HW 3 SDS 315

Somya Krishna, sk55256

2025-01-30

https://github.com/somya-k535/SDS-313-HW3.git

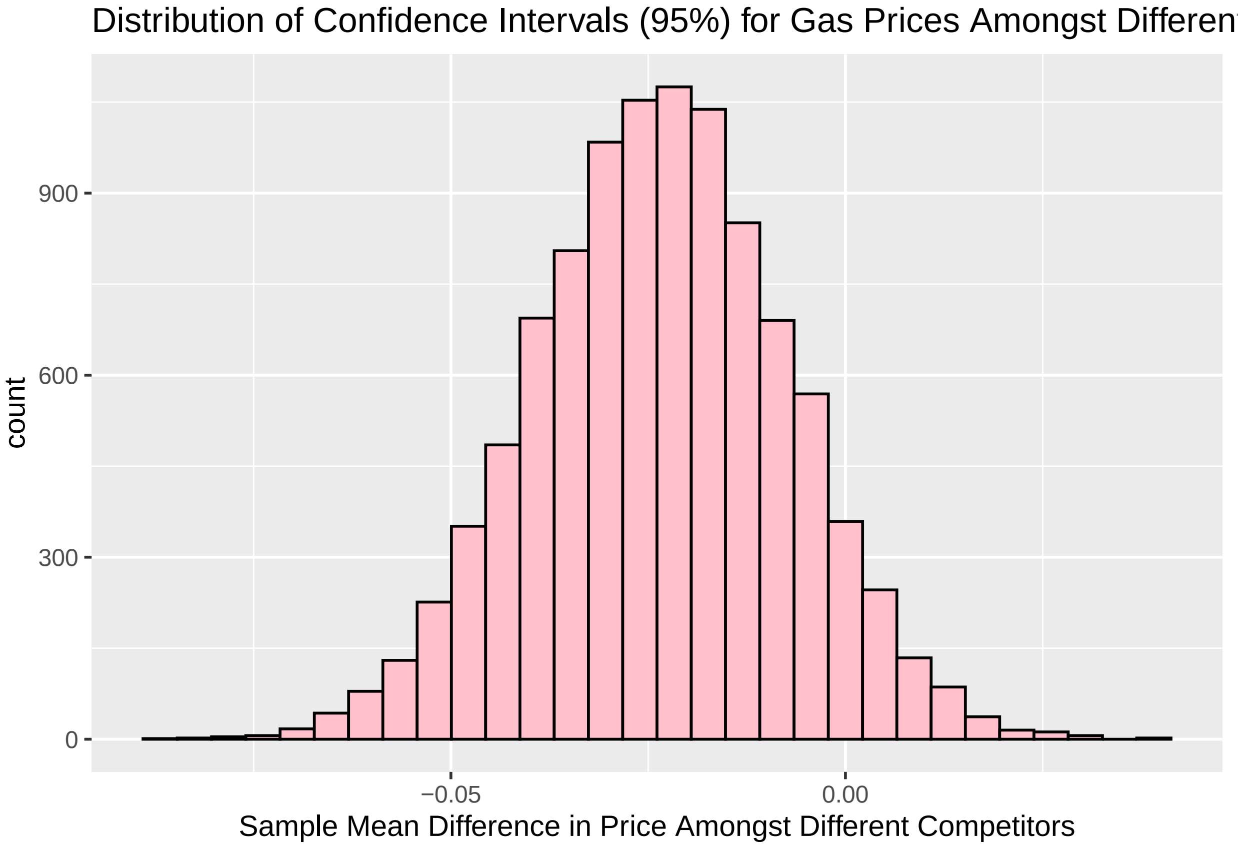
Problem 1: Theories

In this problem, we will dive into how UT Austin students view their professors.

Theory A

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

```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05521221 0.007455161  0.95 percentile -0.05121855
```



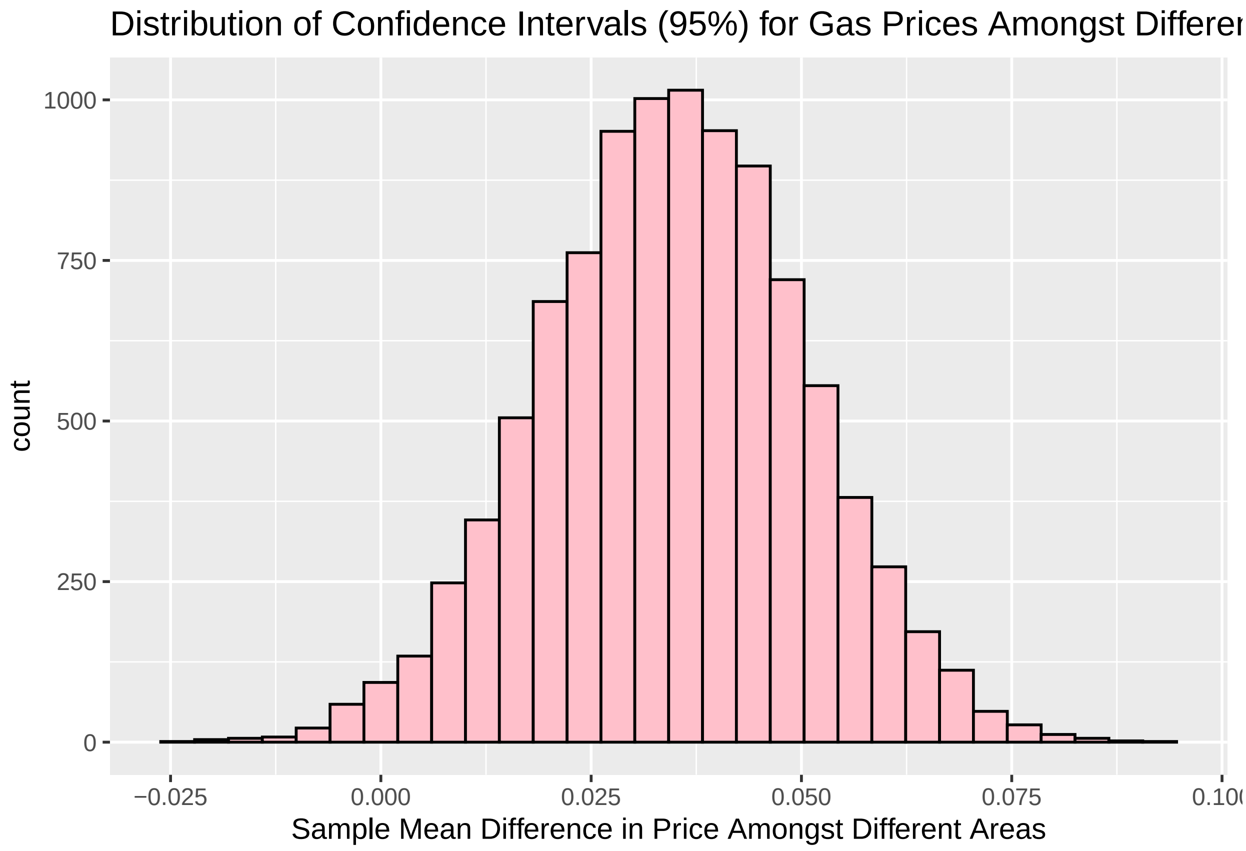
Evidence: To determine if gas stations charge more if they lack direct competition in sight, we must create a confidence interval using the Price (of gas at different gas stations in Austin) and Competitors (if there are competitors nearby) variables. The confidence interval had a lower bound of -0.055 and an upper bound of 0.007.

Conclusion: Because the confidence interval includes 0, there is not enough evidence to support that a change in gas station prices is related to whether or not competitors are in the area of the gas station. The theory is not supported by the data.

Theory B

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

```
##       name       lower     upper level     method   estimate
## 1 diffmean 0.004077953 0.0651072  0.95 percentile 0.04239868
```



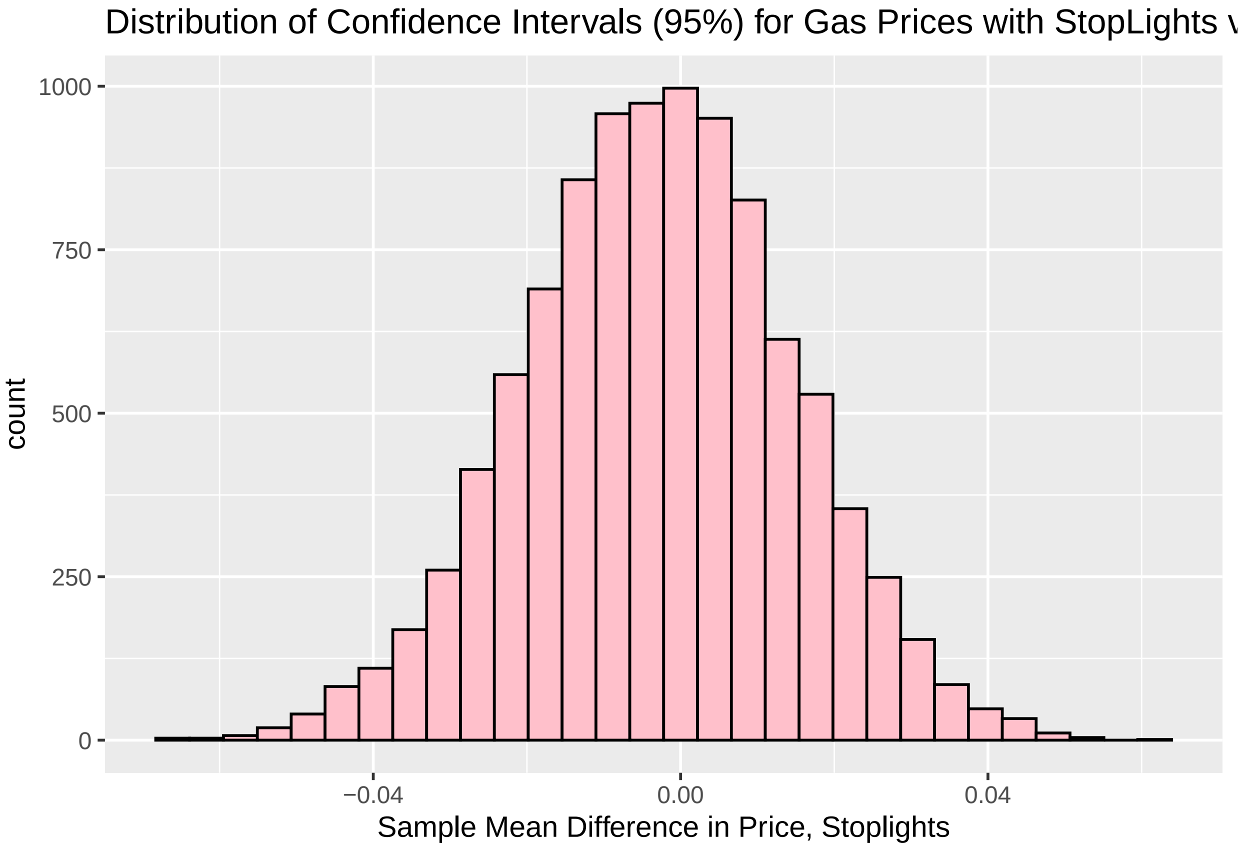
Evidence: To determine if gas stations charge more based on if the area of the gas station is richer, we must create a confidence interval using the Price (of gas at different gas stations in Austin) and is_rich(if the income is greater than the median income of Austin, Texas) variables. The confidence interval had a lower bound of 0.004 and an upper bound of 0.066.

Conclusion: Because the confidence interval does not include 0, there is enough evidence to support that a change in gas station prices is related to the income level of the gas station. The theory is supported by the data.

Theory C

Claim: Gas stations at stoplights charge more.

```
##       name       lower      upper level     method    estimate
## 1 diffmean -0.03787892 0.03078654  0.95 percentile 0.004105991
```



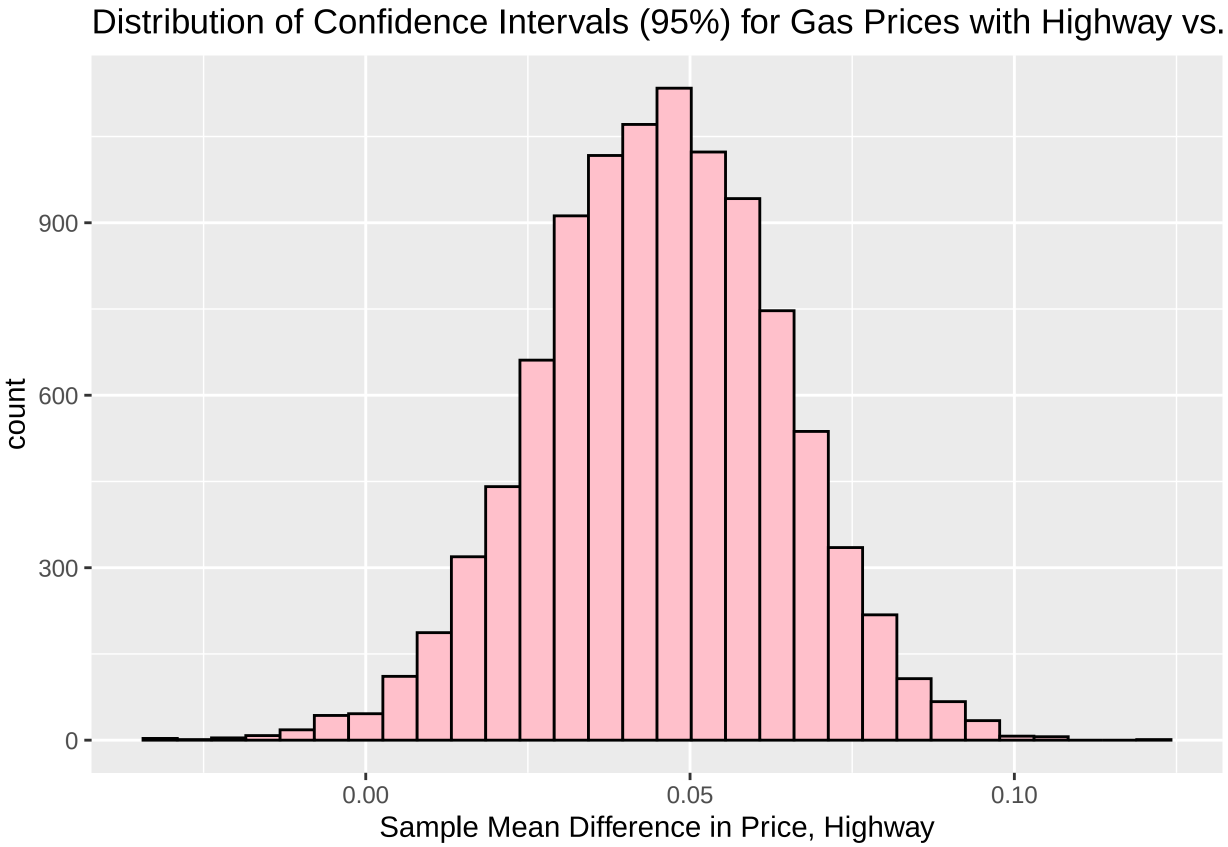
Evidence: To determine if gas stations charge more based on if there is a stoplight in front of the gas station, we must create a confidence interval using the Price (of gas at different gas stations in Austin) and Stoplight(if there is a stoplight in front of the gas station) variables. The confidence interval had a lower bound of -0.038 and an upper bound of 0.0302.

Conclusion: Because the confidence interval includes 0, there is not enough evidence to support that a change in gas station prices is related to whether or not there is a stoplight in front of the gas station. The theory is not supported by the data.

Theory D

Claim: Gas stations with direct highway access charge more.

```
##       name      lower      upper level     method   estimate
## 1 diffmean 0.00860061 0.08100228  0.95 percentile 0.07340513
```



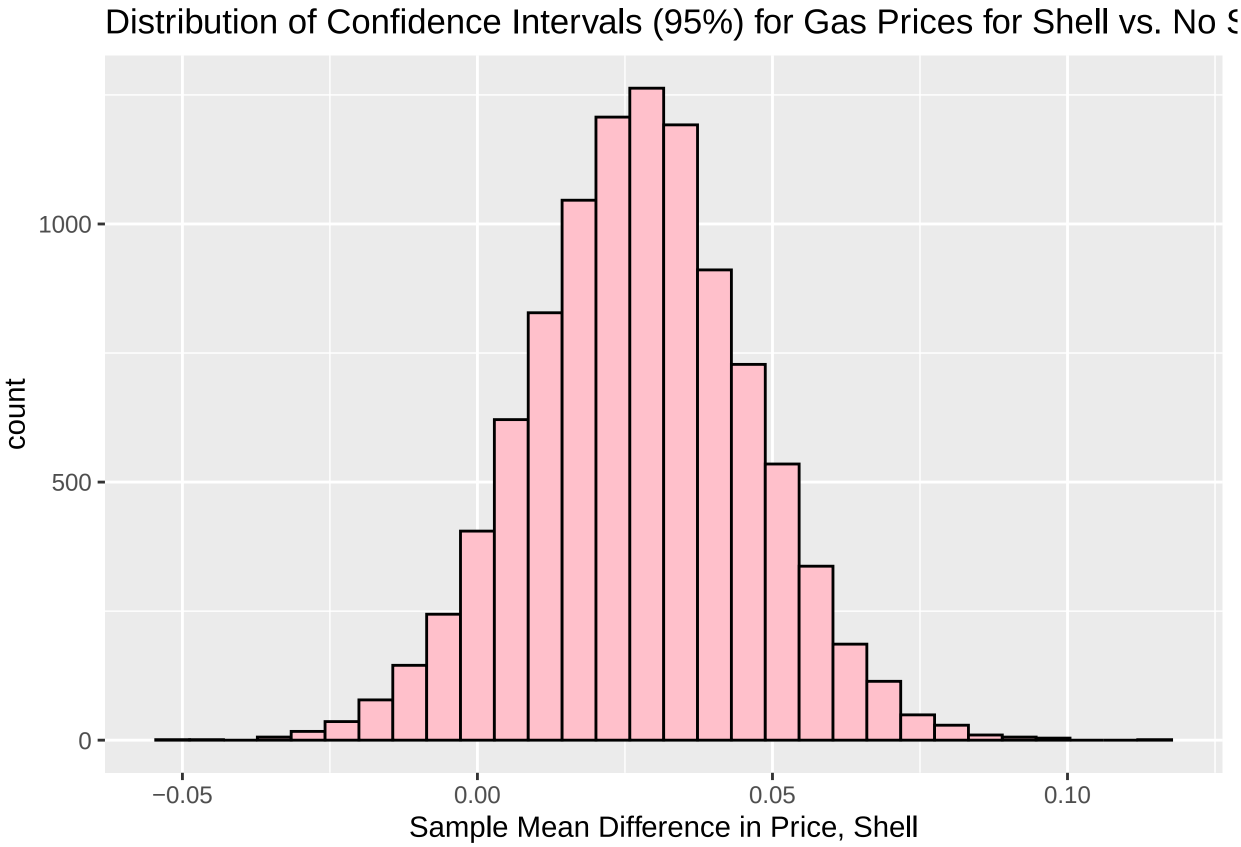
Evidence: To determine if gas stations charge more based on if the gas station is accessible from either a highway or a highway access road, we must create a confidence interval using the Price (of gas at different gas stations in Austin) and Highway(if the gas station is accessible from either a highway or a highway access road) variables. The confidence interval had a lower bound of 0.009 and an upper bound of 0.083.

Conclusion: Because the confidence interval does not include 0, there is enough evidence to support that a change in gas station prices is related to whether or not the gas station is accessible from either a highway or a highway access road. The theory is supported by the data.

Theory E

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

```
##       name        lower      upper level     method    estimate
## 1 diffmean -0.009408493 0.06455993  0.95 percentile 0.008126761
```



Evidence: To determine if Shell charges more than other gas stations, we must create a confidence interval using the Price (of gas at different gas stations in Austin) and is_shell(if the gas station is Shell or not) variables. The confidence interval had a lower bound of 0.009 and an upper bound of 0.082.

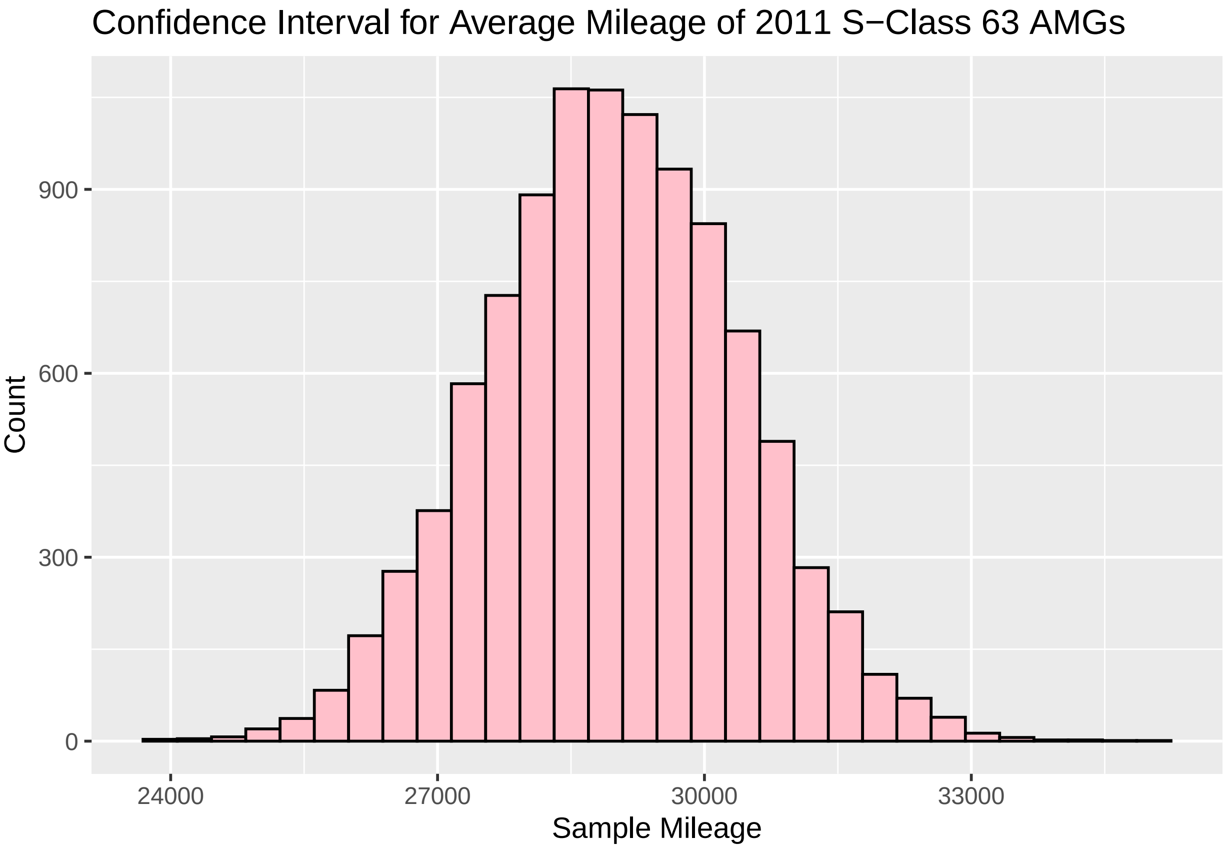
Conclusion: Because the confidence interval does not include 0, there is enough evidence to support that a change in gas station prices is related to whether or not the gas station is a Shell gas station or not. The theory is supported by the data.

Problem 2: Mercedes S-Class Vehicles

In this problem, we will investigate Mercedes S-Class vehicles sold on cars.com.

Part A

```
##   name    lower    upper level     method estimate
## 1 mean 26223.63 31762.52  0.95 percentile 29462.84
```

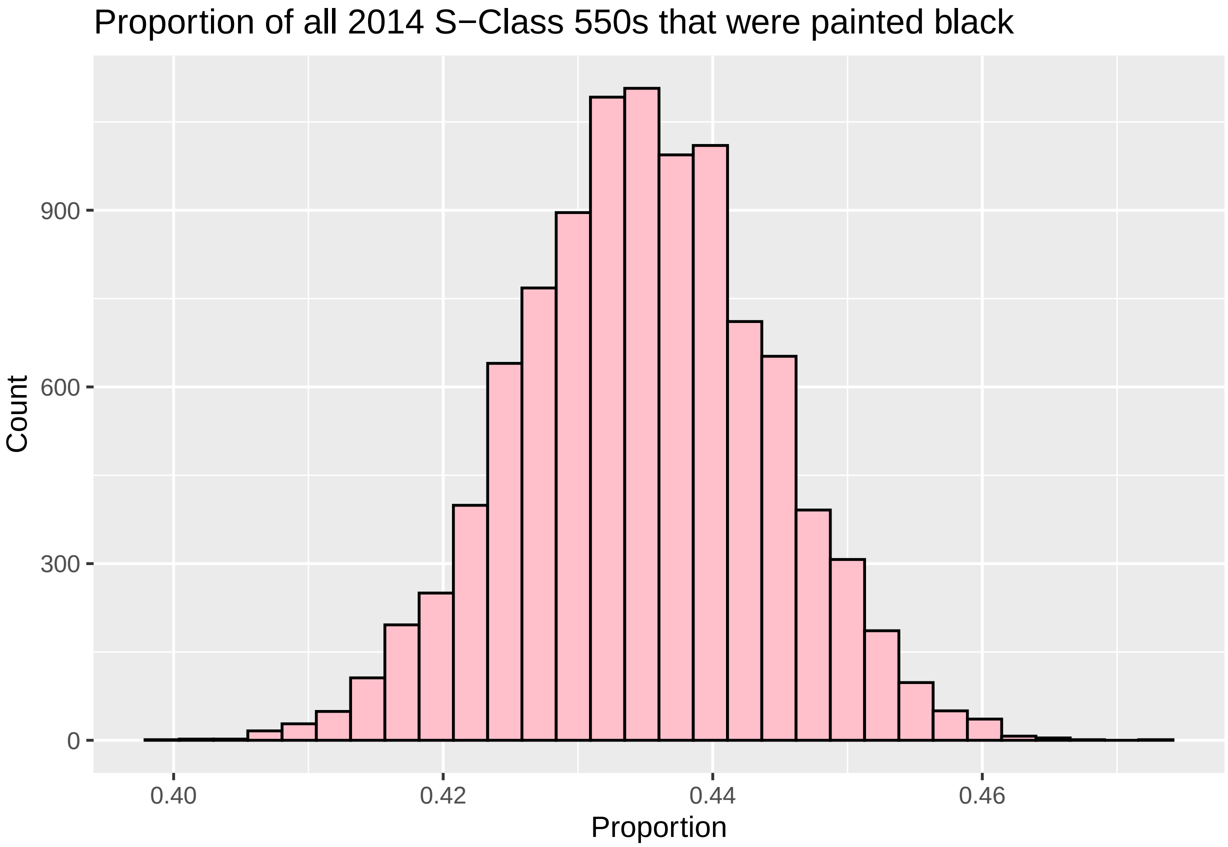


My task was to filter the data set down to include only those cars where year == 2011 and trim == "63 AMG". Based on these 116 cars, compute a 95% bootstrap confidence interval for the average mileage of 2011 S-Class 63 AMGs that were hitting the used-car market when this data was collected.

The lower bound of the confidence interval was 26362.76, and the upper bound was 31734.2. Since the confidence level was 95%, we are 95% confident that the average mileage of 2011 S-Class 63 AMGs that were hitting the used-car market when this data was collected is in between 26362.76 miles and 31734.2 miles.

Part B

```
##        name     lower     upper level     method  estimate
## 1 prop_TRUE 0.4164071 0.4527518  0.95 percentile 0.4323295
```



My task was to filter the data set down to include only those cars where year == 2014 and trim == "550". Based on this sample of 2889 cars, compute a 95% bootstrap confidence interval for the proportion of all 2014 S-Class 550s that were painted black. Hint: you might find this easiest if you use mutate to first define a new variable, isBlack, that is either TRUE or FALSE depending on whether the car is black.

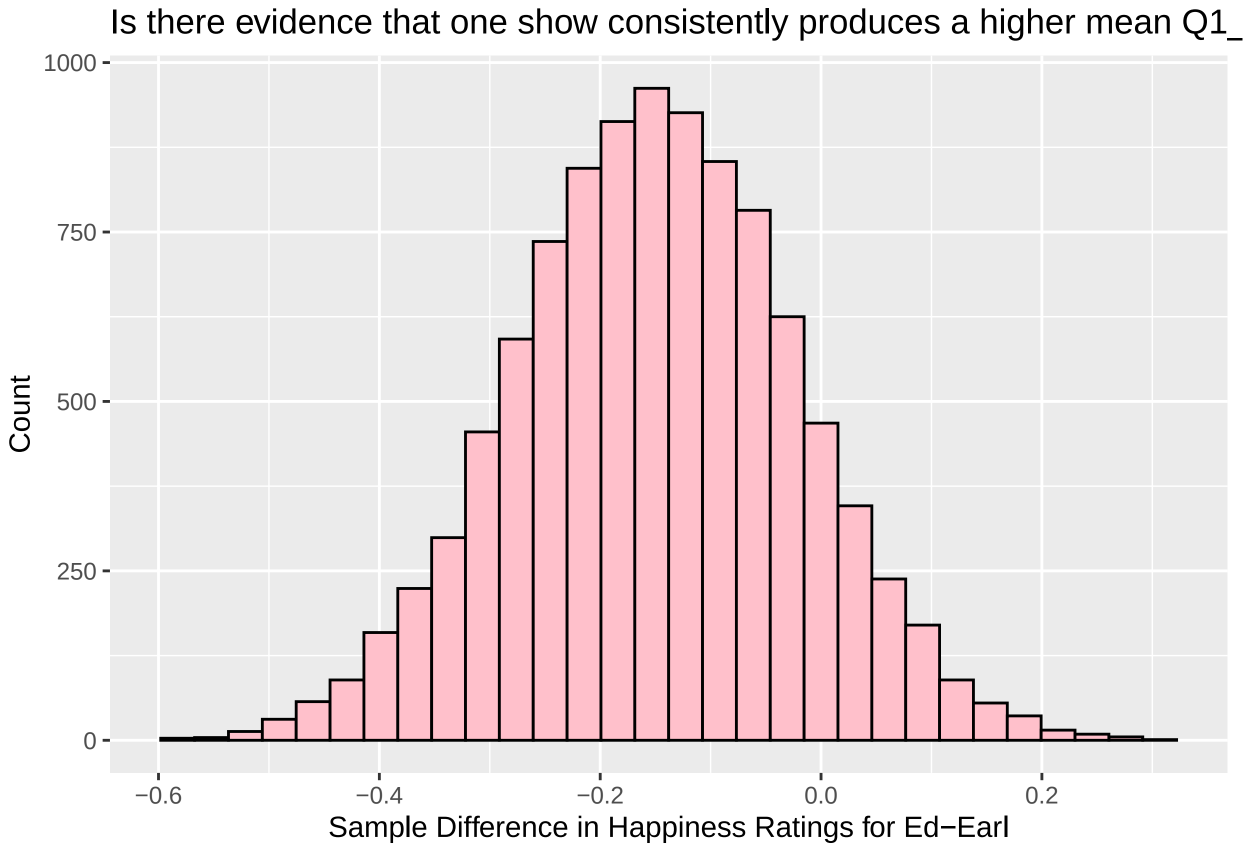
The lower bound of the confidence interval was 0.4167532, and the upper bound was 0.4527518 Since the confidence level was 95%, we are 95% confident that the proportion of all 2014 S-Class 550s that were painted black is in between 0.4167532 and 0.4527518.

Problem 3: NBC TV Viewer Survey

In this problem, we will investigate NBC's market research survey on how viewers respond to TV shows.

Part A

```
##       name      lower      upper level     method   estimate
## 1 diffmean -0.4008732 0.09616395  0.95 percentile -0.1722585
```



1) Question: What question are you trying to answer? I am trying to find out if the show *Living with Ed* or *My Name is Earl* makes people happier.

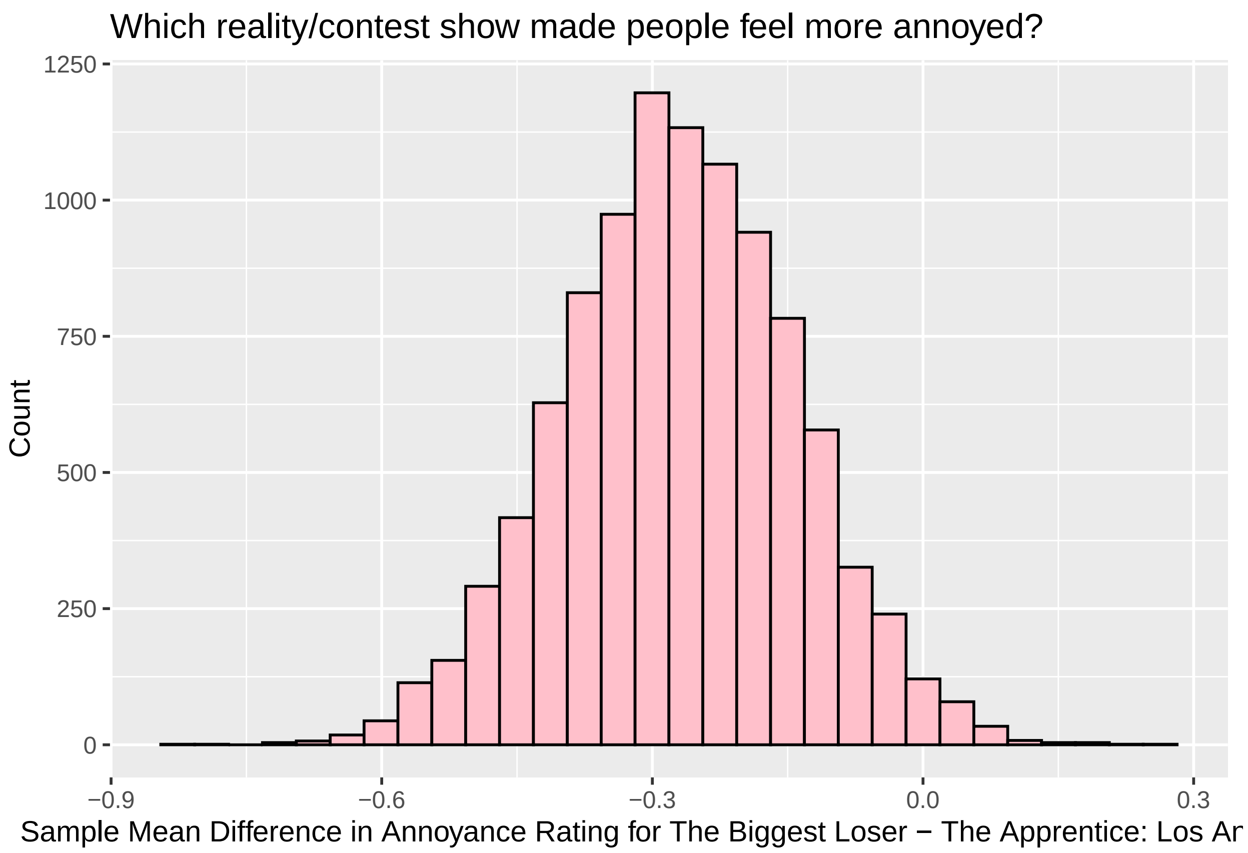
2) Approach: What approach/statistical tool did you use to answer the question? I created a new dataset to just show the shows *Living with Ed* and *My Name is Earl*. I used bootstrapping with diffmean with Q1_Happy (happiness ratings) and Show (Which show it was). I then conducted a confidence interval at 95% confidence and plotted the results using a histogram.

3) Results: What evidence/results did your approach provide to answer the question? (E.g. any numbers, tables, figures as appropriate.) Make sure to include appropriate measures of uncertainty! The confidence interval has a lower bound of -0.3958495 and an upper bound of 0.1067168. The estimated difference was 0.01561044.

4) Conclusion: What is your conclusion about your question? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set. Since the interval contains 0, there is no statistically significant evidence that the difference in happiness ratings is related to whether or not the show is *Living with Ed* or *My Name is Earl*. We are 95% confident that the difference in happiness ratings between the two shows is in between a decrease of 0.3958495 and an increase of 0.1067168.

Part B

```
##       name      lower       upper level     method   estimate
## 1 diffmean -0.5274894 -0.01861493  0.95 percentile -0.4589944
```



1) Question: What question are you trying to answer? I am trying to find out if the show *The Biggest Loser* or *The Apprentice: Los Angeles.* makes people more annoyed.

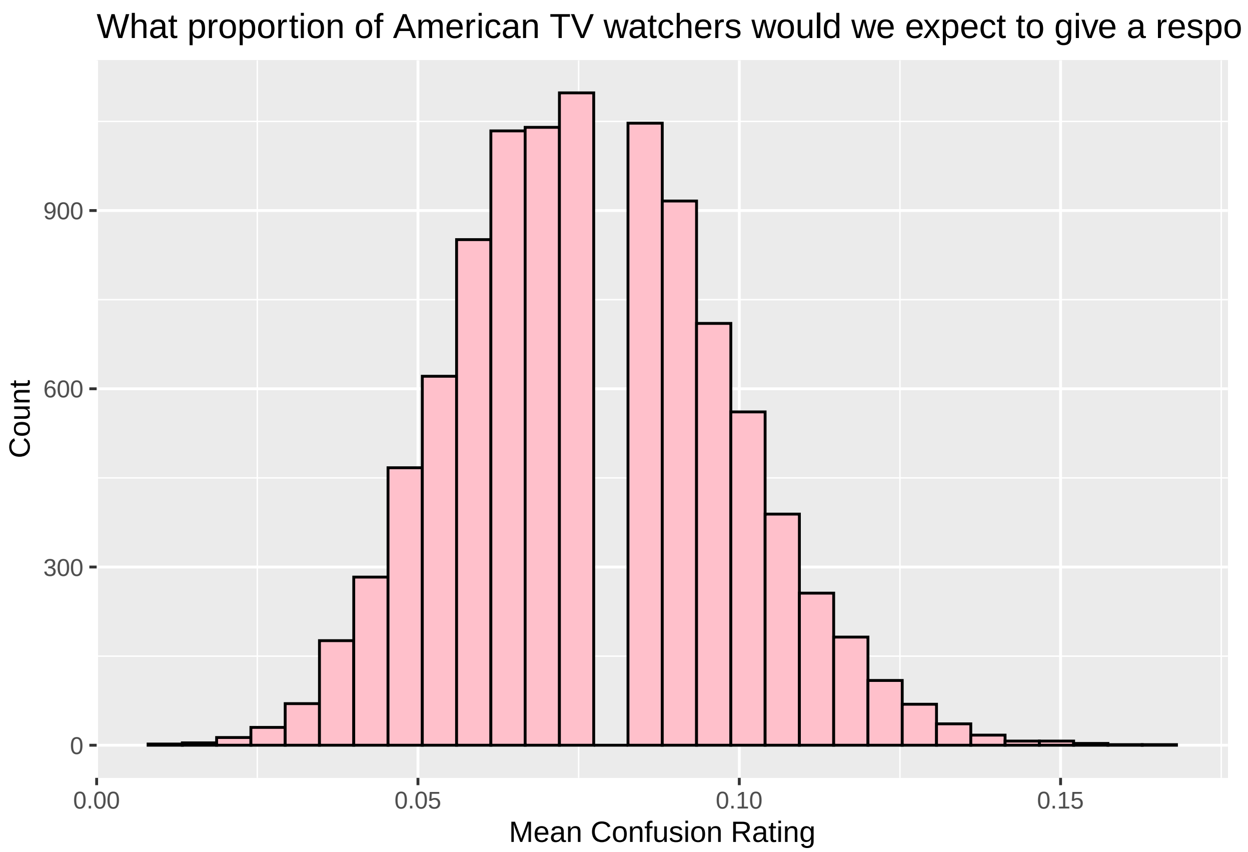
2) Approach: What approach/statistical tool did you use to answer the question? I created a new dataset with the shows filtered to *The Biggest Loser* and *The Apprentice: Los Angeles.* and dropped n/a values. I used bootstrapping with diffmean with Q1_Annoyed (annoyance ratings) and Show (Which show it was). I then conducted a confidence interval at 95% confidence and plotted the results using a histogram.

3) Results: What evidence/results did your approach provide to answer the question? (E.g. any numbers, tables, figures as appropriate.) Make sure to include appropriate measures of uncertainty! The confidence interval has a lower bound of -0.5197406 and an upper bound of -0.01976126. The estimated difference was -0.1803638.

4) Conclusion: What is your conclusion about your question? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set. Since the interval does not contain 0, there is statistically significant evidence that the difference in annoyance ratings is related to whether or not the show is *The Biggest Loser* or *The Apprentice: Los Angeles.*. We are 95% confident that the difference in annoyance ratings between the two shows is in between a decrease of 0.5197406 and a decrease of 0.01976126.

Part C

```
##   name      lower     upper level     method   estimate
## 1 mean 0.03867403 0.1161602  0.95 percentile 0.06629834
```



1) Question: What question are you trying to answer? I am trying to find out what proportion of American TV watchers we would expect to give a response of 4 or greater to the "Q2_Confusing" question for the show *Dancing with the Stars*.

2) Approach: What approach/statistical tool did you use to answer the question? I created a new filtered dataset. I filtered the show to just show Dancing with the Stars, dropped n/a values, and created a variable called confusing to tell whether or not people find Dancing with the Stars confusing. I used bootstrapping with mean with confusing (confused or not). I then conducted a confidence interval at 95% confidence and plotted the results using a histogram.

3) Results: What evidence/results did your approach provide to answer the question? (E.g. any numbers, tables, figures as appropriate.) Make sure to include appropriate measures of uncertainty! The confidence interval has a lower bound of 0.03867403 and an upper bound of 0.1160221 The estimated difference was 0.03314917.

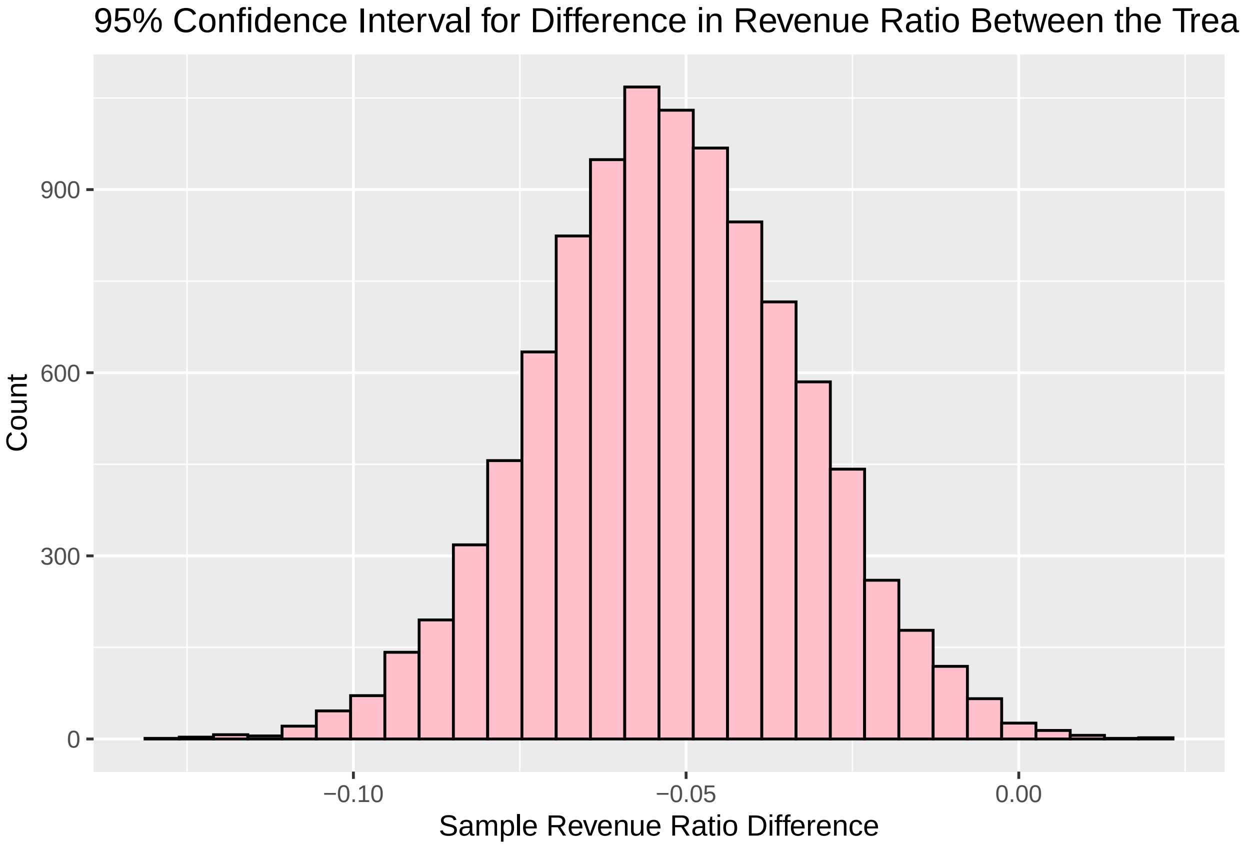
4) Conclusion: What is your conclusion about your question? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set. We are 95% confident that the proportion of American TV watchers we would expect to give a response of 4 or greater to the "Q2_Confusing" question for the show *Dancing with the Stars* is between 0.03867403 and 0.1160221. The value is likely to be around 0.03314917.

Problem 4: EBay

In this problem, we will investigate data from an experiment run by EBay in order to assess whether the company's paid advertising on Google's search platform was improving EBay's revenue

Confidence Interval for Difference in Revenue Ratio between the Treatment and Control DMAs

```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.09139034 -0.01353314  0.95 percentile -0.06220969
```



1) Question: What question are you trying to answer? I want to know 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.
2) Approach: What approach/statistical tool did you use to answer the question? I created a new dataset from the ebay dataset to create a new variable called revenue ratio which divides rev_after by rev_before. I also created a variable called treatment to determine if the DMA was in the treatment group or if the DMA was in the control group.
3) Results: What evidence/results did your approach provide to answer the question? (E.g. any numbers, tables, figures as appropriate.) The confidence interval has a lower bound of -0.09113869 and an upper bound of -0.01287033 The estimated difference was -0.03678399
4) Conclusion: What is your conclusion about your question? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set. Since the interval does not contain 0, there is statistically significant evidence that the revenue ratio is the not the same in the treatment and the control groups. We are 95% confident that the difference in revenue ratio between the treatment group and control group is in between a decrease of 0.09113869 and a decrease of 0.01287033.