Homework 3

Arun Mahadevan Sathia Narayanan

2025-02-12

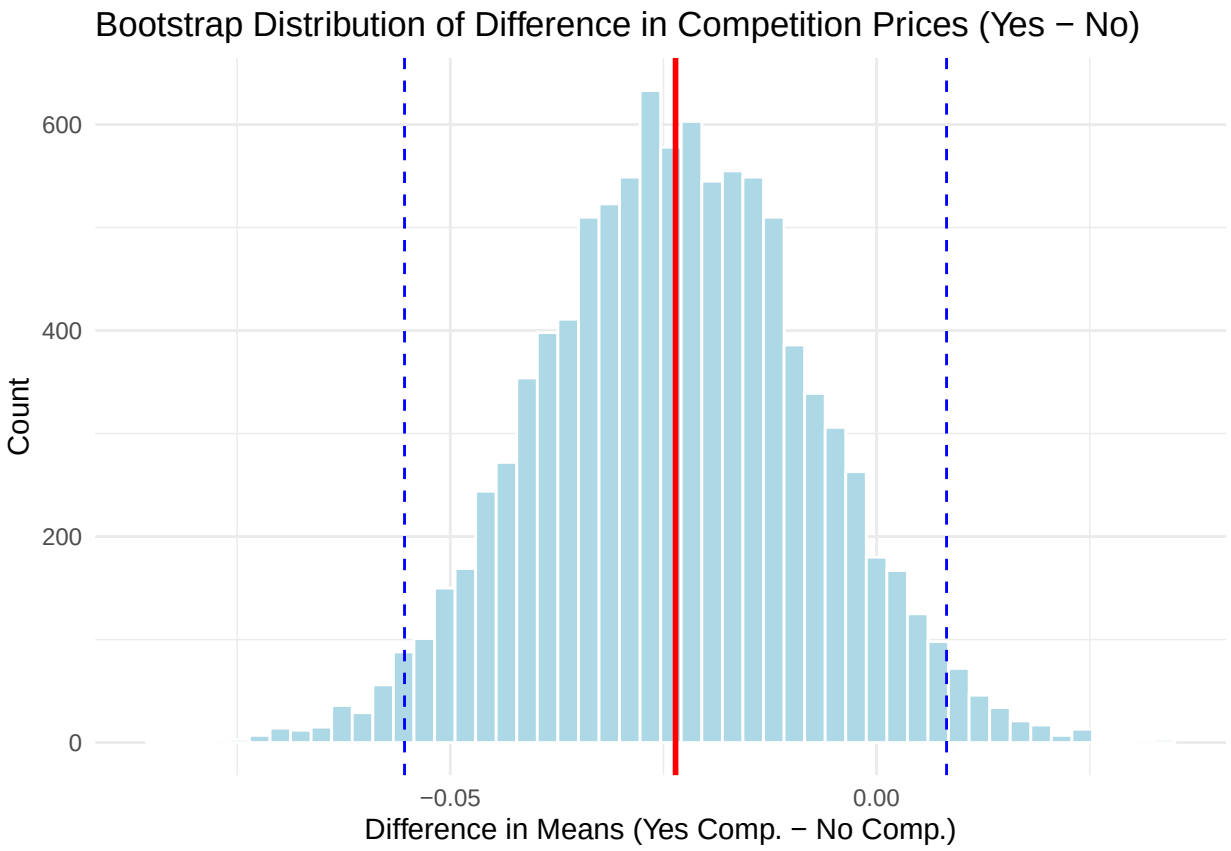
GitHub Link:
To GitHub

GitHub Link (Text Format):
https://github.com/arunmsn/SDS315/tree/main/Week_05/Homework_3

Problem 1

1 - Theory A -

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

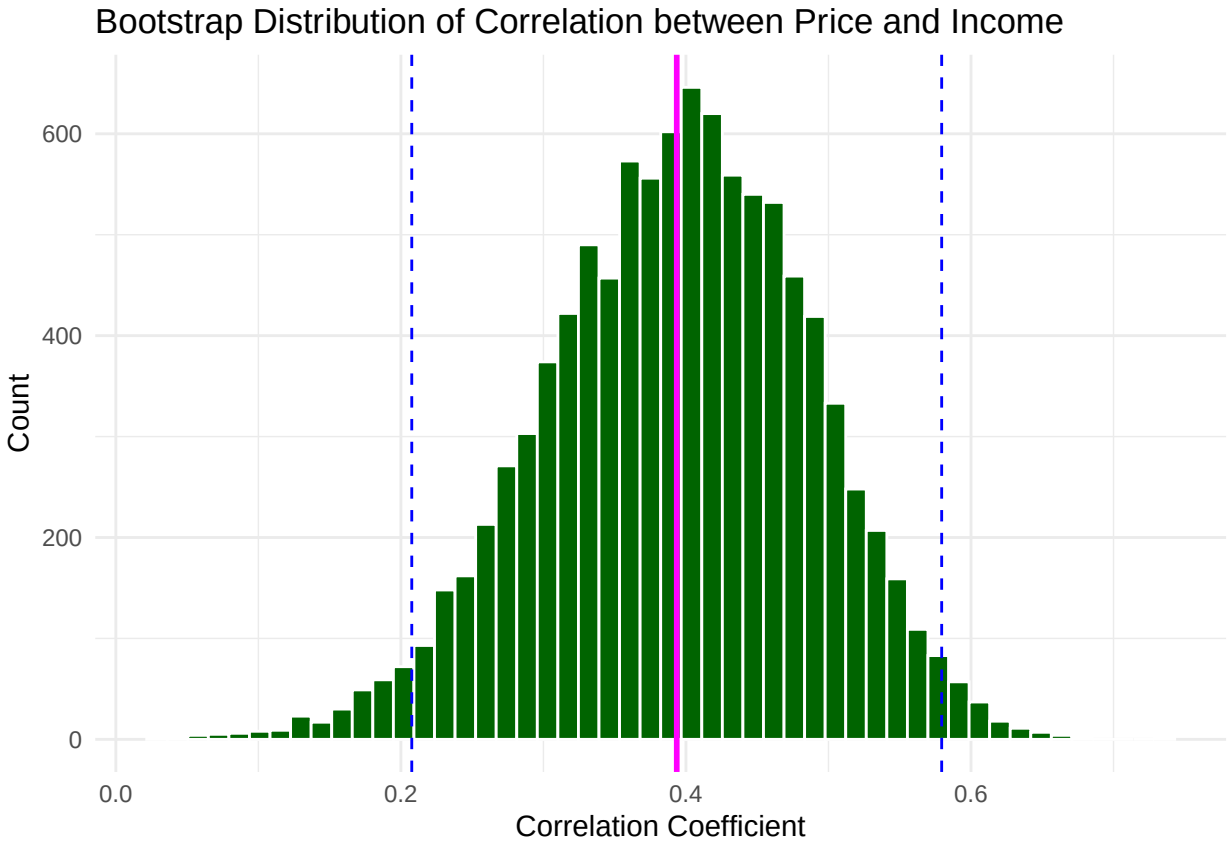



```
## Bootstrap Results:
## Mean difference: -0.02358237
## Standard deviation: 0.01589475
## 95% Confidence Interval: -0.05537188 to 0.008207136
```

The above distribution was made using 10000 bootstrap trials. The net results are the numbers we see below the distribution: the mean difference between the prices (this is Yes - No) is ~`-0.0233`, which indicates that a majority of the bootstrap trials had `No` having higher prices (specifically, around 2 cents higher). From the trials, the difference in gas price between stations with competitors and stations without competitors (Yes - No) will be anywhere between ~`-0.055` (5.5 cents lower) and ~`0.008` (0.8 cents higher), with 95% confidence.

1 - Theory B -

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




```
## Bootstrap Results:
```

```
## Mean correlation: 0.3935298
```

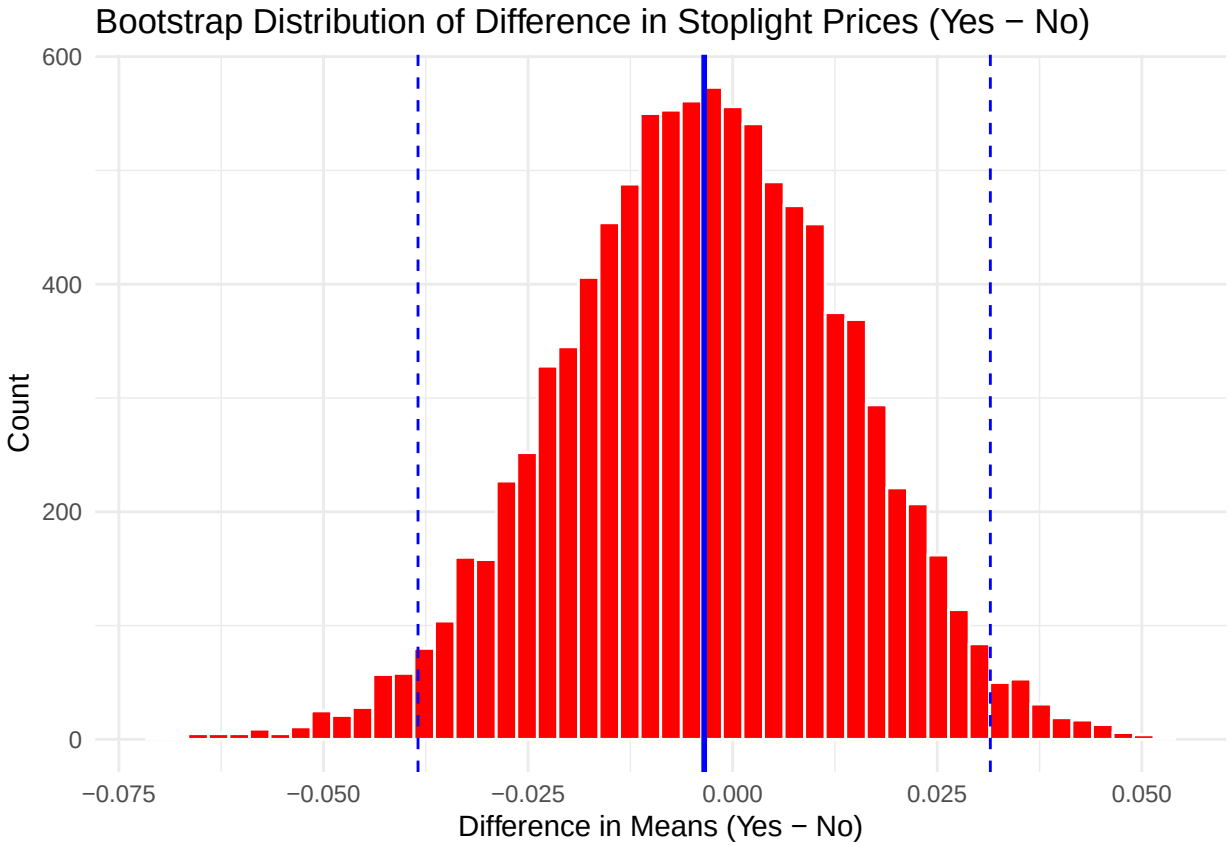
```
## Standard deviation: 0.09294905
```

```
## 95% Confidence Interval: 0.2076317 to 0.5794279
```

Similar to the previous graph, the distribution above is created using a bootstrapped sample 10000 times for the correlation between `Income` (that is, richness of an area) and `Price`. From the data, we found that the mean correlation was ~`0.393`, indicating a majority of the trials to be having a positive, moderately strong relationship. The correlation, with 95% confidence, could be anywhere from ~`0.203` to ~`0.583`. This goes to show that (once again, with 95% confidence), the richer the area, the higher the gas prices for that area will be.

1 - Theory C -

Claim: Gas stations at stoplights charge more.




```
## Bootstrap Results:
```

```
## Mean difference: -0.003480227
```

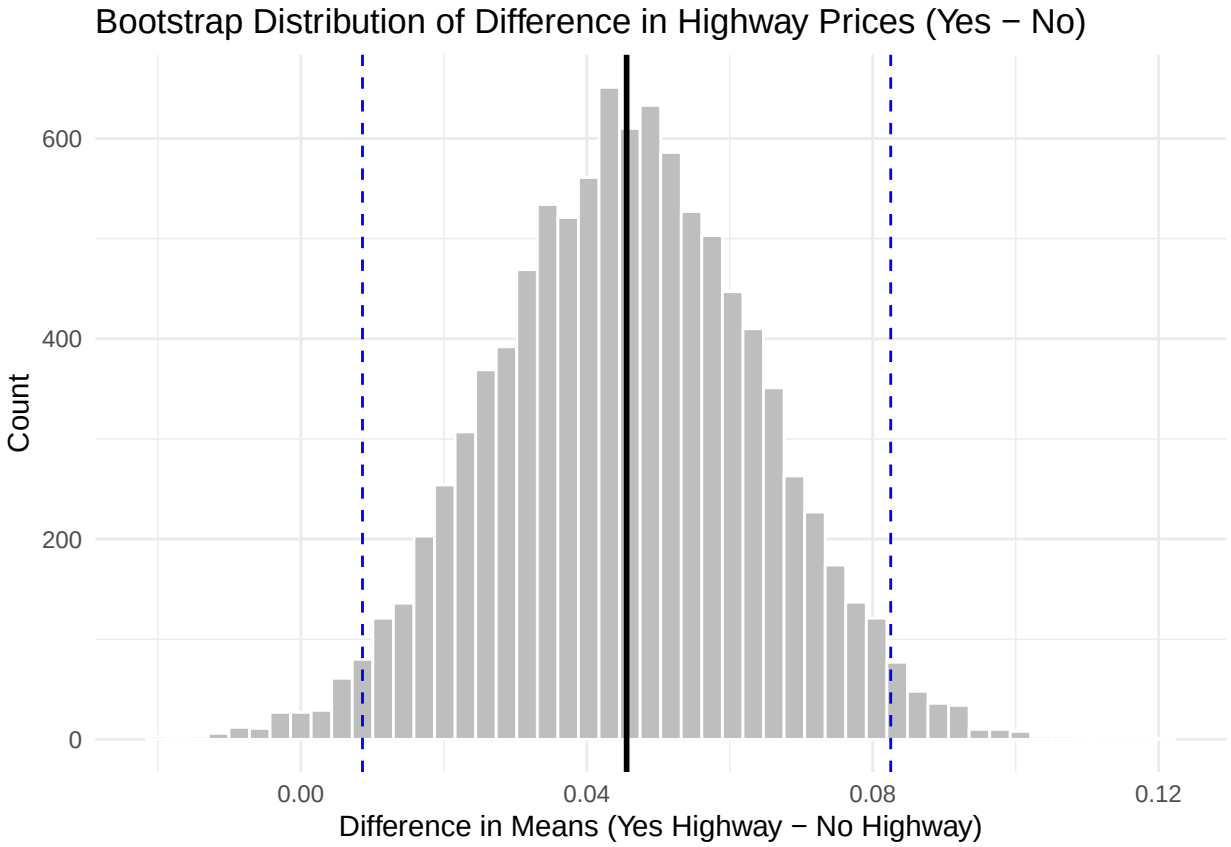
```
## Standard deviation: 0.01748406
```

```
## 95% Confidence Interval: -0.03844835 to 0.03148789
```

Using a bootstrapped sample 10000 times, the above distribution was created to show the mean gas prices at stoplights. The mean difference was ~`-0.003` (0.3 cents less), which means a majority of the trials showed that if a station was to be at a stoplight, they would charge less (however, this value is very close to zero). But I cannot say this with 100% confidence. I can say, with 95% confidence, that the range of values the differences in the mean prices could be is from ~`-0.038` (3.8 cents lower) to ~`0.03` (3 cents higher). Since the range contains both negative and postive values, the confidence in the theory falls.

1 - Theory D -

Claim: Gas stations with direct highway access charge more.




```
## Bootstrap Results:
```

```
## Mean difference: 0.04556885
```

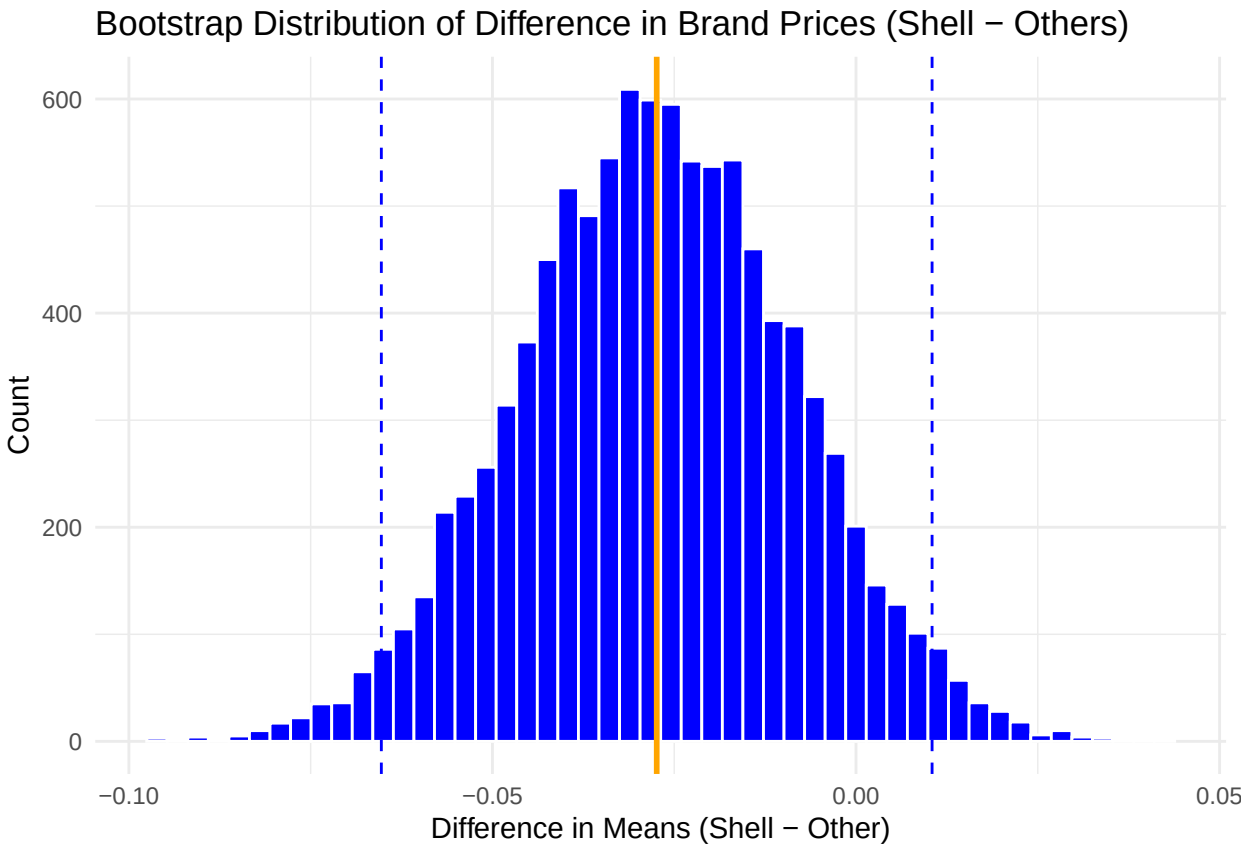
```
## Standard deviation: 0.01847498
```

```
## 95% Confidence Interval: 0.008618899 to 0.08251881
```

Using a bootstrapped sample 10000 times, the above distribution was created to show the mean gas prices dependent upon if they had highway access or not. The mean difference was ~`0.045` (which means a majority of the trials showed that gas stations with highway access charged an average of 4.5 cents compared to those without). This cannot be said with 100% confidence, however. I am 95% confident that the difference in means between stations with highway access compared to those without will be anywhere from ~`0.009` (0.9 cents) to ~`0.082` (8.2 cents). Which also means I am 95% confident that the stations with highway access will charge more than stations without highway access. This shows that there may be some merit for this theory.

1 - Theory E -

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




```
## Bootstrap Results:
## Mean difference: -0.02740872
## Standard deviation: 0.01893498
## 95% Confidence Interval: -0.06527868 to 0.01046123
```

Here, the Brand category was under review, where we wanted to compare if the Shell gas brand charges more than the other gas brands. From the 10000 bootstrapped trials, the mean difference between Shell and the other brands was ~`-0.027` (2.7 cents less). This means several of the trials showed the mean Shell price being cheaper than the mean of the other brands. This is not said with 100% confidence; rather I am 95% confident that the difference in mean prices between Shell and the other brands is anywhere from ~`-0.065` (6.5 cents cheaper) to ~`0.011` (1.1 cents more expensive). This disproves the confidence in the theory, as the range contains both negative and positive values.

Problem 2

2 - Part A

```
## Bootstrap Results:
## Mean: 29020.25
## Standard deviation: 1401.421
## 95% Confidence Interval: 26217.4 to 31823.09
```

The 95% confidence interval for the mileage is anywhere from `26183.85` miles to `31797.55` miles.

2 - Part B

```
## Bootstrap Results:
## Mean: 0.4347367
## Standard deviation: 0.009239012
## 95% Confidence Interval: 0.4162587 to 0.4532147
```

The 95% confidence interval for the proportion is anywhere from `0.416` to `0.453`.

Problem 3

3 - Part A

```
## Bootstrap Results:
```

```
## Observed Mean Difference (Ed - Earl): 0.1490515
```

```
## 95% Confidence Interval: -0.1024291 to 0.3949049
```

The 95% confidence interval of the mean difference between `Q1_Happy` for `"Living with Ed"` and `Q1_Happy` for `"My Name is Earl"` is ~`-0.103` to ~`0.394`. Since the interval contains both negative and positive values, there is no very strong evidence to show that one show consistently produces a high mean `Q1_Happy` response among viewers. However, the edge can go a little toward `"Living with Ed"`, since more of the 95% confidence interval is positive.

3 - Part B

```
## Bootstrap Results:
```

```
## Observed Mean Difference (Loser - App): -0.270997
```

```
## 95% Confidence Interval: -0.5178352 to -0.01680964
```

The 95% confidence interval of the mean difference between `Q1_Annoyed` for `"The Biggest Loser"` and `Q1_Annoyed` for `"The Apprentice: Los Angeles"` is ~`-0.518` to ~`-0.017`. Since the interval contains only negative values, there is strong evidence to show that one show (here, that would be `"The Apprentice: Los Angeles"`) consistently produces a high mean `Q1_Annoyed` response than the other (`"The Biggest Loser"`) among viewers.

3 - Part C

```
## Bootstrap Results:
```

```
## Observed Mean Difference (Loser - App): 0.07734807
```

```
## 95% Confidence Interval: 0.03867403 to 0.1160221
```

The 95% confidence interval of the proportion of viewers who gave the `Q2_Confusion` rating a 4 or higher is anywhere from `0.044` to `0.121`. Therefore, I am 95% confident that anywhere from 4% to 12% will report a `Q2_Confusion` of a 4 or higher. This is a very small proportion relative to the entire population of voters, so a majority (anywhere from 96% to 88%) will report a `Q2_Confusion` that is 3 or lower.

Problem 4

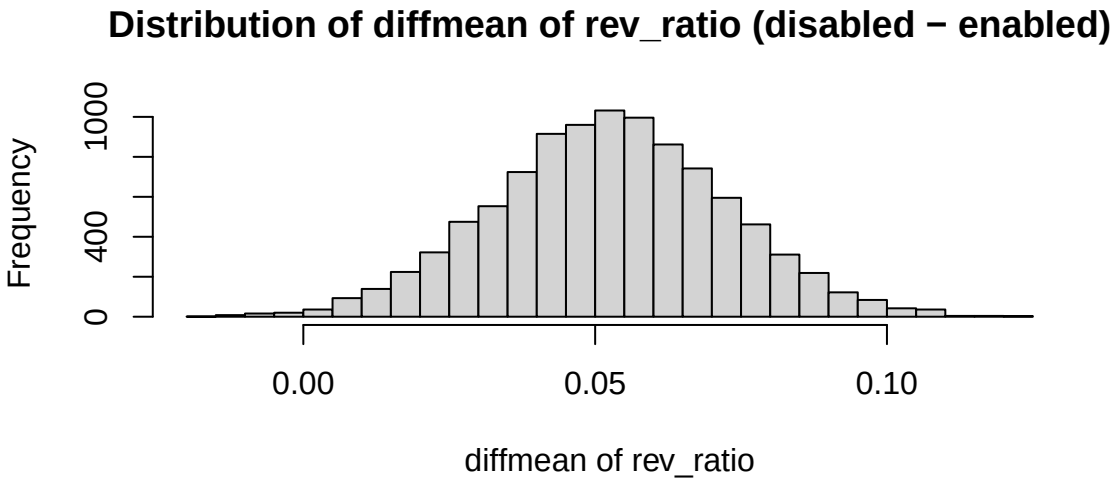
4 - Question

How did the revenue ratio for the designated DMAs without paid search compare to the revenue ratio for the rest of the DMAs with paid search?

4 - Approach

The first step is to mutate a column to include revenue ratio per DMA (`rev_ratio` is the mutated column). The second step is to split the group into if the paid search was paused or not, making it easier for future processes. The third step is to bootstrap on the individual populations (bootstrap for paid search, bootstrap for non-paid search). Then we would get the difference in the rev_ratio between the two populations, and that would be one data point. I am doing this by getting the mean rev_ratio per population, then finding the diffmean (difference in means) between the populations. The fourth step is to repeat this process for a total of 10,000 times, and then plot the distribution of differences in rev_ratio.
The fifth step is to look at the results, which is next.

4 - Results




```
## 95% Confidence Interval: 0.01298506 to 0.09129611
```

Here, we can see the distribution of the differences in mean rev_ratios between the DMAs without paid search and the DMAs with paid search. The values in the 95% confidence interval shows a range of ~`0.013` to ~`0.091`.

4 - Conclusion

As we the saw above, the 95% confidence interval has a range of ~`0.013` to ~`0.091`. This shows that the differences in the rev_ratios between the paid search paused and the paid search still enabled was positive for 95% of the distribution (i.e. the confidence interval never included negative values), and thus a majority of the results actually showed that turning off the paid search led to an increase in revenue by a slight margin. For the stakeholders who would want to enable a paid search (i.e. the ad for the product would be sponsored), this data actually shows that for larger companies like `eBay`, it might not actually be necessary for the ad to be paid for. I am 95% confident that the paid search feature for eBay products (and other larger companies) can be turned off.