HW 3

Genavieve Middaugh

2025-02-07

GitHub link - https://github.com/crrankyypants/HW3-

Problem 1 - Gas Station Prices

In Problem 1 we are assessing multiple theory given the prices of gas stations with multiple outside factors. The goal is to try to understand the variation in gas station prices and what causes that variation. The data given by GasPrices.csv a dataset made by students in spring of 2016 on a project.

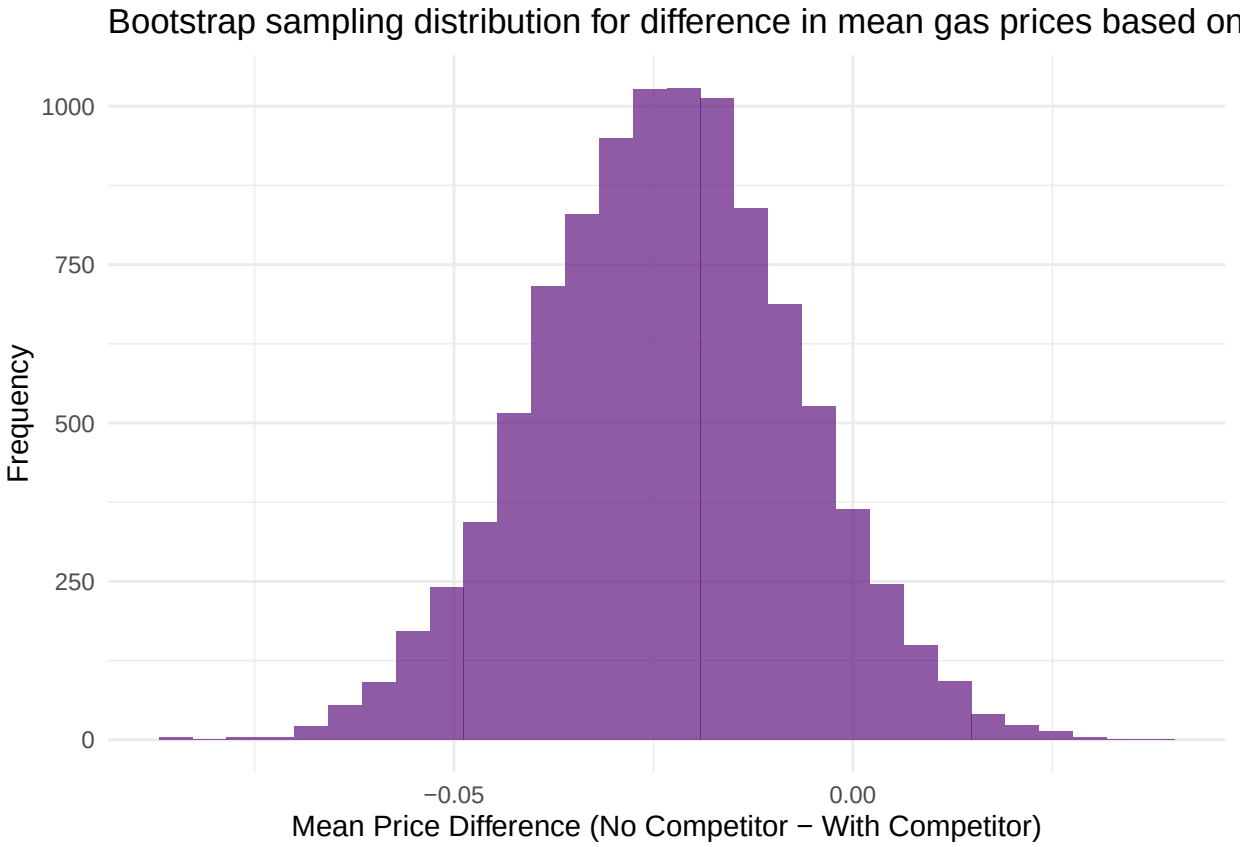
The important variables are below:

- ID: Order in which gas stations were visited
- Name: Name of gas station
- Price: Price of regular unleaded gasoline, gathered on Sunday, April 3rd, 2016
- Highway: Is the gas station accessible from either a highway or a highway access road?
- Stoplight: Is there a stoplight in front of the gas station?
- Competitors: Are there any other gas stations in sight?
- Zipcode: Zip code in which gas station is located
- Income: Median Household Income of the ZIP code where the gas station is located based on 2014 data from the U.S. Census Bureau
- Brand: ExxonMobil, ChevronTexaco, Shell, or Other.

The theorys we are trying to either prove or disprove are below:

A) Gas stations charge more if they lack direct competition in sight.
B) The richer the area, the higher the gas prices.
C) Gas stations at stoplights charge more.
D) Gas stations with direct highway access charge more.
E) Shell charges more than all other non-Shell brands.

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



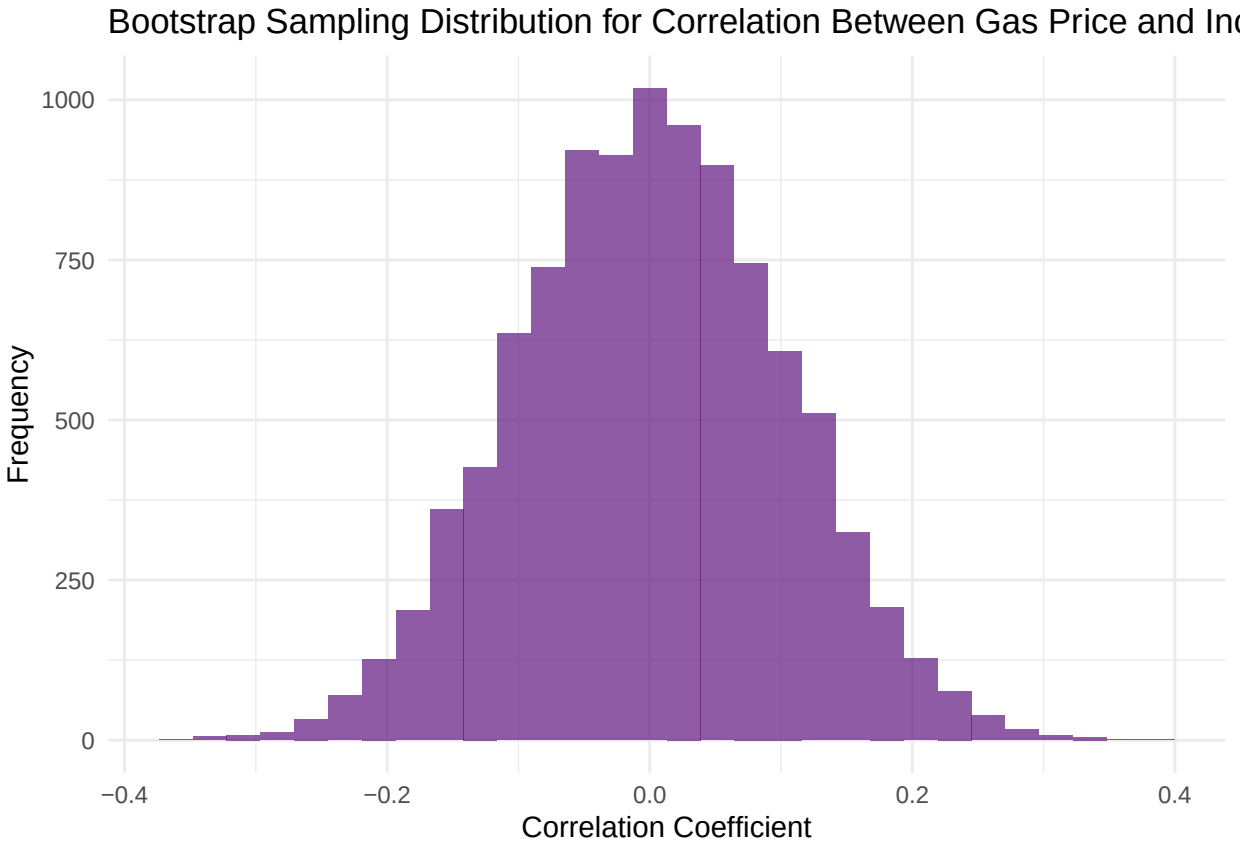
If you have ever driven around, anywhere in america im sure you've seen your fair share of gas stations, the fluctuation of prices varies from place to place, but one may wonder if having competitors near by lead to gas stations having lower prices. The Competitors and Price variables from gasprices data was used to understand if there were competitors around or not and what the prices were for each gas station.

The above graph shows a bootstrap sampling distribution of the difference in the means of these 2. The graph skews to the left of 0 only slightly, so we can say with 95% confidence that the average difference in gas price between gas stations with no competitors and with competitors is between -0.055 and 0.008 cents. Since the graph is so close to 0, this means it is a common occurrence for there to be no difference, meaning the average prices were the same or at least almost the same. It seems that there is very little correlation especially considering the mean is -0.023, a very low number and when you think about the difference in gas prices this looks like 2.34 and 2.32 difference is quite small.

To conclude, the data does not present any real evidence of the idea that gas stations prices are correlated to whether there are competitors nearby. But we must remember that this data was only done over the course of 1 day, in the spring of 2016 in Austin. This data does not cover enough time for us to draw conclusions on if over time this is the same conclusion.

2. The richer the area, the higher the gas prices.

Because gas stations have various different pricing depending on where you go in America, its reasonable to quesiton if the general income level of the community the gas station is in has any affect on the prices of gas. Below is a graph that is a distribution of samples looking for correlation between gas price and income of the community it is located in.



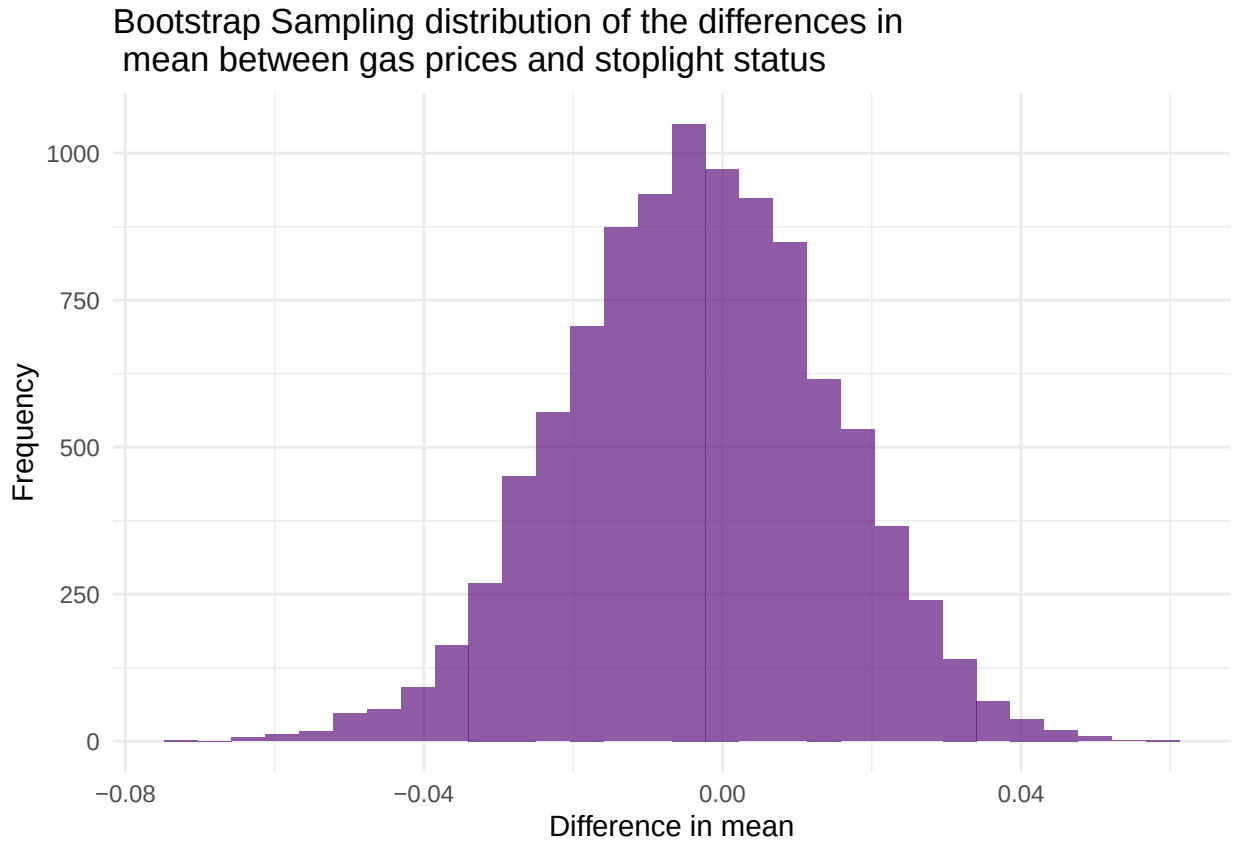
Given the Distribution we can say with 95% confidence that the correlation between gas prices and the average income of the area around it is between -0.194 and 0.198. This confidence interval includes 0 meaning that there is not much evidence of these variables causing one another. The income of the area does not seem to suggest the price of gas in the same area. There is a wide range of correlation values between incomr and gas prices, and 0 is right in between the bell shaped distribution, which suggest there is no correlation between the two.

This graph shows the correlation between the two variables, which is 0.396. The correlation is moderate or at least above 0, which suggest some relationship between the two variables.

After looking at these two facts, the initial correlation between the two variables shows there is a moderate correlation between the two, but the sampling distribution shows that it varies and that there is almost no correlation. I think with this information we can conclude that the data is two small, and done over such a short amount of time, we cannot conclude anything about gas stations and their prices based on their average income level using this data.

3. Gas stations at stoplights charge more.

The next theory is that some Gas stations charge more for gas if it is near a stop light. I think this is a reasonble idea, but lets see if it follows through.

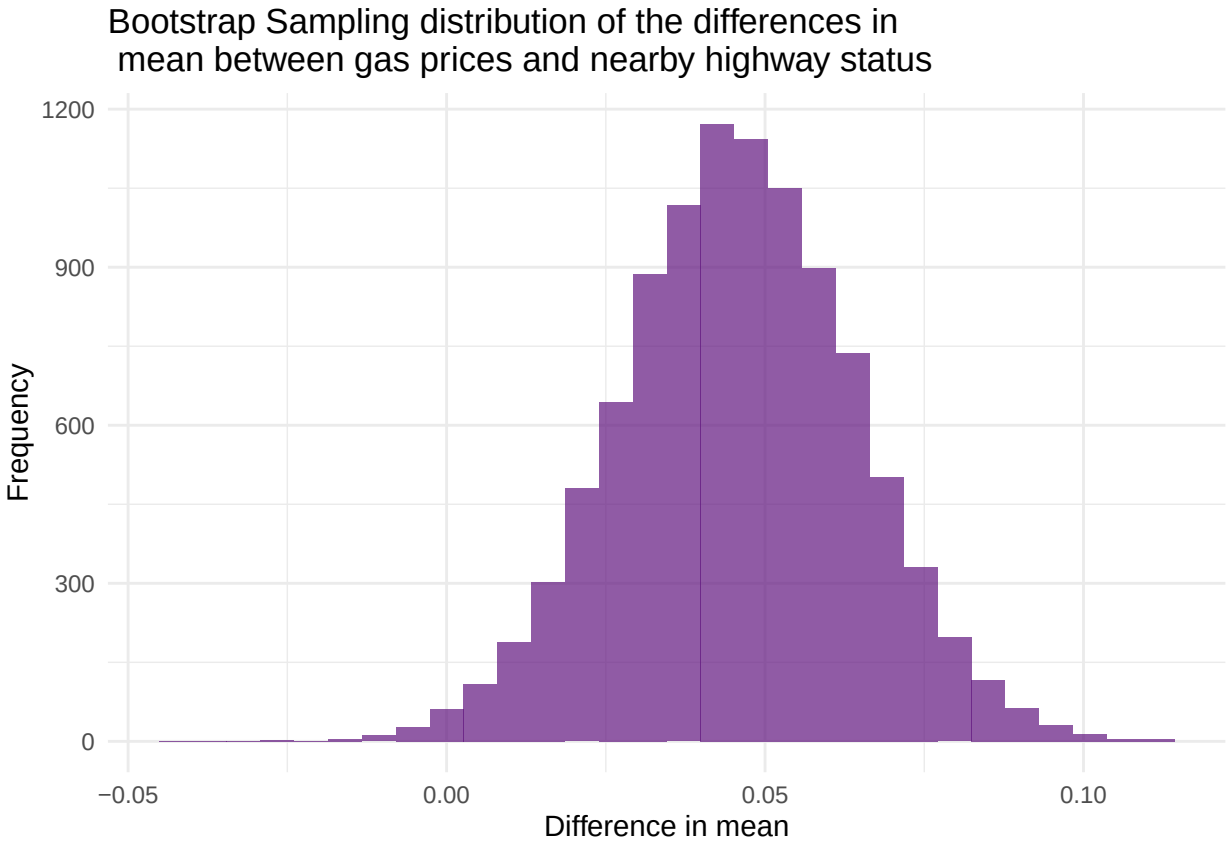


After bootstrap sampling the distribution of the Price and Stoplight variables we find that the mean price for gas stations if there is no light is 1.866 and for gas stations near a light were 1.863. This graph is clearly bell shaped with a center ranging near zero, meaning that the difference in means often ends up being zero or close, this means there is little difference between the means.I can say with 95% confidence that the difference in mean between the stores with stop lights and without gas prices was between -0.038 and 0.03 cents. Zero is included in the interval once again meaning there is too many outcome and a lot of variance.

The conclusion here is that we once again cant say for sure if there is any evidence of gas stations charging more for gas if they are near a gas station.

4. Gas stations with direct highway access charge more.

Related to the last theory this one asks if gas stations with direct highway access charge more for gas.

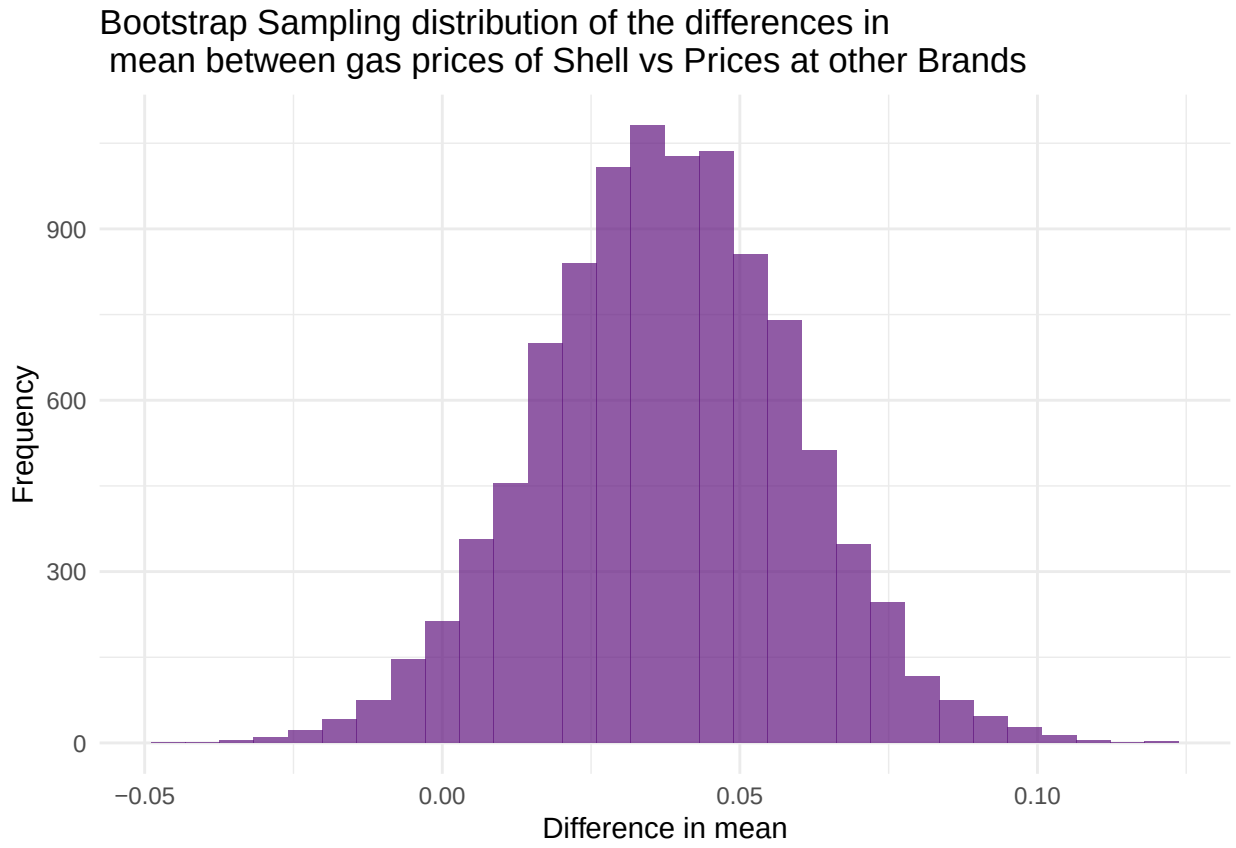


First we can see that the general mean from the data for stores not near a highway was 1.854 and for stores near a highway 1.9. They are quite close in range and don't seem to differ, although we can acknowledge that the stores not near highways do have a lower average in price. In the Graph above it shows the sampling distribution of the difference in means from these two variables. We see that it is bell shaped and consistent, and we also see that these are higher values, meaning that there were differences in the mean, suggesting that stores near highways charge more. So I am 95% confident that the difference of means between gas prices and their highway status is between 0.009 and 0.082 cents.

In conclusion , given that the confidence interval is above 0, The graph does show that there is some evidence, stating that stores near highways charge more for gas,but its not a strong difference in means so its only a possibility. So we can say that on the day the data was gathered and the area (2016, and in Austin), that on this day there was slight evidence that gas stations near highways charged more for gas.

5. Shell charges more than all other non-Shell brands.

This last theory asks if Shell charges more for gas than any other brand. I have always personally avoided Shell, becasue they seem to always be the most expensive gas around me, so I am curious what the information will show.



The graph above is consistent and slightly above zero, this could suggest that there is a slight difference in price between Shell and other brands gas prices. When I get the confidence interval I am 95% confident that the difference of mean prices between Shell brand and other brands gas prices is between NA and NA cents. This interval does not include zero, but it is pretty low meaning there is a slight significance to this evidence. Shells average price given this data is 1.88 which the others Chevron is 1.88 , exxon is 1.86 and other is 1.85 approximatly.

To conclude, this data shows that there is a slight significance to the evidence of Shell being more expensive than other brands, but not very signifigant. The difference in mean of price is low, and doesn't have a high range.

Problem 2 - Mercedes S-Class

The dataset has data on 30,000 used Mercedes S-Class vehicles sold on cars.com. Mercedes is a popular luxury car brand and here we will be looking at

- price
- mileage
- trim
- color
- year

Part A

In part A, we need to look at cars from 2011 with the trim being 63 AMG, and we are going to simulate a boostrap sampling distribution to find the average mileage of the 2011 S-Class 63 AMG at this time when the data was gathered.

After making those calculations, I am 95% confident that the average mileage of 2011 S-Class 63 AMG cars hitting the used car market falls between 2.6317×10^4 and 3.1816×10^4.

Part B

In part B, we are looking for the proportion of all 2014 S-Class 550s that were Black. To do this we will filter the dat, mutate the data, and find a 95% confidence interval.

I filtered the data to only contain 2014 years and trim 550 and also added a variable that indicated if the car was black or not using true and false. After bootstrapping to find the proportion, so I am 95% confident that the proportion of all 2014 S-Class 550's that are black is between 41.68% and 45.28%.

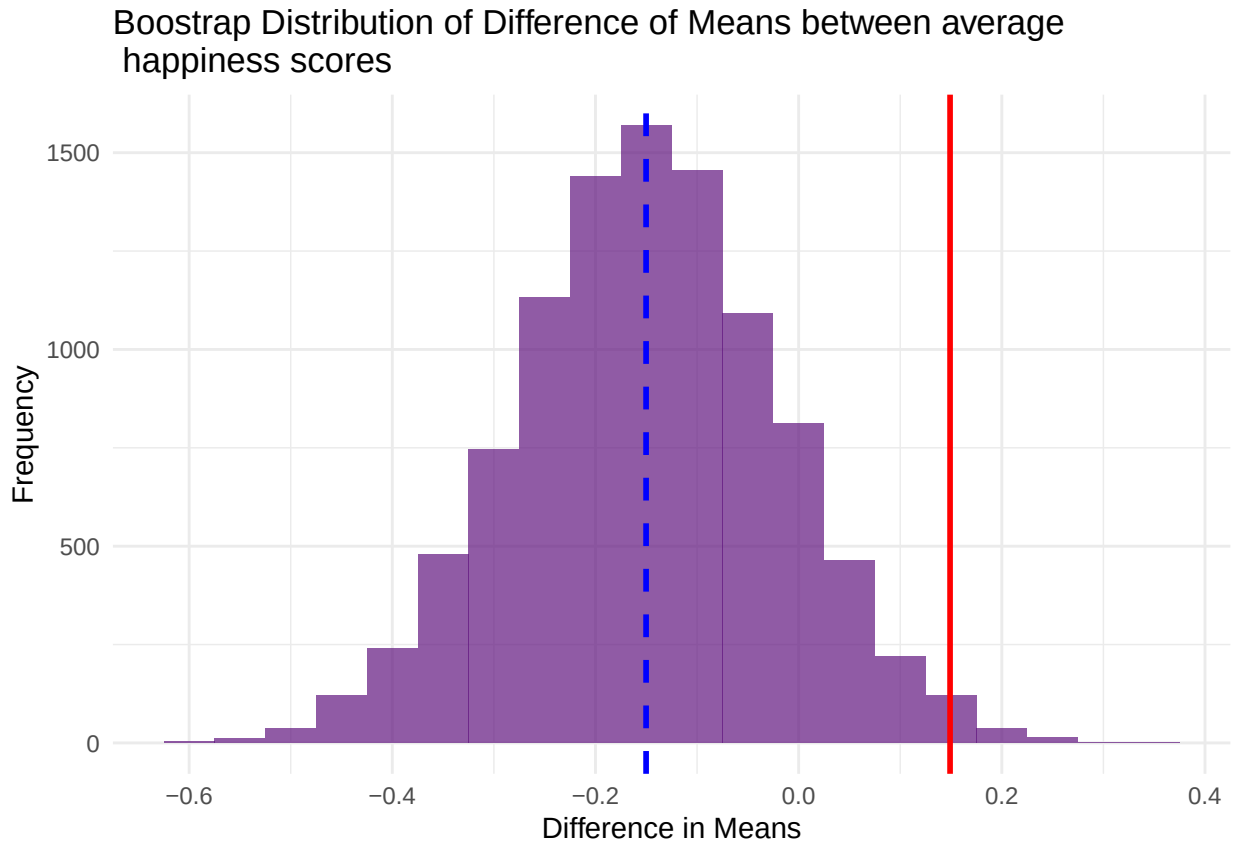
Problem 3 - NBC Viewer Response to TV shows

This problem uses nbc_pilotsurvey.csv which is a data set that contains information on a research on how viewers respond to TV shows. Each row of this data frame shows the responses of a single viewer (the Viewer variable) to the "pilot" episode1 of a single TV show (Show variable). The remaining variables encode the viewers reactions to the show.

1) Question: What question are you trying to answer?
2) Approach: What approach/statistical tool did you use to answer the question?
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!
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. These questions are fairly simple, so we'd expect each of these four sections for each part to be quite short—likely no more than 1-3 sentences each

Part A.

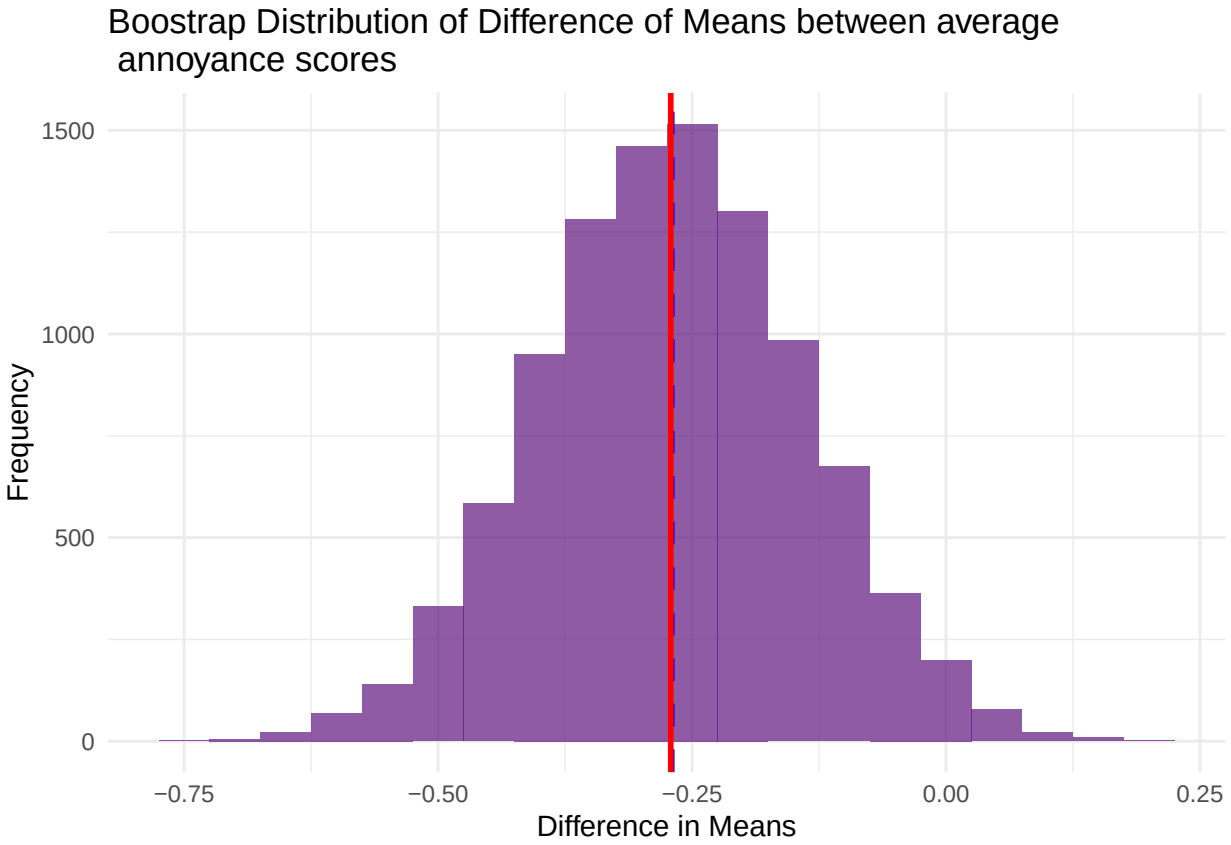
Here we are focusing on the shows "Living with Ed" and "My Name is Earl". Which one makes people happier. Is there evidence that one show consistently produces a higher mean Q1 Happy response among viewers?



To find the answer to this question, first we will filter the data to contain only "Living with Ed" or "My Name is Earl". We create a bootstrap to simulate this distributions difference of means between the certain show and the Q1_happy variable that has the rating of is this show made the viewer happy at all. We can then say with 95% confidence that the difference of means between the happiness rating and the Show watched is between -0.41 and 0.1 points. This data gives answers to the question in that we can see the interval includes 0 which means implies that there was no difference but the range goes to almost -0.5 to positive .25. The range is large and the red line marking where the observed difference lies which is no where near the center of the data. the observed difference from the data is 0.1490515. But overall we see that its center is in the negatives, and go into deeper negatives than the positive side. this could mean that My name is Earl is making people happier overall. But since the confidence interval includes 0 there is too much variety and suggest there is no statistically significant difference in the happiness scores. Overall there is no difference that we can say for sure.

Part B.

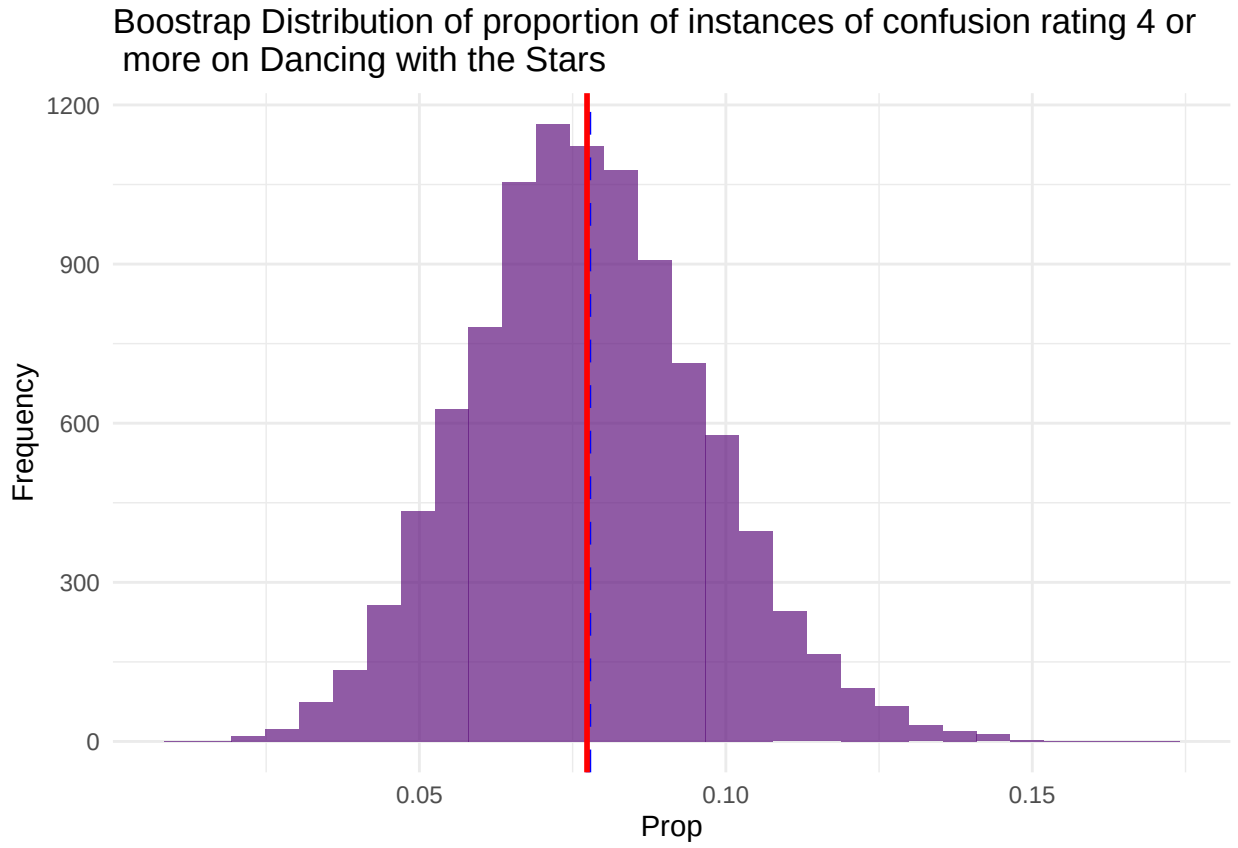
Consider "The Biggest Loser" and "The Apprentice: Los Angeles." Which annoyed people more? Is there evidence that one show consistently produces a higher mean Q1_annoyed response among viewers?



Using a bootstrap sampling distribution of the Q1_Annoyed variable on data containing only "The Biggest Loser" and "The Apprentice: Los Angeles", we got a observed difference of means between these two shows Q1_Annoyed variable which is -0.270997. With 95% confidence we can say that the difference of means between the Q1_annoyed variable and the two shows "The Biggest Loser" and "The Apprentice: Los Angeles" is between -0.52 and -0.01 points. The confidence interval does not contain 0 which means that The Apprentice has higher average annoyed variable. People generally on average think that this show is annoying in contrast to their annoyance to The Biggest Loser. The Observed difference is right in the middle so we see that the data in this data set is in line with what the average bootstrap mean is.

Part C.

"Dancing with the Stars" is a dancing competition between couples,being a celebrity paired with a professional. What proportion of American TV watchers would expect to give response of 4 or greater to Q2_Confusing question? Any response of 4 or 5 indicated that the survey participant either Agreed (4) or Strongly Agreed (5) that "Dancing with the Stars" was a confusing show.'



Based on the data our observed proportion is 0.0773481. After a bootstrap sampling distribution of this data we are 95% confident that the proportion of the amount of confused ratings that were 4 or more on Dancing with the Stars is between 0.04 and 0.12 points. Zero is not included in this interval which alludes to the fact that there is statistically significant evidence of there being a lot of confusion surrounding the show Dancing with the Stars. The interval stays low though not reaching past .2 so it is only moderately significant evidence for this.

Problem 4 - Ebay

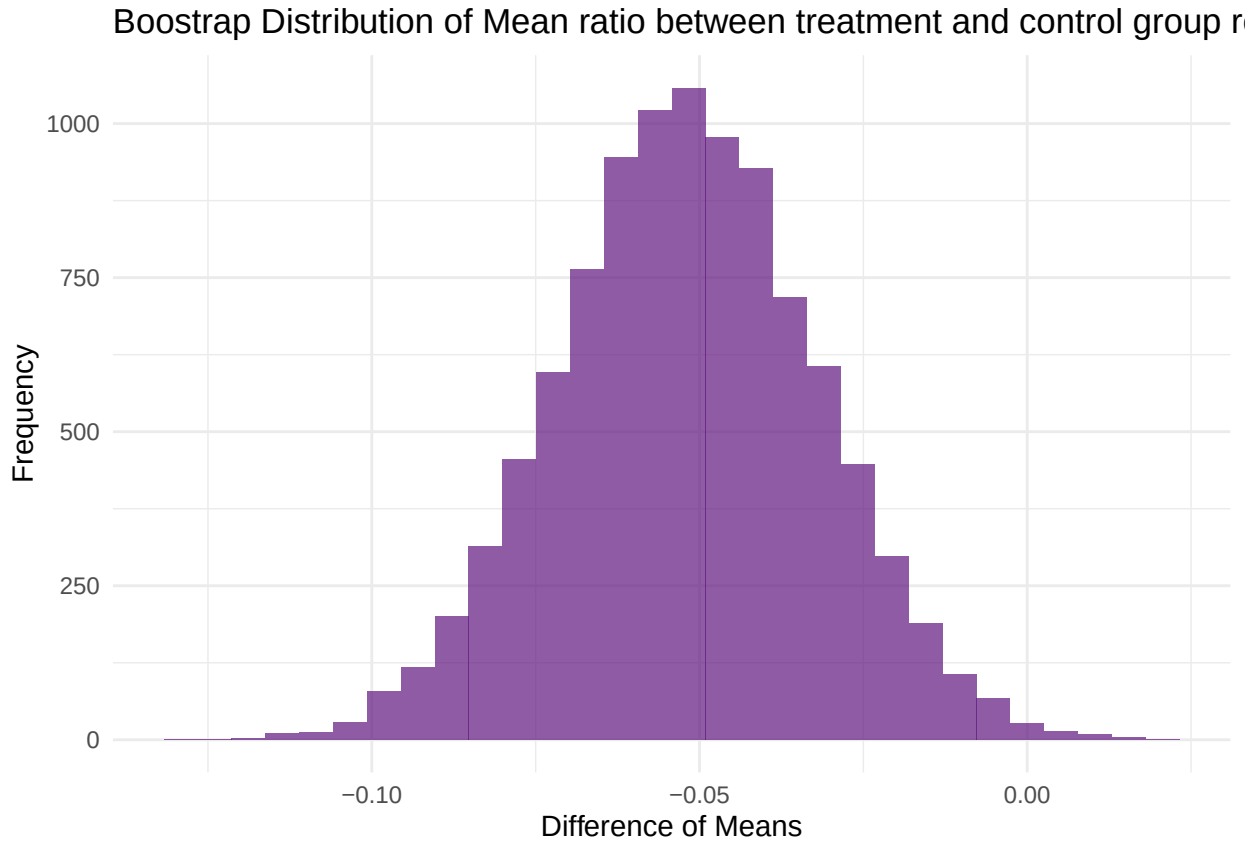
The data collected by Ebay in order to assess whether the companys paid advertising on Google search platform was improving ebays revenue In the experiment, E Bay randomly assigned each of the 210 DMAs to one of two groups:

• the treatment group, where advertising on Google AdWords for the whole DMA was paused for a month, starting on May 22. • the control group, where advertising on Google AdWords continued as before.

In ebay.csv you have the results of the experiment. The columns in this data set are:

• DMA: the name of the designated market area, e.g. New York • rank: the rank of that DMA by population • tv_homes: the number of homes in that DMA with a television, as measured by the market research firm Nielsen (who defined the DMAs in the first place) • adwords_pause: a 0/1 indicator, where 1 means that DMA was in the treatment group, and 0 means that DMA was in the control group. • rev_before: EBay's revenue in dollars from that DMA in the 30 days before May 22, before the experiment started. • rev_after: EBay's revenue in dollars from that DMA in the 30 days beginning on May 22, after the experiment started.

Is the revenue ratio the same in the treatment and control groups? Or does the data favor the idea that paid search advertising on Google creates extra revenue for Ebay?



Above we are using boostrap sampling distribution of the difference of means between the revenue ratio for treatment group and the control groups. The above graph tells us that with 95% confidence the difference of revenue ratio means for the treatment group minus the control group is between $-0.09 and $-0.01. Zero is not in this interval so it is suggested there is evidence that the effect of Ad Words is weaker afterwards. Because of this we have to say that there is evidence that AdWords is not contributing to extra revenue.