Hw_3

Pranita

2025-02-12

Name: Pranita Chaudhury

UT EID: pc28377

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

to do later- label graph axis, fix 1b, write up all conclusions for 1 AND for conclusion add the stakeholder pov

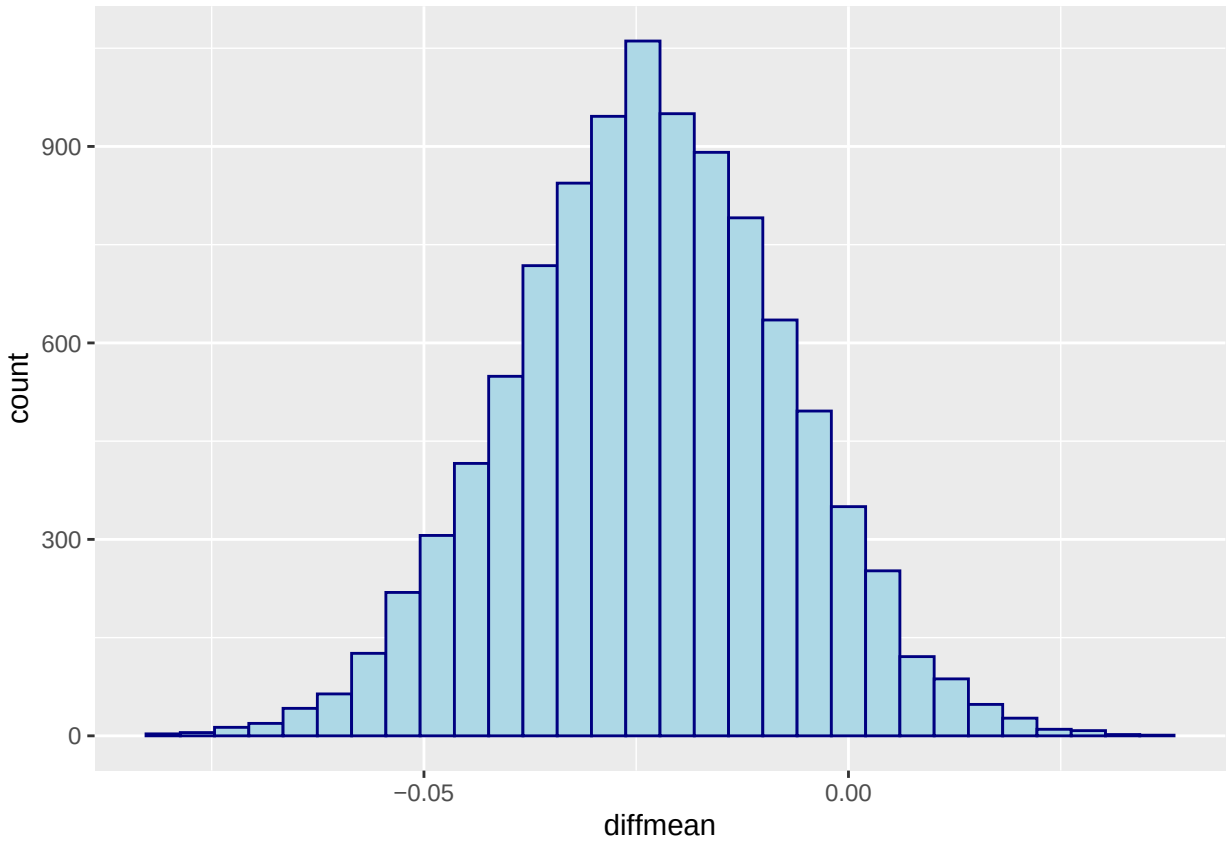
in #4 does 10,000 monte sims just mean do(10,000)

can i answer the questions in paragrpah format or do i need the 1) 2) format

Problem 1

Theory A

The first theory is that gas stations charge more if they lack direct competition in sight.

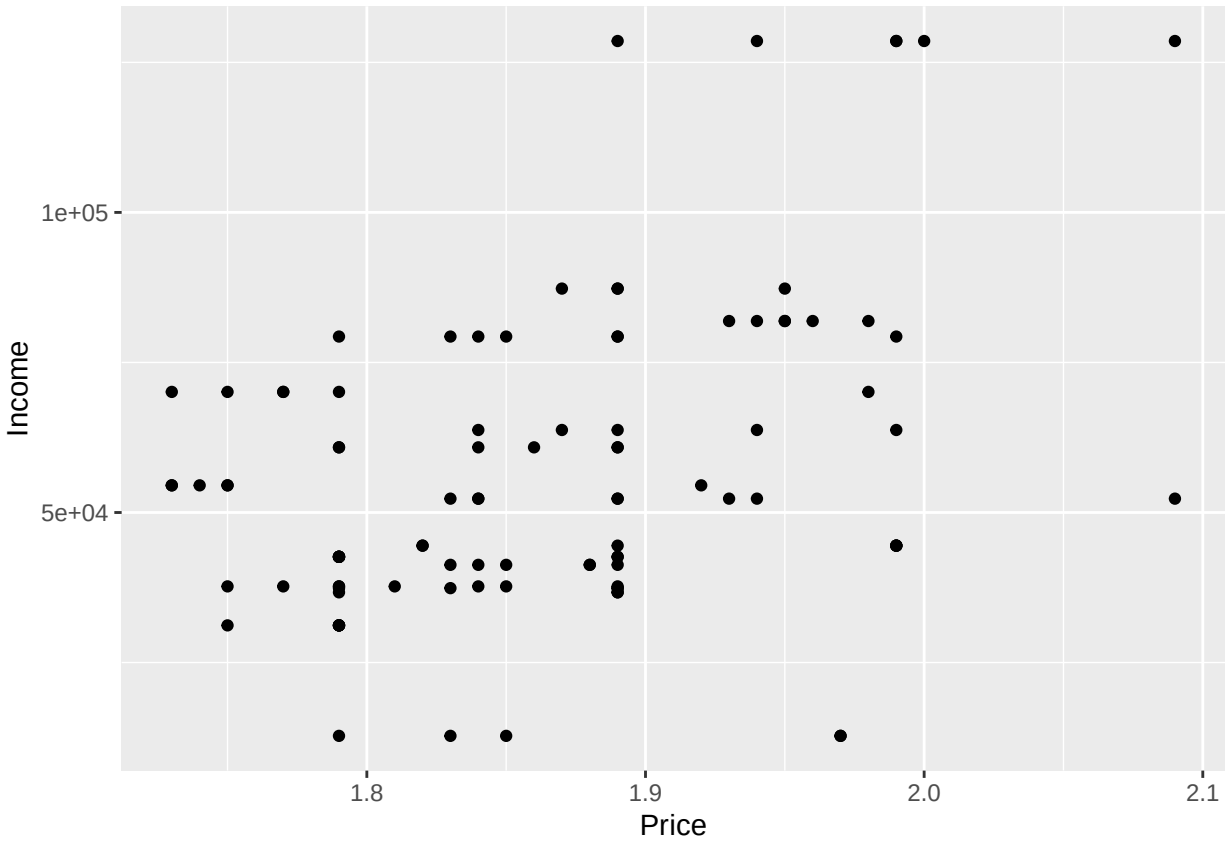



```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05507495 0.007751219  0.95 percentile -0.02348235
```

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 for this claim by comparing the mean price for gas stations with and without competitors in sight. Using a 95% confidence interval, there is a -0.054 to 0.008 difference in prices. Since 0 is included in this interval, it should be noted that there is no significant difference 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




```
##  (Intercept)       Income
## 1.793442e+00 1.248341e-06
```

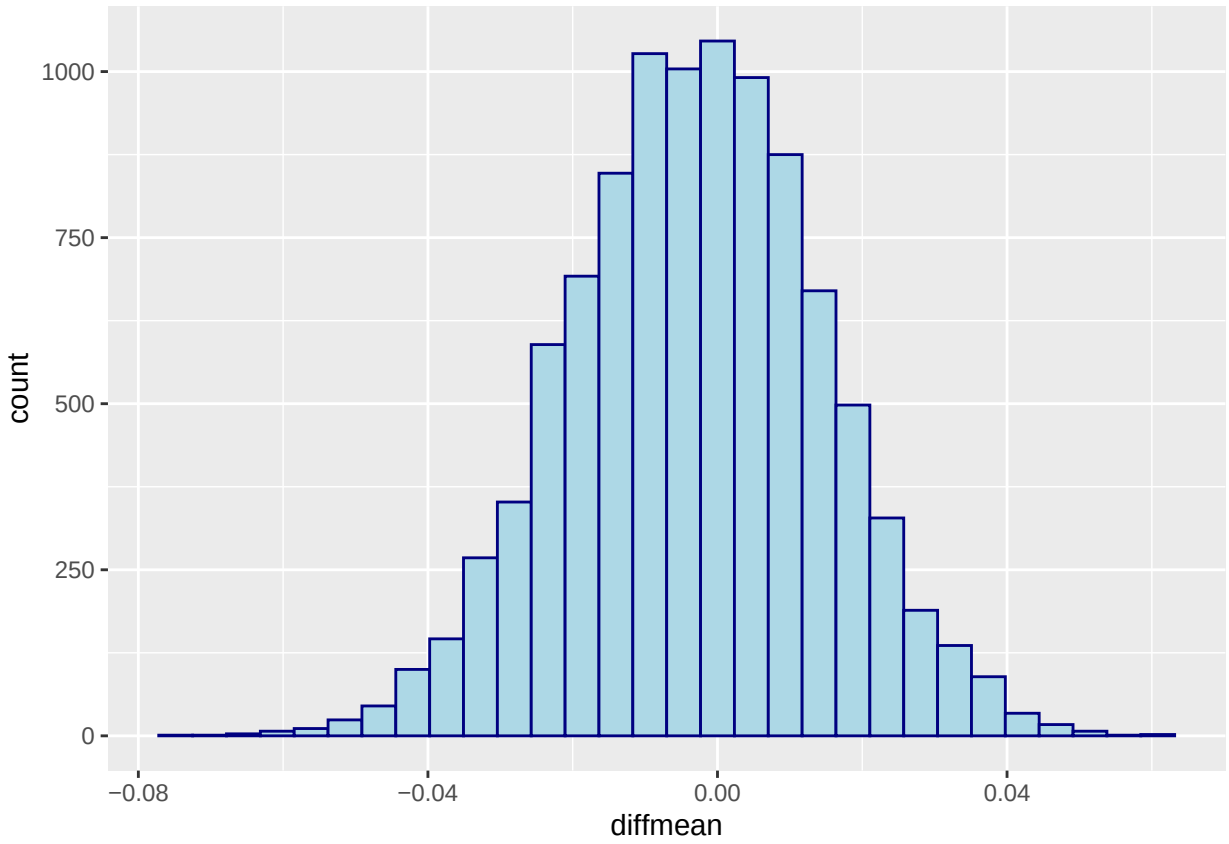
```
##        name        lower        upper level     method     estimate
## 1 Intercept 1.756008e+00 1.827530e+00  0.95 percentile 1.819702e+00
## 2    Income 6.676096e-07 1.783731e-06  0.95 percentile 8.696826e-07
## 3     sigma 6.494887e-02 8.262315e-02  0.95 percentile 7.257700e-02
## 4 r.squared 3.433521e-02 3.323077e-01  0.95 percentile 8.617654e-02
## 5         F 3.525927e+00 4.927206e+01  0.95 percentile 9.336023e+00
```

```
## [1] 0.08151499
```

The claim is that the higher the median income of an area, the price of gas will be more expensive. Evidence can be collected for this by comparing the median income of an area to the price of gas in said area. By plotting the graph there is a slight positive correlation visible, but it is not obvious so we can try to fit a linear model comparing the gas prices and income of the area. The coefficient for income shows a very small positive correlation, suggesting there is a weak correlation between gas prices of 1.248e-6.

Theory C

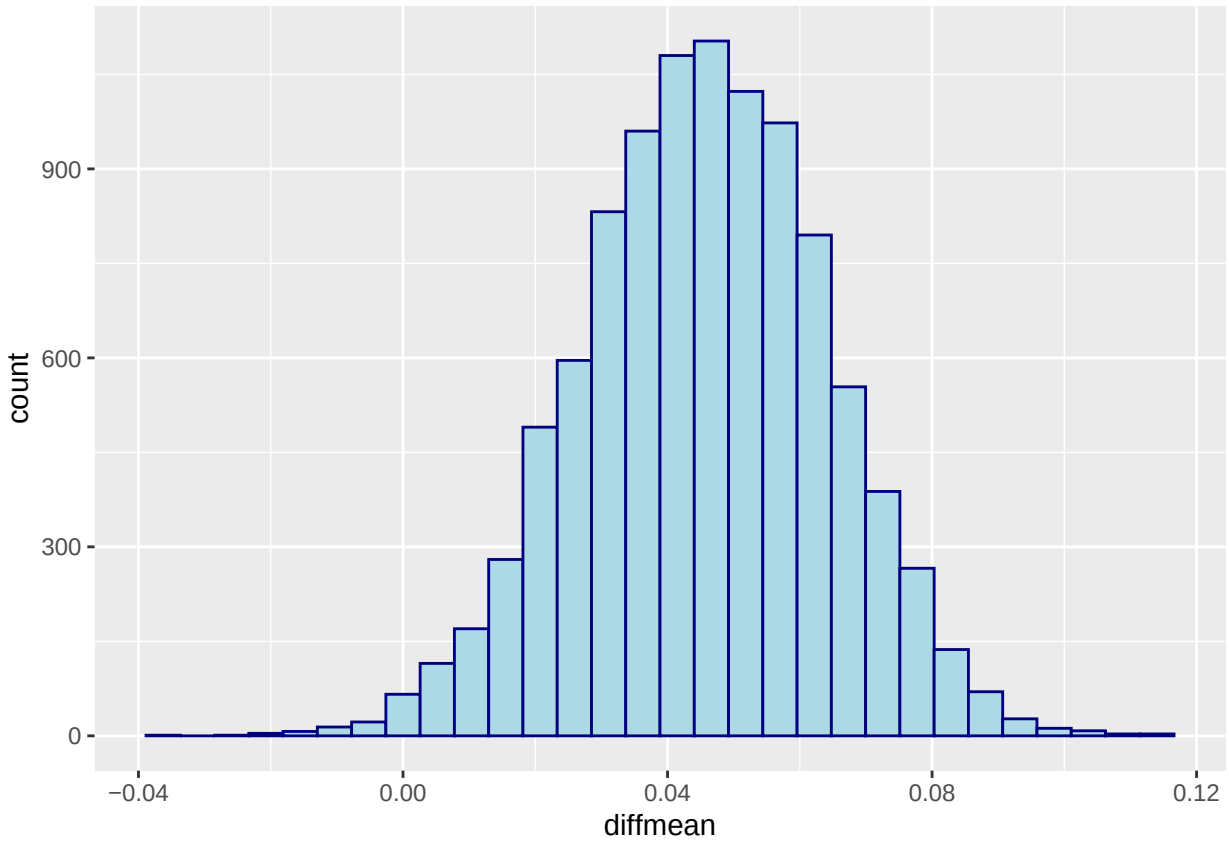
Gas stations at stoplights charge more




```
##       name       lower      upper level     method     estimate
## 1 diffmean -0.03753939 0.03147818  0.95 percentile -0.003299916
```

Theory D

Gas stations with direct highway access charge more

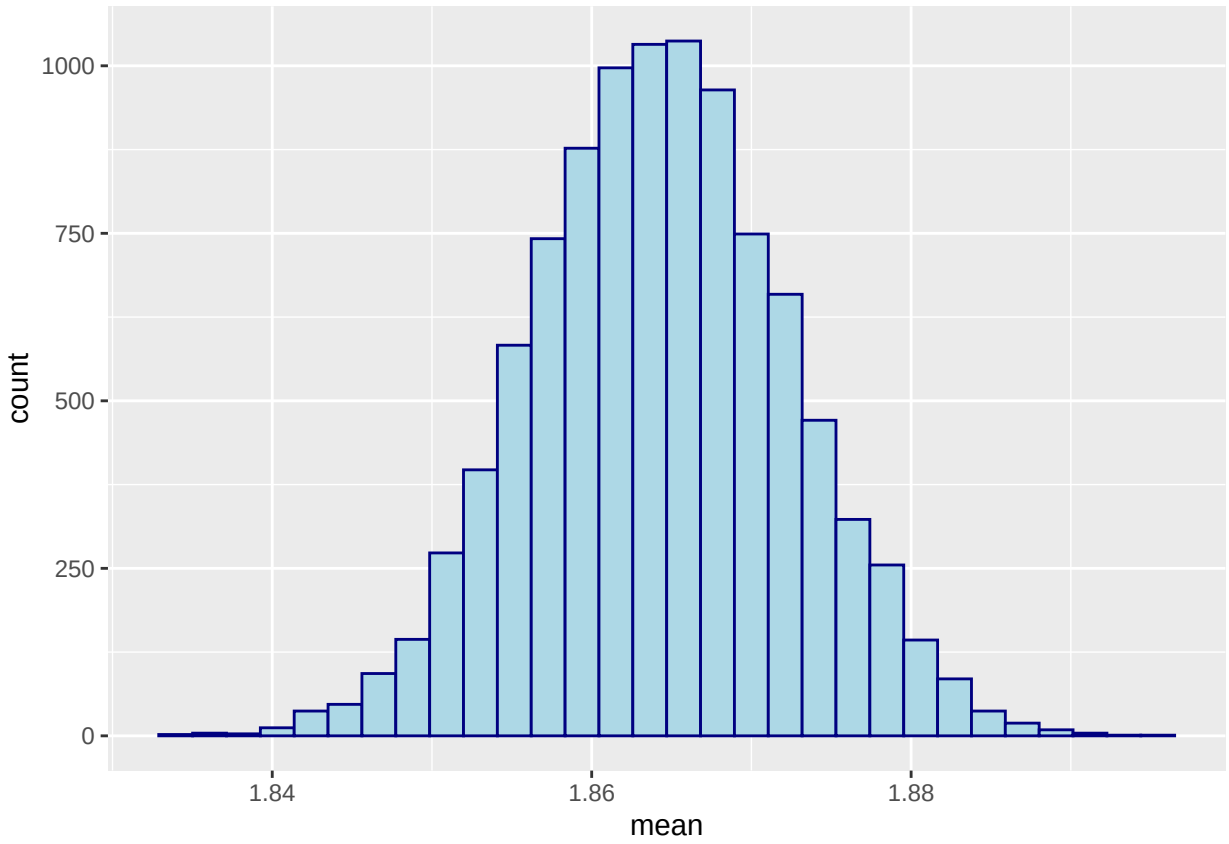



```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.008741557 0.08061048  0.95 percentile 0.0456962
```

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 for this by using 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 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 somwhere 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

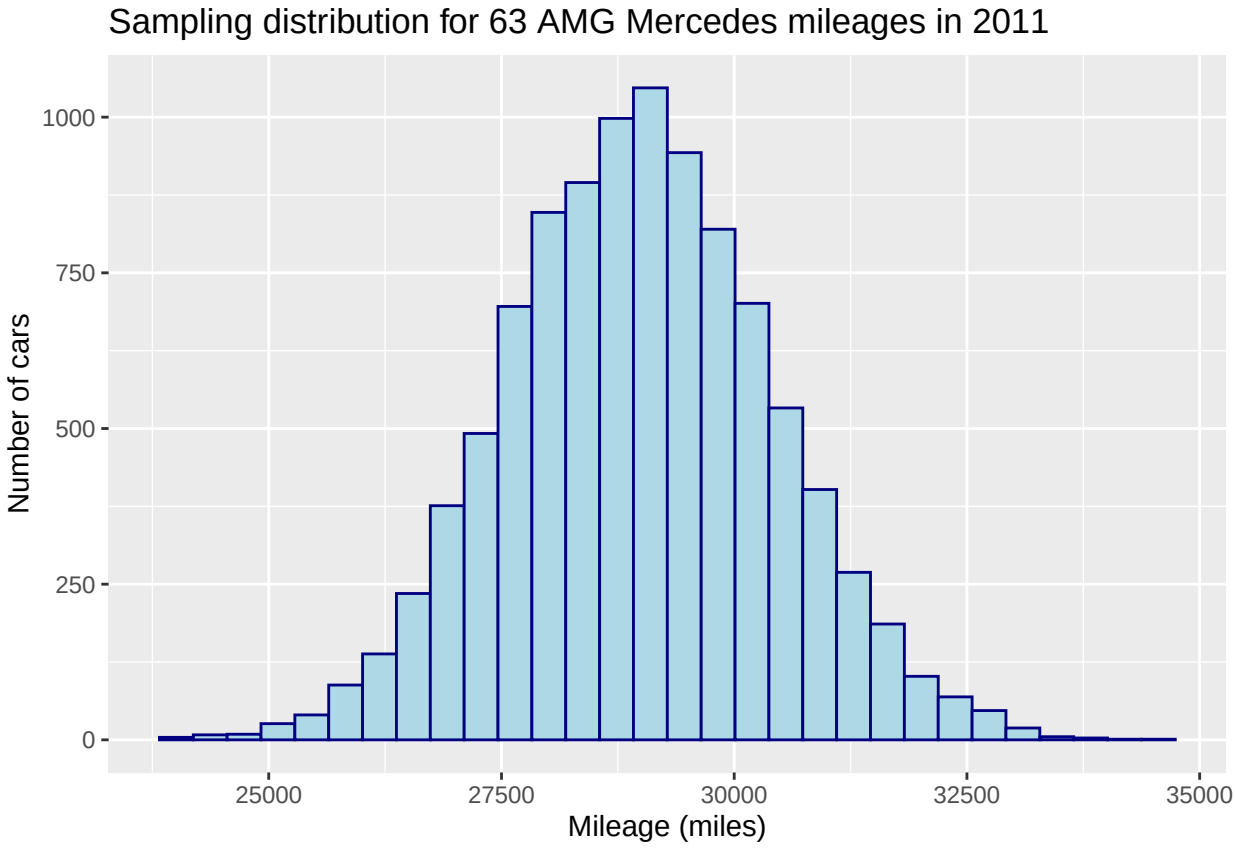



```
##   name    lower    upper level     method estimate
## 1 mean 1.848611 1.880198  0.95 percentile 1.864257
```

The claim is the brand Shell charges more than other brands of gas. This can be tested through 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 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 claim is supported.

Problem 2

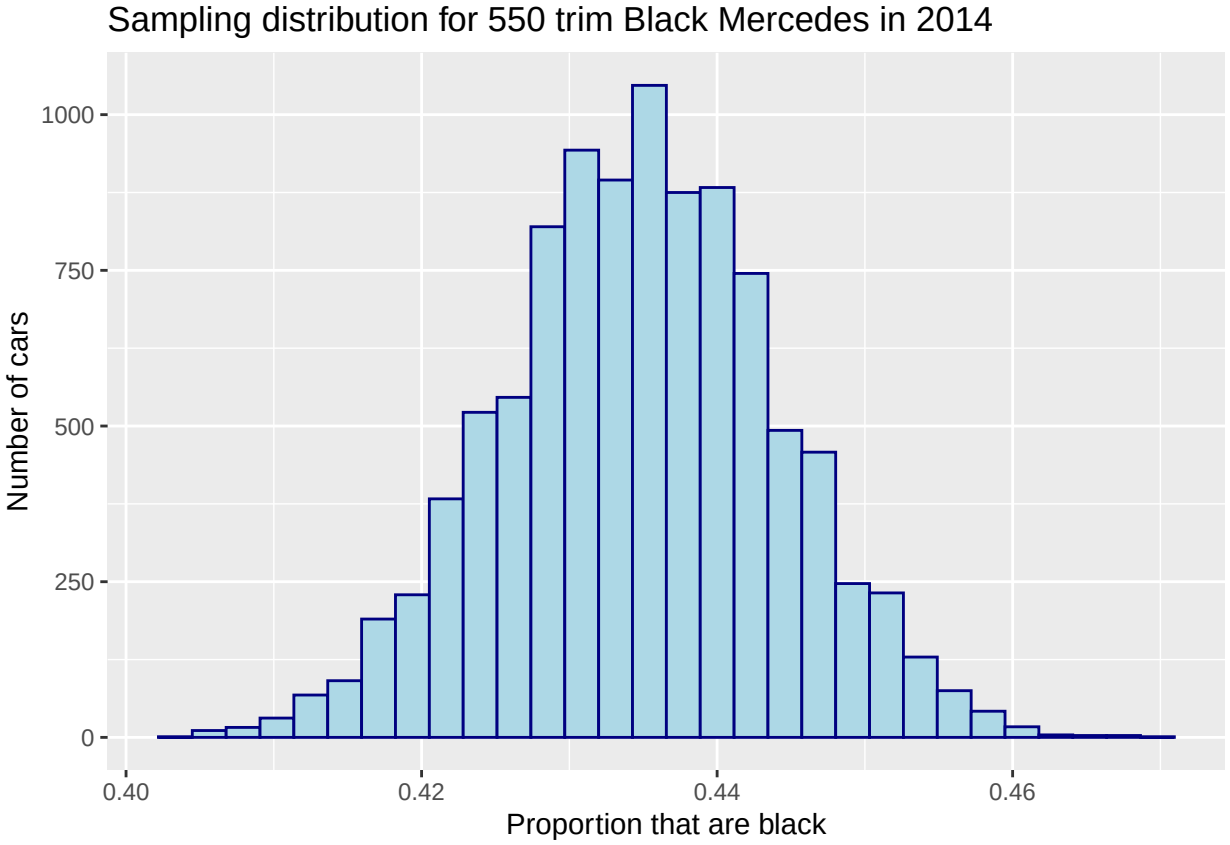
Part A




```
##   name    lower    upper level     method estimate
## 1 mean 26238.28 31813.78  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 0.4164071 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 result -0.396473 0.1018651  0.95 percentile -0.1490515
```

I am answering the question of which show makes people happier: Living with Ed or My Name is Earl. In order to answer this question I used the bootstrapping method. I made two data sets (one for each show) which only contained the name of the show and viewer's answer to the question Q1_Happy, which indicated how happy the show made them. I then bootstrapped the means of the data sets in order to compare them, and used 10,000 simulations in order to conclude my 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 -0.3541 to 0.01039. However, since our confidecne interval includes zero we cannot say that one show makes people more happy than the other. 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.

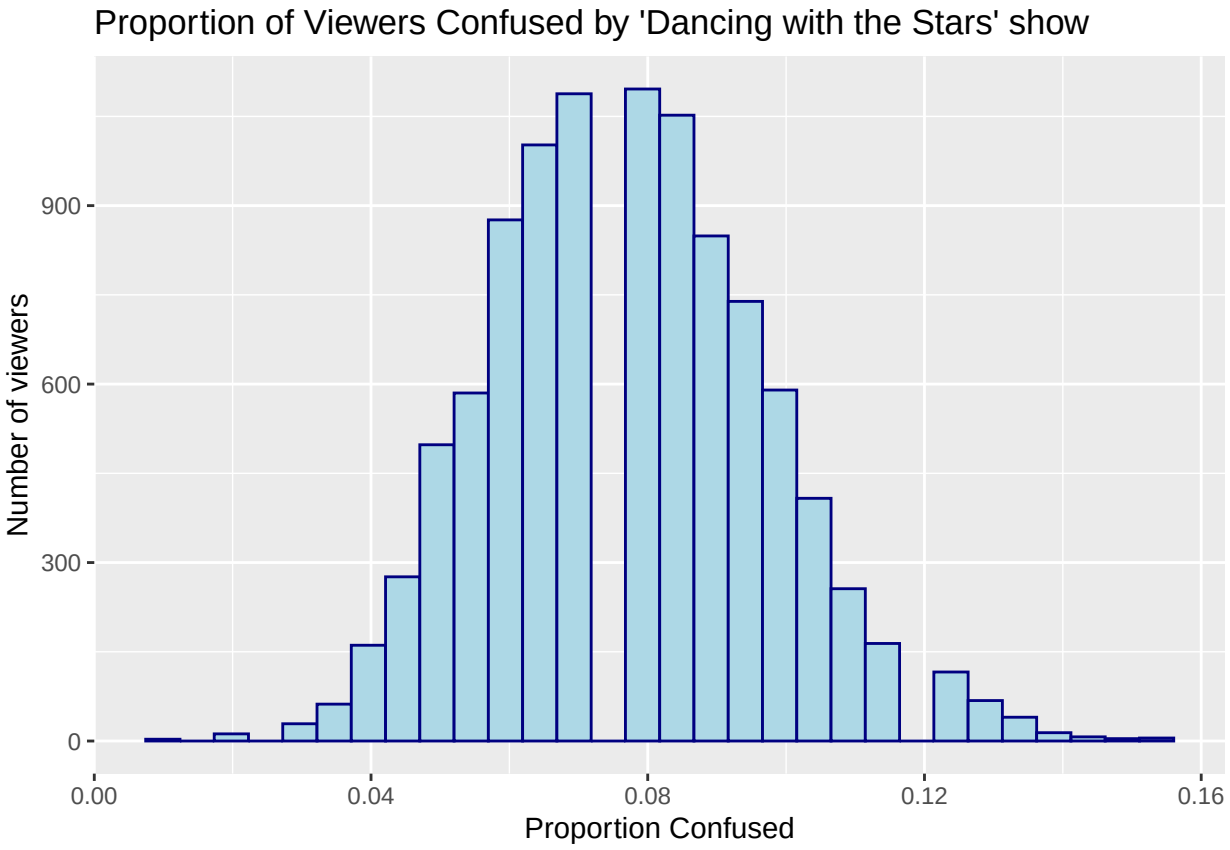
Part B

```
##     name      lower     upper level     method   estimate
## 1 result -0.3777957 0.1145485  0.95 percentile -0.1490515
```

I am answering the question of which show makes people annoyed: The Biggest Loser or The Apprentice: Los Angeles. In order to answer this question I used the bootstrapping method. I made two data sets (one for each show) which only contained the name of the show and viewer's answer to the question Q1_Annoyed, which indicated how happy the show made them. I then bootstrapped the means of the data sets in order to compare them, and used 10,000 simulations in order to conclude my 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 -0.393 to 0.104 However, since our confidence interval includes zero we cannot say that one show makes people more happy than the other. Therefore, we can conclude that based on our results there is no statistical significance that The Biggest Loser or The Apprentice: Los Angeles makes the average population of viewers more annoyed than the other show.

Part C

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



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". 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).

My approach provided me with a table containing the 95% confidence interval. We 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. 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".

Problem 4

```
##     name       lower       upper level     method    estimate
## 1 result -0.09063378 -0.01349869  0.95 percentile -0.05228145
```

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.

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. 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.

Using these results, it can be concluded with 95% confidence that DMAs where Ebay turned their Google ads off had, on average, revenue ratios that were 0.09086329 to 0.01336305 lesser than DMAs where Ebay did not turn their Google ads off.