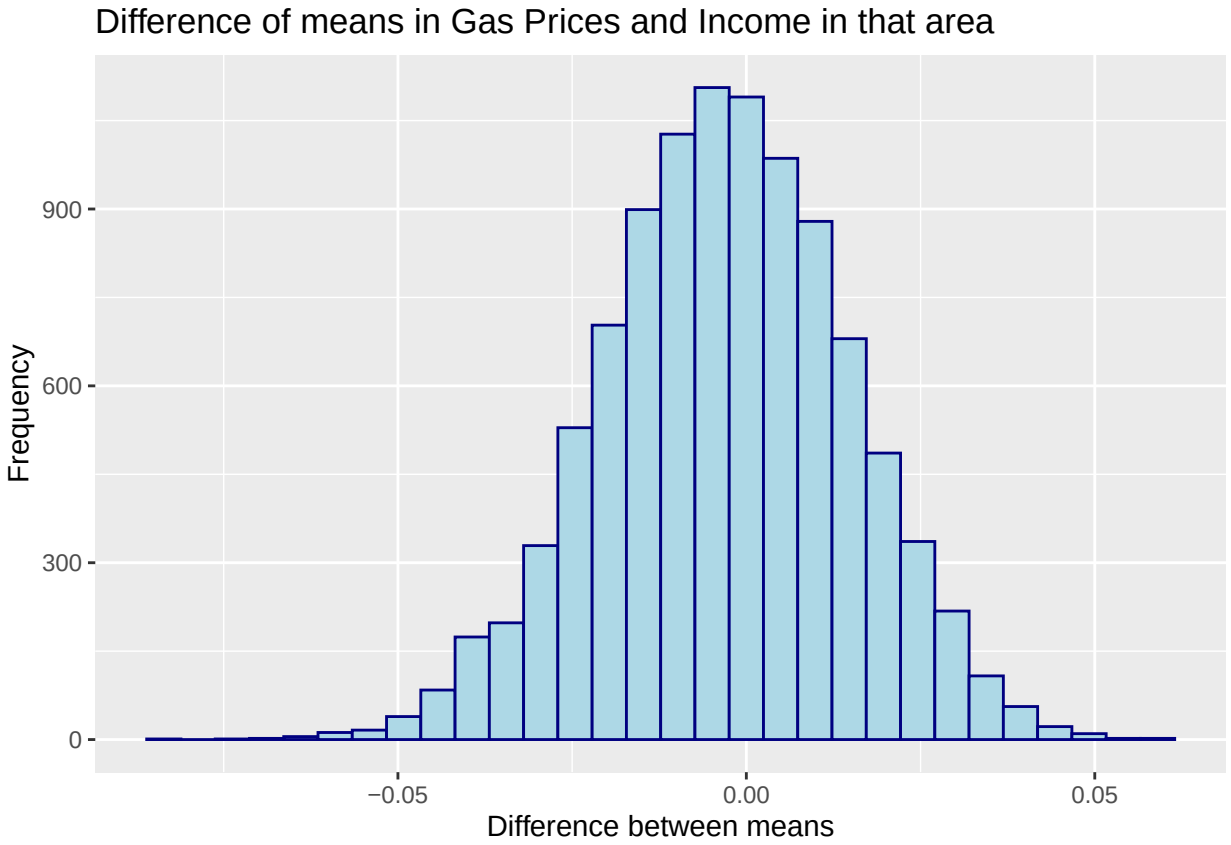
Theory B - The richer the area, the higher the gas price




```
##       name       lower      upper level     method   estimate
## 1 diffmean -0.01187101 0.05375173  0.95 percentile 0.02132411
```

The claim is that the more rich an area is, the higher the gas prices will be. Evidence can be collected for this by comparing the income, defined as the Median Household Income in the area, to the price of gas in said area. The evidence against this claim can be found through creating a 95% confidence interval and looking at the interval. In order to do this I created the rich variable, and used the median Income found in the dataset ($52,306) defined as the variable mid, and bootstrapped that against the Price variable. The resulting confidence interval was -0.01138222 to 0.05392115, and since it contained zero it means there is no statistical significance to support the claim. Therefore, using a 95% confidence interval there is no evidence to support the claim that higher income areas have more expensive gas based on this dataset while using the median Income of this dataset to differentiate between rich and not rich.

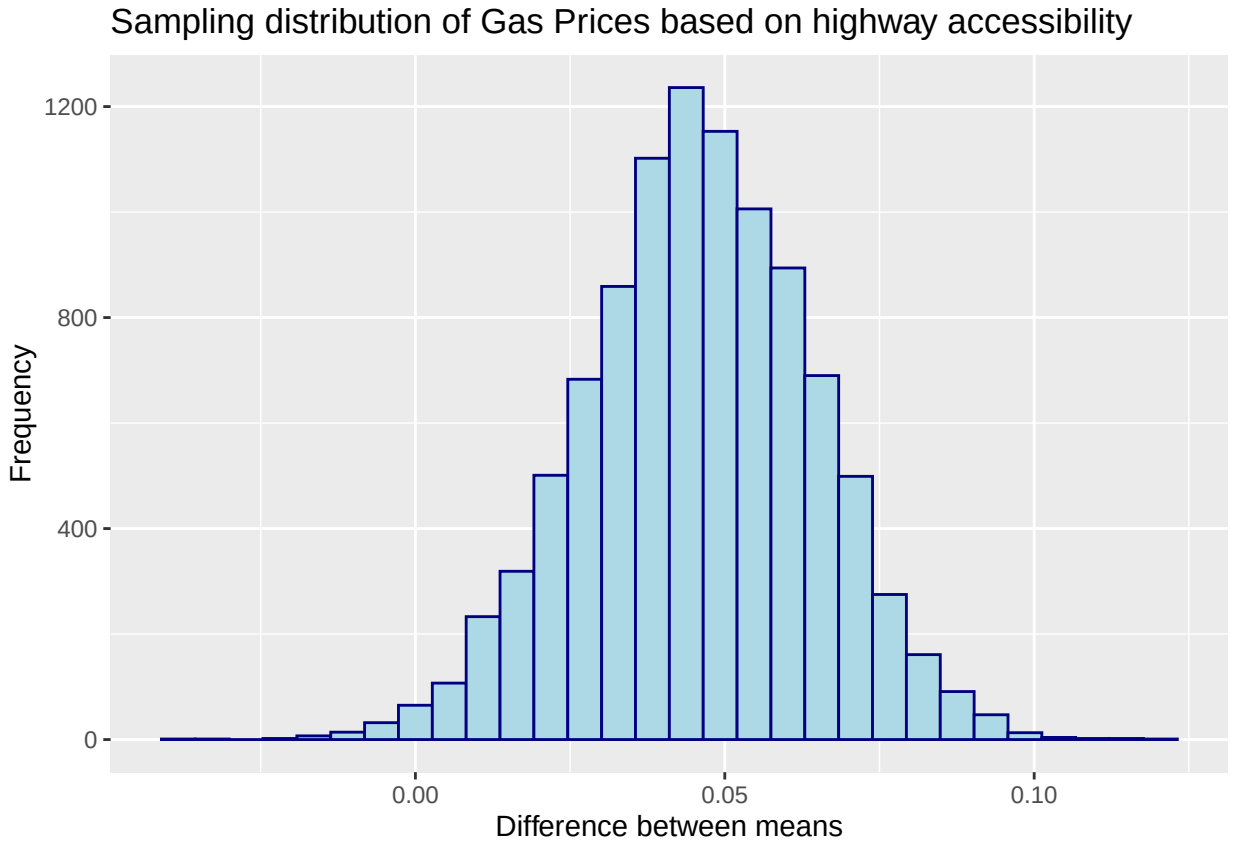
Theory C - Gas stations at stoplights charge more




```
##       name       lower      upper level     method     estimate
## 1 diffmean -0.03838553 0.03063937  0.95 percentile -0.003299916
```

The claim is that gas stations at stoplights tend to charge more for gas. Evidence that can be used to dispute this claim can come from constructing a 95% confidence interval comparing prices for a gas station at a stoplight to gas stations that are not. The result of that procedure shows that there is between a -0.03804523 to 0.02977349 price difference in means between gas stations that are and are not next to a stoplight. Therefore since zero is included in the confidence interval, we can conclude with 95% confidence that the initial claim is not supported.

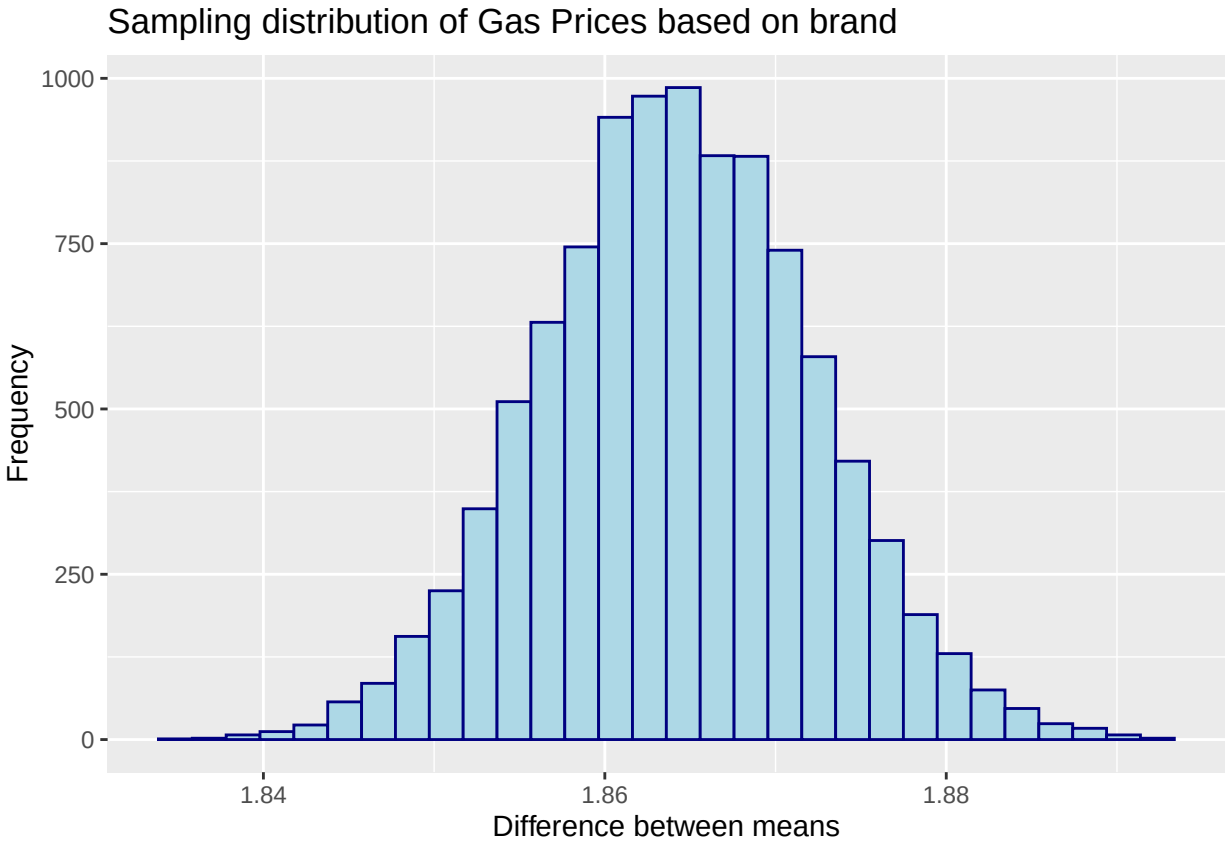
Theory D - Gas stations with direct highway access charge more




```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.008923484 0.08153093  0.95 percentile 0.0456962
```

The claim is that gas stations with direct highway access tend to charge more than gas stations further away from the highway. Evidence can be gathered to support this claim by constructing a 95% confidence interval to compare gas prices for gas stations with and without direct highway access. The result of that procedure shows that there is between a 0.009 and 0.081 price difference in means between gas stations with and without direct highway access. Therefore, we can conclude with a 95% confidence interval that gas stations with direct highway access do tend to charge somewhere between $0.009 and $0.081 more than gas stations without direct highway access.

Theory E - Shell charges mroe than all other non-Shell brands

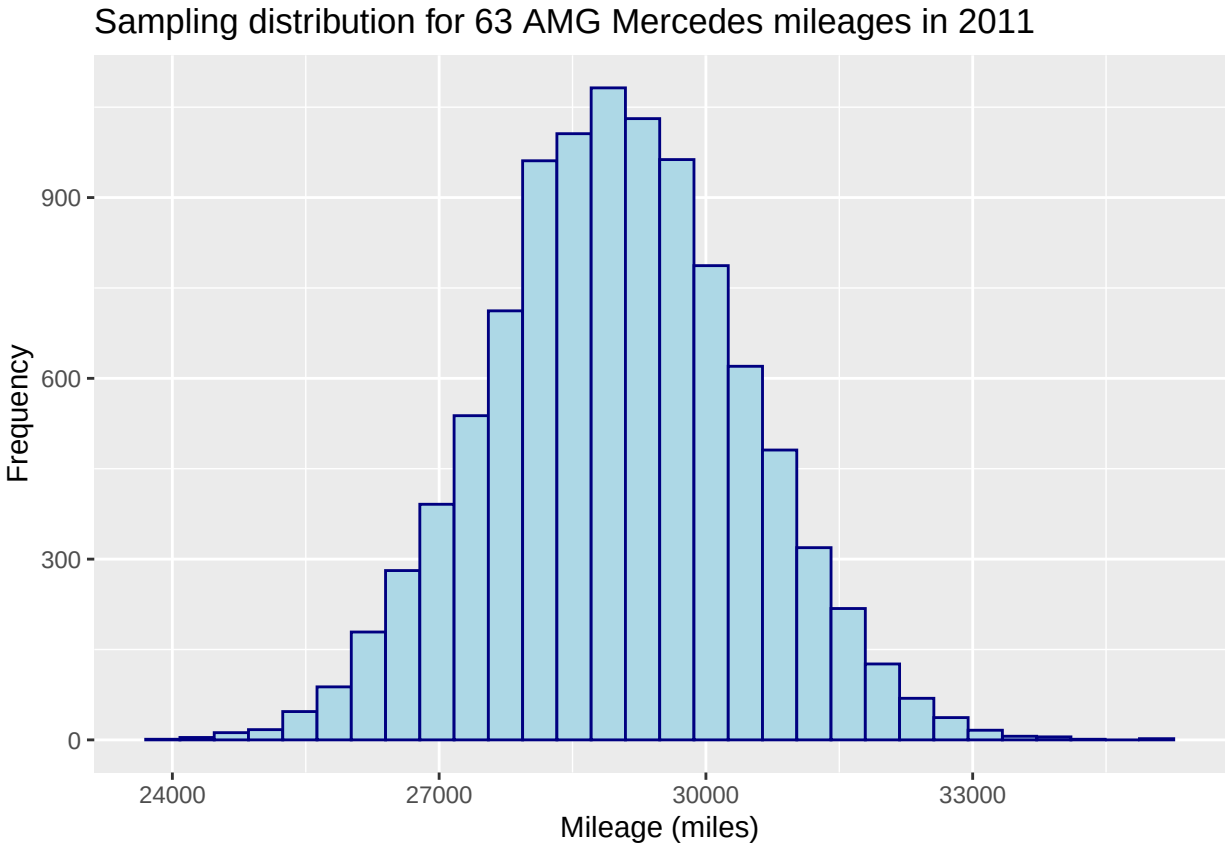



```
##   name    lower    upper level     method estimate
## 1 mean 1.848614 1.880099  0.95 percentile 1.864257
```

The claim is the brand Shell charges more than other brands of gas. Evidence for this claim can be shown by comparing the mean prices of the price of gas that Shell charges to all other gas brands. Using a 95% confidence interval it can be determined that there is a 1.85 to 1.88 difference in mean prices. Therefore, we can conclude that on average the difference in prices between gas from Shell compared to other gas brands is somewhere between $1.85 and $1.88 with 95% confidence, and the initial claim is supported.

Problem 2

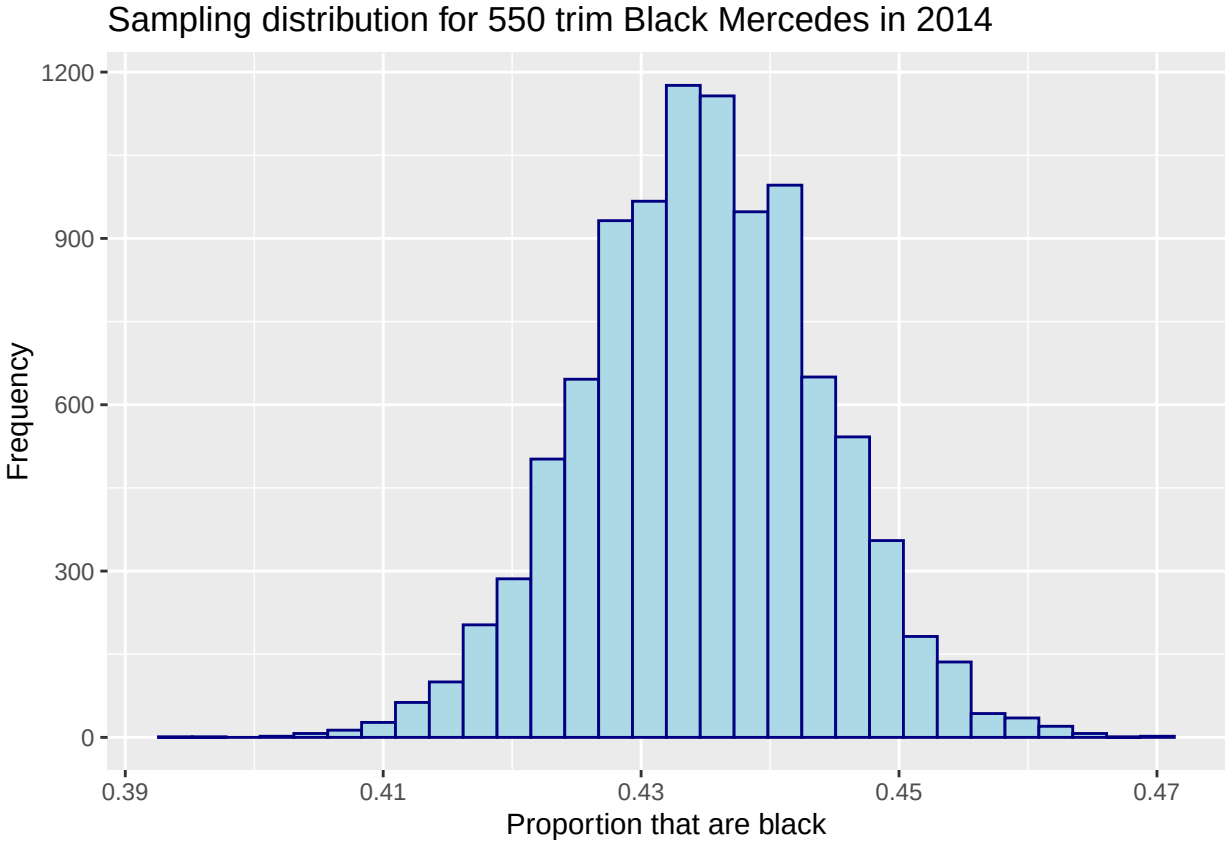
Part A




```
##   name    lower    upper level     method estimate
## 1 mean 26210.52 31826.24  0.95 percentile 28997.34
```

We are looking at used Mercedes S-Class vehicles sold on cars.com with a trim of 63 AMG in the year 2011. The graph shows the distribution of their mileage. Based on this data we can say with 95% confidence that the average mileage of cars that fit in this category is somewhere between 26293.74 and 31777.48 miles.

Part B




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

We are looking at used Mercedes S-Class vehicles sold on cars.com with a trim of 550 in the year 2014. The graph shows the distribution of the proportion of cars colored black. Based on this data we can say with 95% confidence that the average proportion of black cars that fit in this category is somewhere between 0.4164 and 0.4531.

Problem 3

Part A

```
##       name      lower    upper level     method    estimate
## 1 diffmean -0.4483169 0.299056  0.95 percentile -0.08119384
```

Question: I am answering the question of which show makes people happier: Living with Ed or My Name is Earl. Approach: I approached this question by bootstrapping and creating a 95% confidence interval in order to compare the difference of their means. I filtered the shows by name and created a new dataset with only those two shows and factored it by show. I then bootstrapped the difference of means of the different shows by how happy viewers said it made them, and created a 95% confidence interval. Results: My approach provided me with a table containing the 95% confidence interval. We can say with 95% confidence that there difference in happiness for the average viewer ranges from -0.4572 to 0.3033. However, since our confidence interval includes zero we cannot say that one show makes people more happy than the other. Conclusion: Therefore, we can conclude that based on our results there is no statistical significance that Living with Ed or My Name is Earl makes the average population of viewers more happy than the other show. For stakeholders such as the producers of the shows, since there is no significant difference in viewer happiness between these

two shows they may choose to air these shows on TV for the same amount or at similar times since they generate a similar amount of happiness for the viewers.

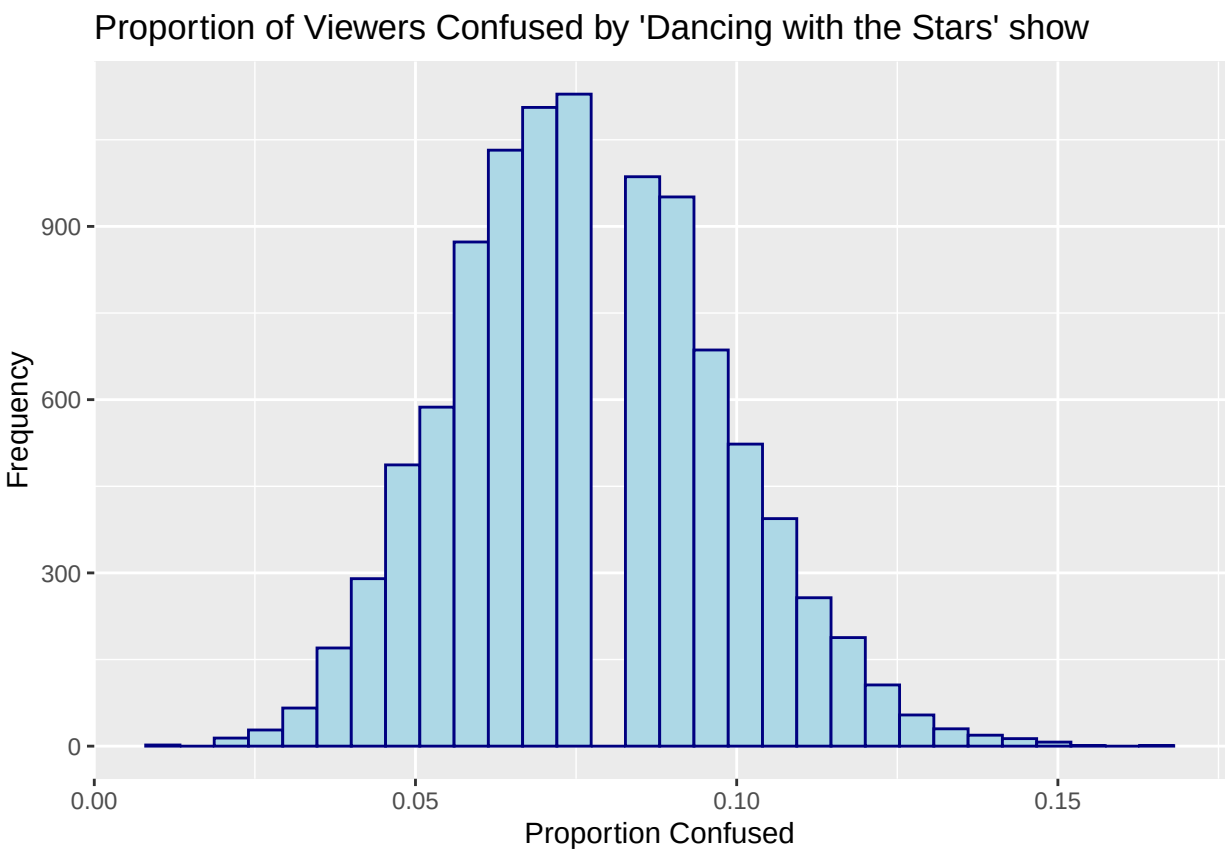
Part B

```
##       name      lower       upper level     method  estimate
## 1 diffmean -0.5294143 -0.02004238  0.95 percentile -0.270997
```

Question: I am answering the question of which show makes people annoyed: The Biggest Loser or The Apprentice: Los Angeles. Appraoch: I approached this question by bootstrapping and creating a 95% confidence interval in order to compare the difference of their means. I filtered the shows by name and created a new dataset with only those two shows and factored it by show. I then bootstrapped the difference of means of the different shows by how happy viewers said it made them, and created a 95% confidence interval. Results: My approach provided me with a table containing the 95% confidence interval of -0.5217547 to -0.02603958. Conclusion: Therefore we can say with 95% confidence that the difference in annoyance between the average viewer of The Biggest Loser and The Apprentice: Los Angeles lies within -0.5217547 to -0.02603958, and that people are less annoyed by the show The Biggest Loser. For stakeholders such as the producers of the shows, since The Biggest Losers causes less annoyance among viewers, so they may want to air that on TV more than The Apprentice: Los Angeles.

Part C

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



Quesiton: I am answering the question of what proportion of American TV watchers would we expect to give a response of 4 or greater to the "Q2_Confusing" question, as in what proportion of viewers are confused by the premise of the TV show "Dancing with the stars".

Approach: In order to answer this question I used the bootstrapping method. I made a data set containing the viewer's answer to the question Q2_Confusing, which indicated how confused the show made them, and set answers of 4 or 5 to "True" and anything lower than that to "False". I then simulated 10,000 samples of that through bootstrapping in order to find the the proportion of viewers that responded with a 4 or 5 (indicating they were confused).

Results: My approach provided me with a table containing the interval 0.03867403 to 0.1160221 with 95% confidence. Additionally, there is a graph which shows the sampling distribution of the average proportion of American TV watchers confused by the premise of the show "Dancing with the stars".

Conclusion: Therefore, ee can say with 95% confidence that the average proportion of viewers confused by the show "Dancing with the stars" is somewhere from 0.0387 to 0.1160. For stakeholders such as the producers of the shows, they may want to invest in ways to make the show more self explanatory in order to have new viewers tune in. Since current viewers are confused it may be more difficult for future viewers to tune in.

Problem 4

```
##       name      lower       upper level     method    estimate
## 1 diffmean -0.0906901 -0.01331123  0.95 percentile -0.05228145
```

Question: During May of 2013 Ebay ran an experiment regarding their revenue brought in by ads, and turned off their Google ads in certain locations (DMA). They compared the revenue ratios between DMAs that still had Ebay ads running (control group) and DMAs where Ebay turned their ads off (treatment group). The question I am trying to answer is whether the revenue ratio for Ebay is the same between their treatment and control groups.

Approach: My approach in order to answer this question was to use the bootstrapping method, Firstly I created a revenue_ratio variable in order to compare the revenue ratios between the two groups and not just raw revenue numbers. Next I made two different data sets, one for the control group and the other for the treatment group. Then I created 100,000 simulations using the bootstrap method in order to compare the means of the revenue ratios of two groups.

Results: Finally I created a 95% confidence interval using the result of the bootstrapping, and I got a confidence interval of -0.09086329 to -0.01336305. This result is displayed above in a table.

Conclusion: Using these results, it can be concluded with 95% confidence that DMAs where Ebay turned their Google ads off had revenue ratios that were on average -$0.09086329 to -$0.01336305 lesser than DMAs where Ebay did not turn their Google ads off. This means that Ebay earns more profit when their ads on Google are turned on. Stakeholders, referring to Ebay higher ups or any large business that uses the same type of ads on Google, would be interested in knowing these results as it shows that ads on Google do infact positively affect revenue of the company.