Homework 4

Juhi Malwade - jm97555

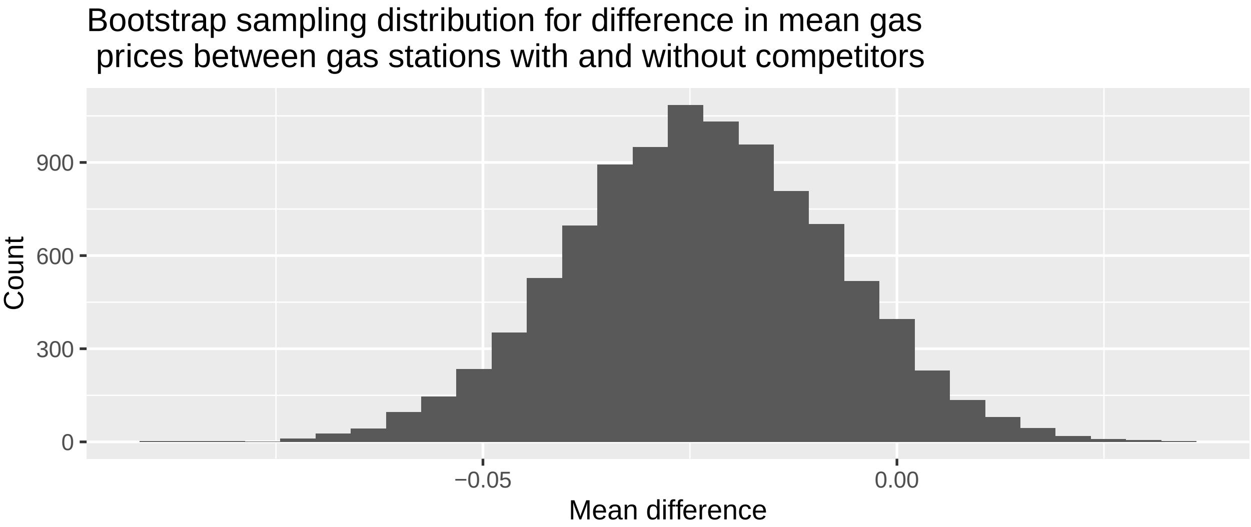
2/18/2024

Github Link

Problem 1: Gas Prices

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

Evidence:



name	lower	upper	level	method	estimate
diffmean	-0.0553252	0.0077236	0.95	percentile	-0.0234824

The histogram above represents the bootstrap sampling distribution for the difference in mean gas prices between gas stations with competitors and no competitors. After calculating a confidence interval with a level of 0.95, I found that the difference in price between gas stations with and without competitors is somewhere between $-0.06 and $0.01, with 95% confidence.

Conclusion:

Because the 95% confidence interval contains 0, indicating that there are both positive and negative differences in gas prices within the confidence interval, the theory that gas stations charge more if they lack direct competition in sight is not supported.

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

Evidence:

name	lower	upper	level	method	estimate
Intercept	1.7593161	1.8296623	0.95	percentile	1.7844702
Income	0.0000007	0.0000018	0.95	percentile	0.0000016
sigma	0.0638783	0.0848692	0.95	percentile	0.0688935
r.squared	0.0401482	0.3219138	0.95	percentile	0.3179953
F	4.1409248	46.9991458	0.95	percentile	46.1602957

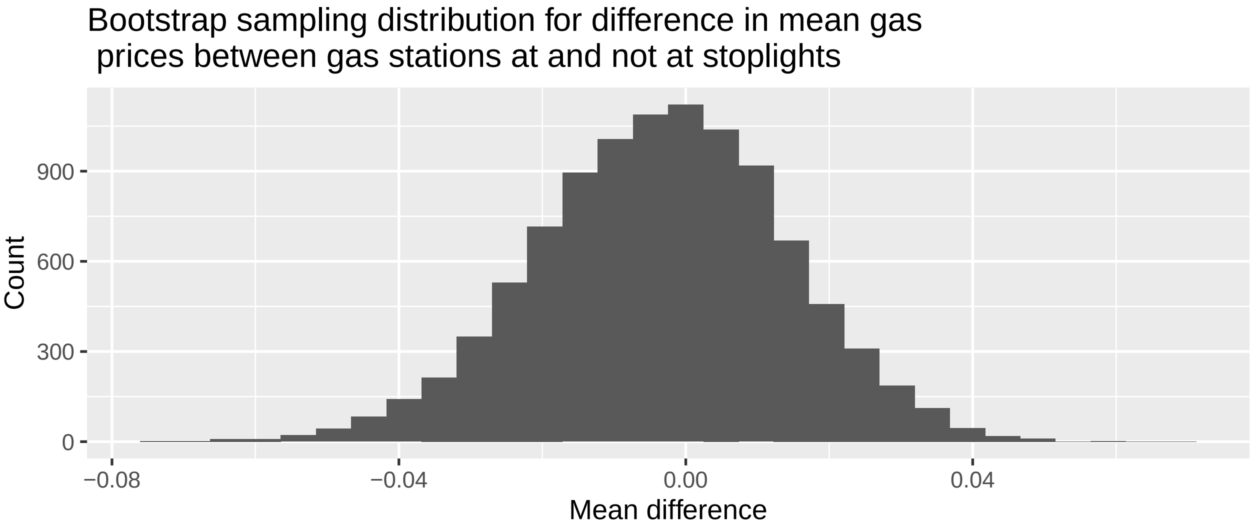
To collect evidence for the claim, I created 10,000 bootstrap samples from the original dataset. For each bootstrap sample, I refit a linear regression model for Income versus Gas Price. Based on the 95% confidence interval shown above, we are 95% confident the slope of Income versus Gas Price is between $\$6.7020173 \times 10^{-7}$ and $\$1.7796629 \times 10^{-6}$ - or equivalently, the gas price increases between $0.07 and $0.18 for every $100,000 increase in income with 95% confidence.

Conclusion:

Because the confidence interval is positive and does not contain zero, there is a statistically significant relationship between how rich an area is and gas prices. Thus, the theory that "the richer the area, the higher the gas prices" is supported. However, because the increase in gas prices with the increase in income is very small, this theory generally does not matter that much in the real world.

Claim 3: Gas stations at stoplights charge more.

Evidence:



name	lower	upper	level	method	estimate
diffmean	-0.0384934	0.0298429	0.95	percentile	-0.0032999

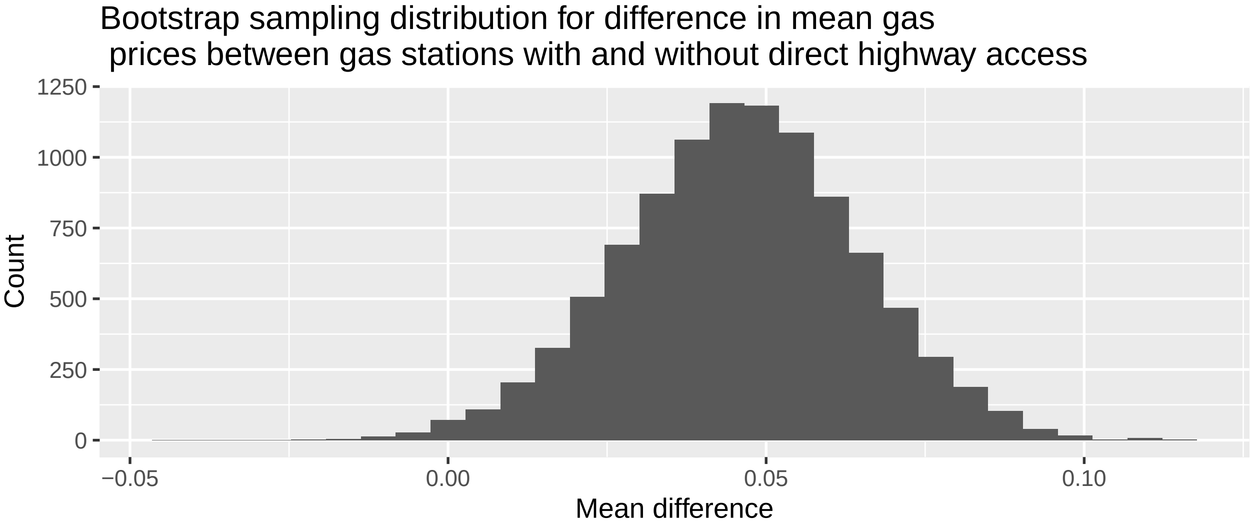
To collect evidence for the claim, I created 10,000 bootstrap samples from the original dataset. For each bootstrap sample, I calculated the mean difference in gas price for gas stations at stoplights and not at stoplights. The histogram above represents the bootstrap sampling distribution. Based on the 95% confidence interval, the price difference between gas stations with and without stoplights in front is somewhere between $-0.04 and $0.03, with 95% confidence.

Conclusion:

Because the 95% confidence interval contains 0, indicating that there are both positive and negative differences in gas prices within the confidence interval, the theory that gas stations charge more if they are at a stoplight is not supported.

Claim 4: Gas stations with direct highway access charge more

Evidence:



name	lower	upper	level	method	estimate
diffmean	0.0093039	0.082119	0.95	percentile	0.0456962

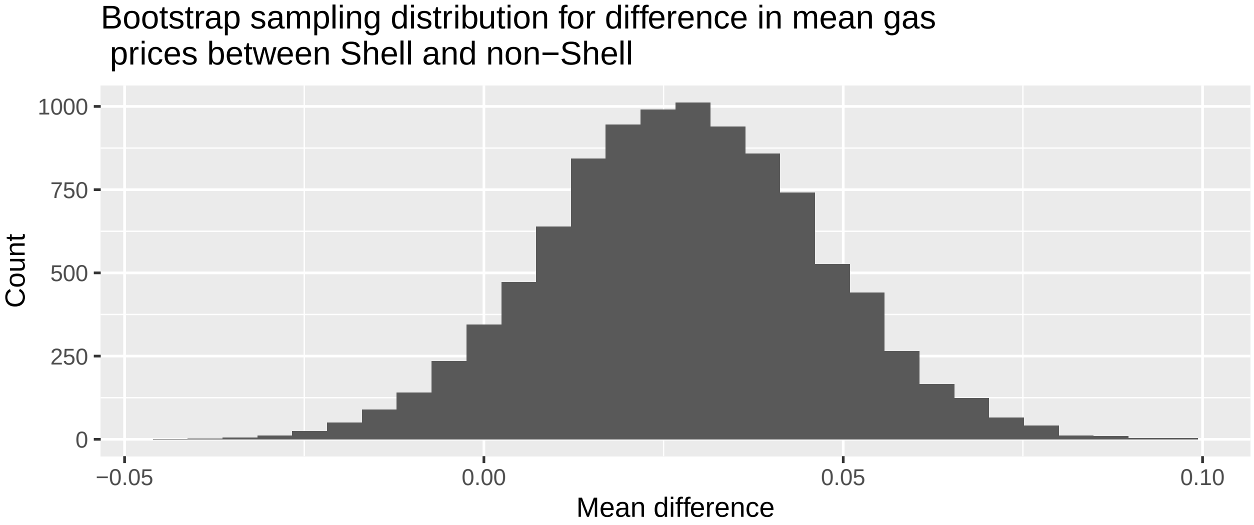
To collect evidence for the claim, I created 10,000 bootstrap samples from the original dataset. For each bootstrap sample, I calculated the mean difference in gas price for gas stations near highways and not near highways. The histogram above represents the bootstrap sampling distribution. Based on the 95% confidence interval, the price difference between gas stations with direct highway access and no direct highway access is somewhere between $0.01 and $0.08, with 95% confidence.

Conclusion:

Because the 95% confidence interval only contains positive values, there is a statistically significant difference between gas prices for gas stations with and without direct highway access. Thus, the theory that "gas stations with direct highway access charge more" is supported.

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

Evidence:



name	lower	upper	level	method	estimate
diffmean	-0.0094409	0.0656739	0.95	percentile	0.0274042

To collect evidence for the claim, I created a new column "Shell" that categorized the dataset between Shell and non-Shell gas stations. Then, I created 10,000 bootstrap samples from this dataset. For each sample, I calculated the mean difference in gas price for Shell and non-Shell gas stations. The histogram represents the bootstrap sampling distribution. Based on the 95% confidence interval displayed above, the price difference between Shell and non-Shell gas stations is somewhere between $-0.01 and $0.07.

Conclusion:

Because the 95% confidence interval contains 0, indicating that there are both positive and negative differences in gas prices within the confidence interval, the theory that Shell charges more than all other non-Shell brands is not supported.

Problem 2: Mercedes S-Class

Part A:

name	lower	upper	level	method	estimate
mean	26306.11	31837.19	0.95	percentile	28997.34

Based on the confidence interval displayed above, the average mileage of 2011 S-Class 63 AMGs that were hitting the used-car market when this data was collected is somewhere between 26306.11 miles and 31837.19 miles with 95% confidence.

Part B:

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

Based on the confidence interval displayed above, the proportion of all 2014 S-Class 550s that were painted black is somewhere between 0.42 and 0.45 with 95% confidence.

Problem 3: NBC Pilot Survey

Part A:

Question: Is there evidence that "Living with Ed" or "My Name is Earl" consistently produces a higher mean Q1_Happy response among viewers?

Approach: I first filtered the data set to only include ratings of the shows Living with Ed and My Name is Earl. Then, I created 10,000 bootstrap samples from this dataset. For each sample, I calculated the mean difference in Q1_Happy responses for My Name is Earl and Living with Ed using "diffmean". Lastly, I constructed a 95% confidence interval for the difference in mean viewer response to the Q1_Happy question for these two shows.

name	lower	upper	level	method	estimate
diffmean	-0.400424	0.0994183	0.95	percentile	-0.1490515

Results: According to the confidence interval displayed above, the mean difference in Q1_Happy responses for My Name is Earl and Living with Ed is between -0.4 and 0.1, with 95% confidence.

Conclusion: Because we are 95% confident the mean difference in ratings is between -0.4 and 0.1, we cannot conclude with statistical significance that one show produced higher mean Q1_Happy ratings than the other. This is because the confidence interval contains both positive and negative mean differences.

Part B:

Question: Is there evidence that "The Biggest Loser" or "The Apprentice: Los Angeles" consistently produces a higher mean Q1_Annoyed response among viewers?

Approach: I first filtered the data set to only include ratings of the shows The Biggest Loser and The Apprentice: Los Angeles. Then, I created 10,000 bootstrap samples from this dataset. For each sample, I calculated the mean difference in Q1_Annoyed responses for The Biggest Loser and The Apprentice: Los Angeles using "diffmean". Lastly, I constructed a 95% confidence interval for the difference in mean viewer response to the Q1_Annoyed question for these two shows.

name	lower	upper	level	method	estimate
diffmean	-0.5261479	-0.0218374	0.95	percentile	-0.270997

Results: According to the confidence interval displayed above, the mean difference in Q1_Annoyed responses for The Biggest Loser and The Apprentice: Los Angeles is between -0.52 and -0.02, with 95% confidence.

Conclusion: We are 95% confident the mean difference in ratings is between -0.52 and -0.02. Thus, because the interval only contains negative mean differences in ratings, we can conclude with 95% confidence that The Apprentice: Los Angeles consistently produces a higher mean Q1_Annoyed response compared to The Biggest Loser.

Part C:

Question: What proportion of American TV watchers would we expect to give a response of 4 or greater to the “Q2_Confusing” question for the show “Dancing with the Stars?”

Approach: I first filtered the data set to only include ratings of the show Dancing with the Stars and created a new variable “Confused” that was TRUE if Q2_Confusing was 4 or greater and FALSE if it was less than 4. Then, I created 10,000 bootstrap samples from this dataset. For each sample, I calculated the proportion of Confused values that were TRUE using “prop”. Lastly, I constructed a 95% confidence interval for the proportions of responses that were 4 or greater for the “Q2_Confusing” question.

name	lower	upper	level	method	estimate
prop_TRUE	0.038674	0.1160221	0.95	percentile	0.0773481

Results: According to the confidence interval displayed above, the proportion of responses to the “Q2_Confusing” that are 4 or greater is between 0.04 and 0.12, with 95% confidence.

Conclusion: We are 95% confident the proportion of responses to the “Q2_Confusing” that are 4 or greater is between 0.04 and 0.12. The proportion is estimated to be in the middle of this interval at 0.08.

Problem 4: Ebay

Question:

Does paid search advertising on Google create extra revenue for EBay?

Approach:

I first created 10,000 bootstrap samples from this dataset. For each sample, I calculated the mean difference in revenue ratio for the treatment group and the control group using “diffmean” - for the treatment group, advertising on Google AdWords for the DMA was paused for a month and for the control group, advertising on Google AdWords continued as before. Lastly, I constructed a 95% confidence interval for the difference in mean revenue ratio for the treatment and control group.

name	lower	upper	level	method	estimate
diffmean	-0.0910824	-0.0132482	0.95	percentile	-0.0522815

Results:

According to the confidence interval displayed above, the mean difference in revenue ratio for the treatment and control group is between -0.09 and -0.01, with 95% confidence.

Conclusion:

In conclusion, there is statistically significant evidence that paid search advertising on Google creates extra revenue for EBay. This is because we are 95% confident the the mean difference in revenue ratio for the treatment and control group is between -0.09 and -0.01. Because the confidence interval only includes negative mean differences, the treatment group (in which advertising was paused) has a systemically lower revenue ratio than the control group.