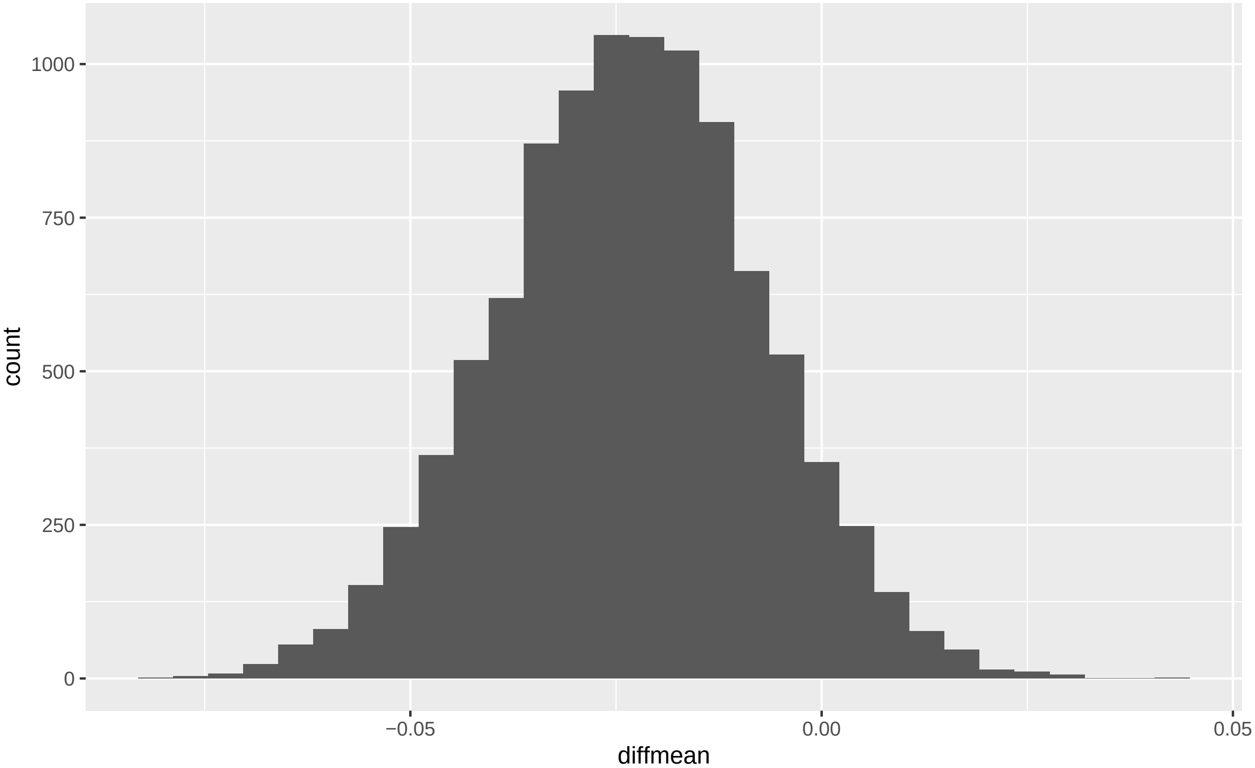
SDS HW 4

Gianluca Bollo (gb25625) - https://github.com/gianlucabollo/HW4-SDS315

2/20/2024

Austin Gas Prices Analysis

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

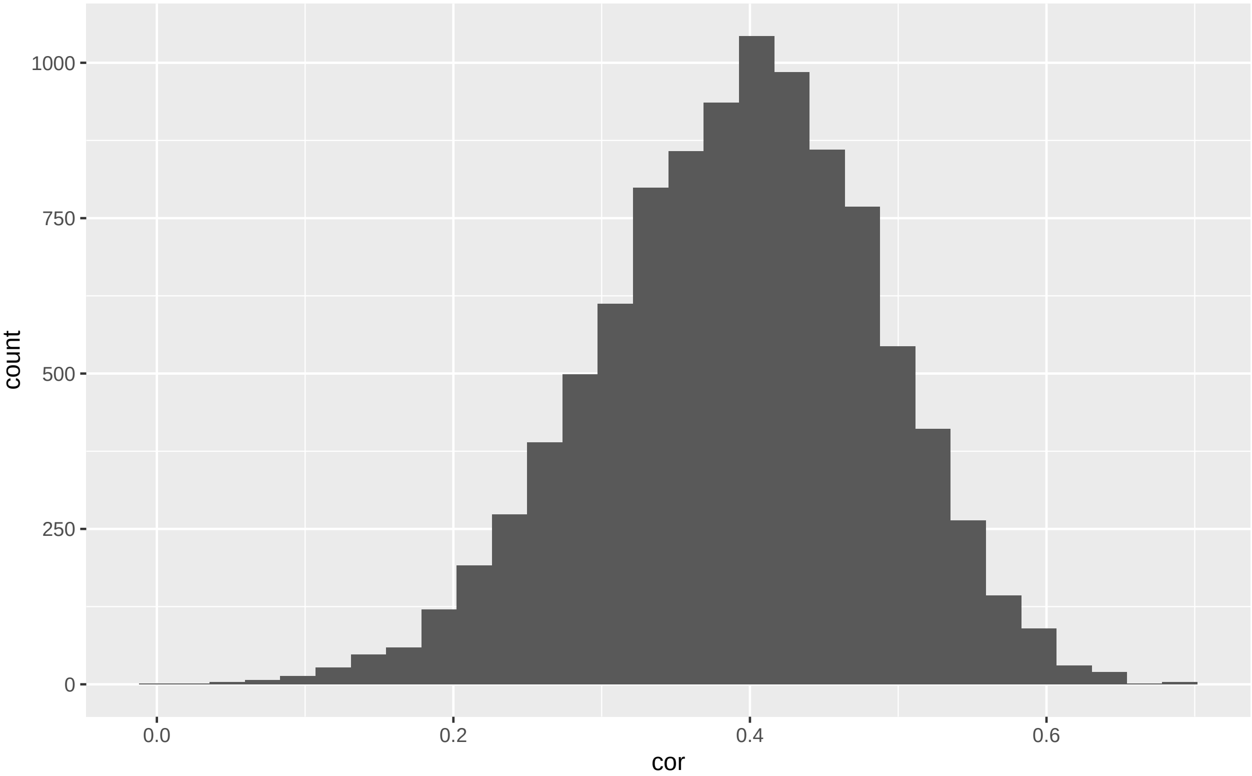


```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05483886 0.007575794  0.95 percentile -0.02348235
```

Evidence: The difference in price between gas stations with and without competition in sight is somewhere between -0.055 and 0.007, with 95% confidence.

Conclusion: Due to this confidence interval containing 0 (not statistically significant) and the fact that the interval itself contains all relatively small values, the claim that "Gas stations charge more if they lack direct competition in sight" is not supported by the data.

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

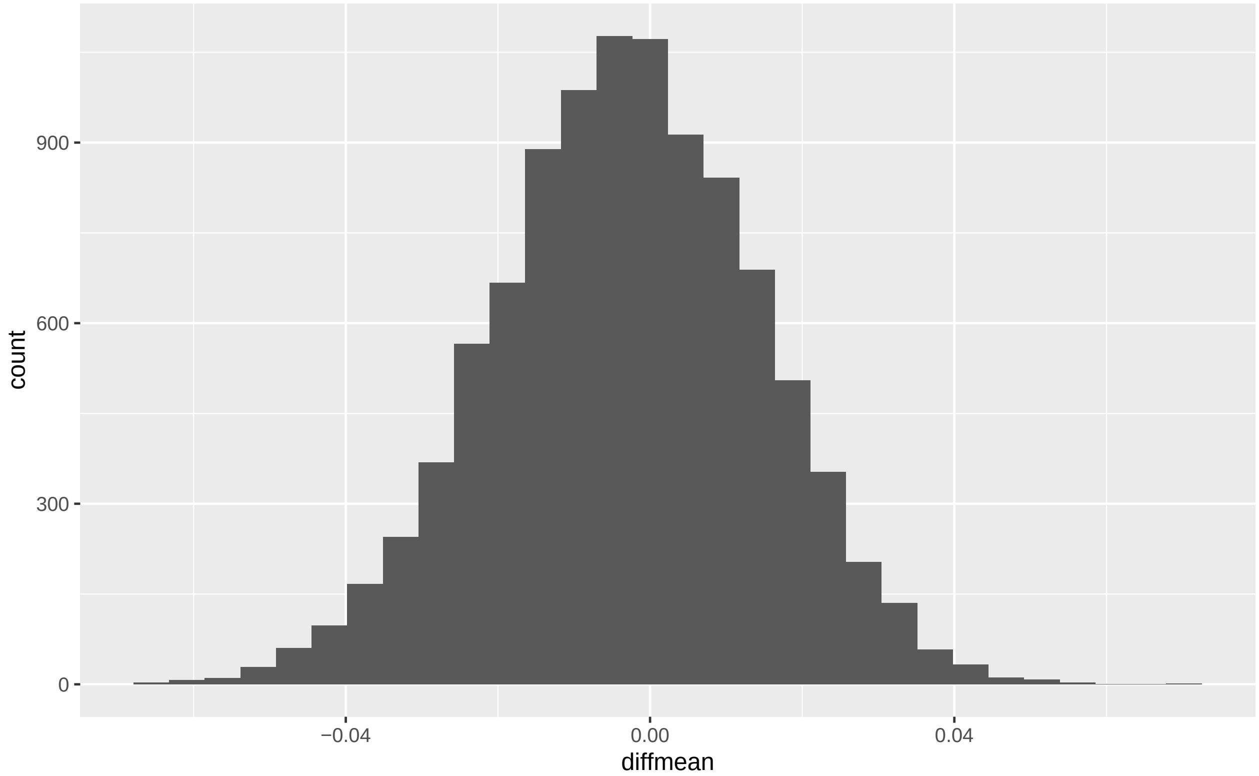


```
##   name     lower     upper level     method  estimate
## 1  cor 0.1981271 0.5636909  0.95 percentile 0.3961546
```

Evidence: The correlation coefficient between gas price and median household income of a certain area is somewhere between 0.197 and 0.564, with 95% confidence.

Conclusion: Due to this confidence interval not containing 0 (statistically significant) and the fact that the interval represents a weak to moderate correlation, the claim that "The richer the area, the higher the gas prices" is slightly supported by the data, in terms of correlation.

Claim C: Gas stations at stoplights charge more.

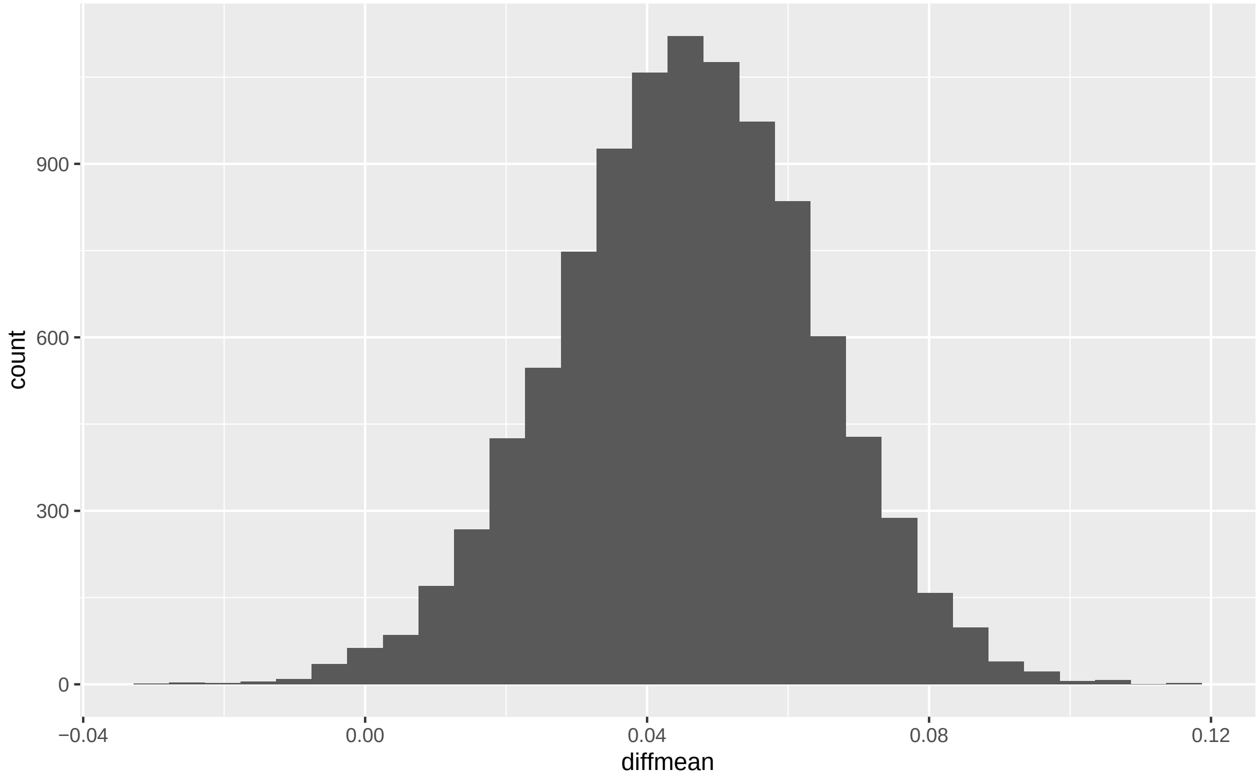


```
##       name       lower     upper level     method     estimate
## 1 diffmean -0.03783605 0.0303702  0.95 percentile -0.003299916
```

Evidence: The difference in price between gas stations with and without stoplights in front of them is somewhere between -0.039 and 0.030, with 95% confidence.

Conclusion: Due to this confidence interval containing 0 (not statistically significant) and the fact that the interval itself contains all relatively small values, the claim that “Gas stations at stoplights charge more” is not supported by the data.

Claim D: Gas stations with direct highway access charge more.

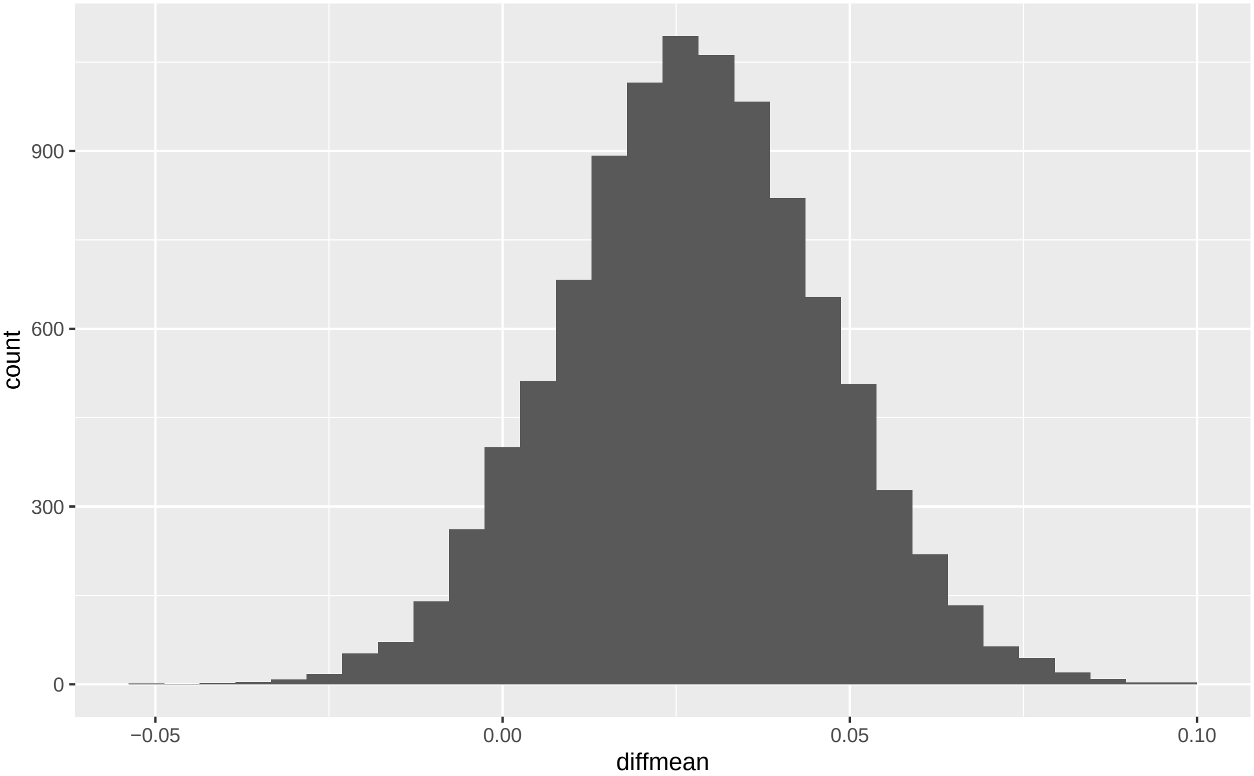


```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.009187102 0.08028171  0.95 percentile 0.0456962
```

Evidence: The difference in price between gas stations with and without direct highway access is somewhere between 0.009 and 0.081, with 95% confidence.

Conclusion: Although this confidence interval does not contain 0 (statistically significant), the mean difference of price, according to the interval, is very small. So practically speaking, the claim that "Gas stations at stoplights charge more" is not supported by the data.

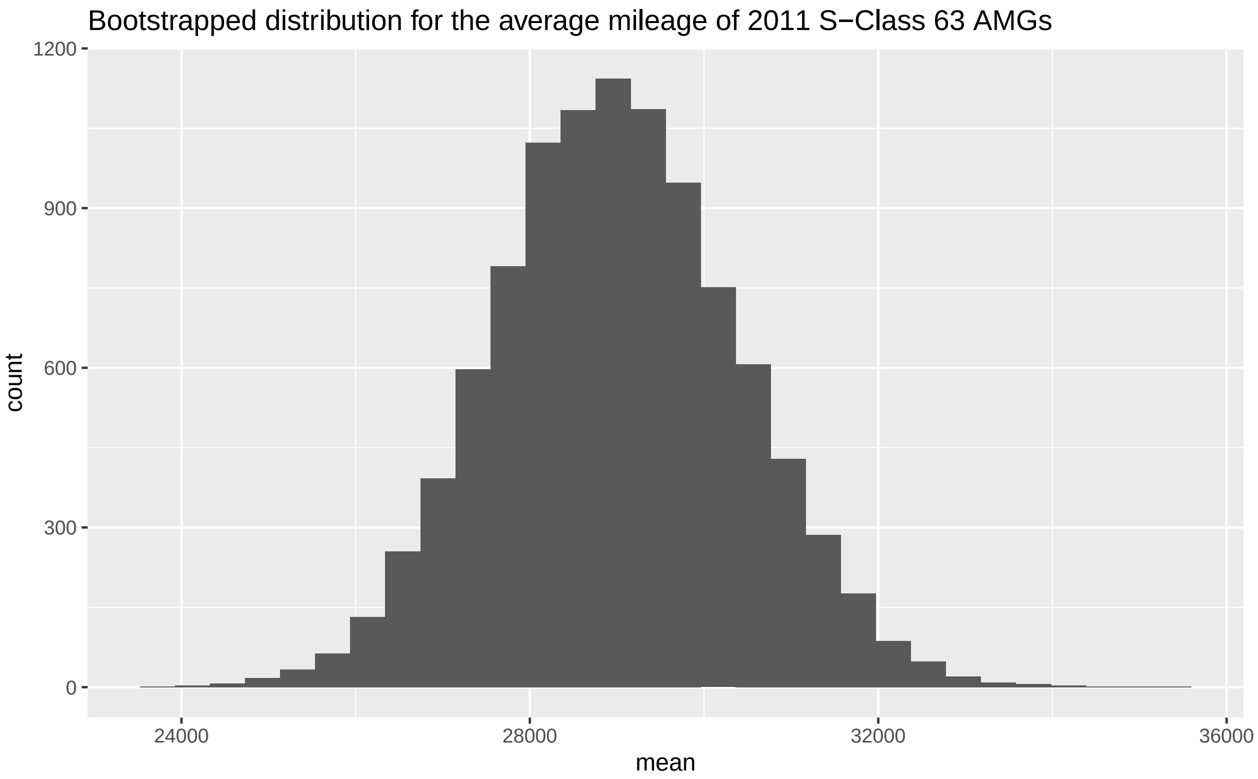
Claim E: Shell charges more than all other non-Shell brands.



```
##     name       lower      upper level     method   estimate
## 1 result -0.00910433 0.06499558  0.95 percentile 0.02740421
```

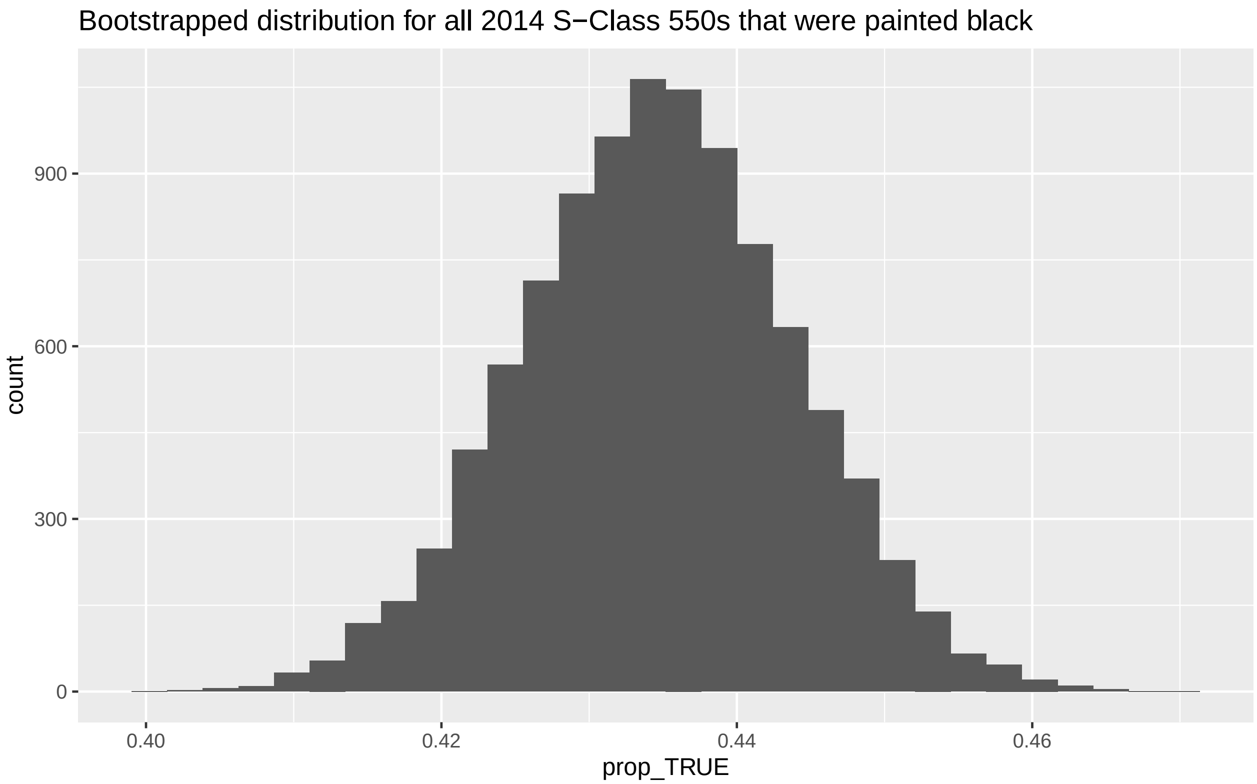
Evidence: The difference in price between shell gas stations and all other non-Shell brands is somewhere between -0.011 and 0.065, with 95% confidence.

Conclusion: Due to this confidence interval containing 0 (not statistically significant) and the fact that the interval itself contains all relatively small values, the claim that “Shell charges more than all other non-Shell brands” is not supported by the data.



```
##   name    lower    upper level     method estimate
## 1 mean 26324.71 31759.54  0.95 percentile 28997.34
```

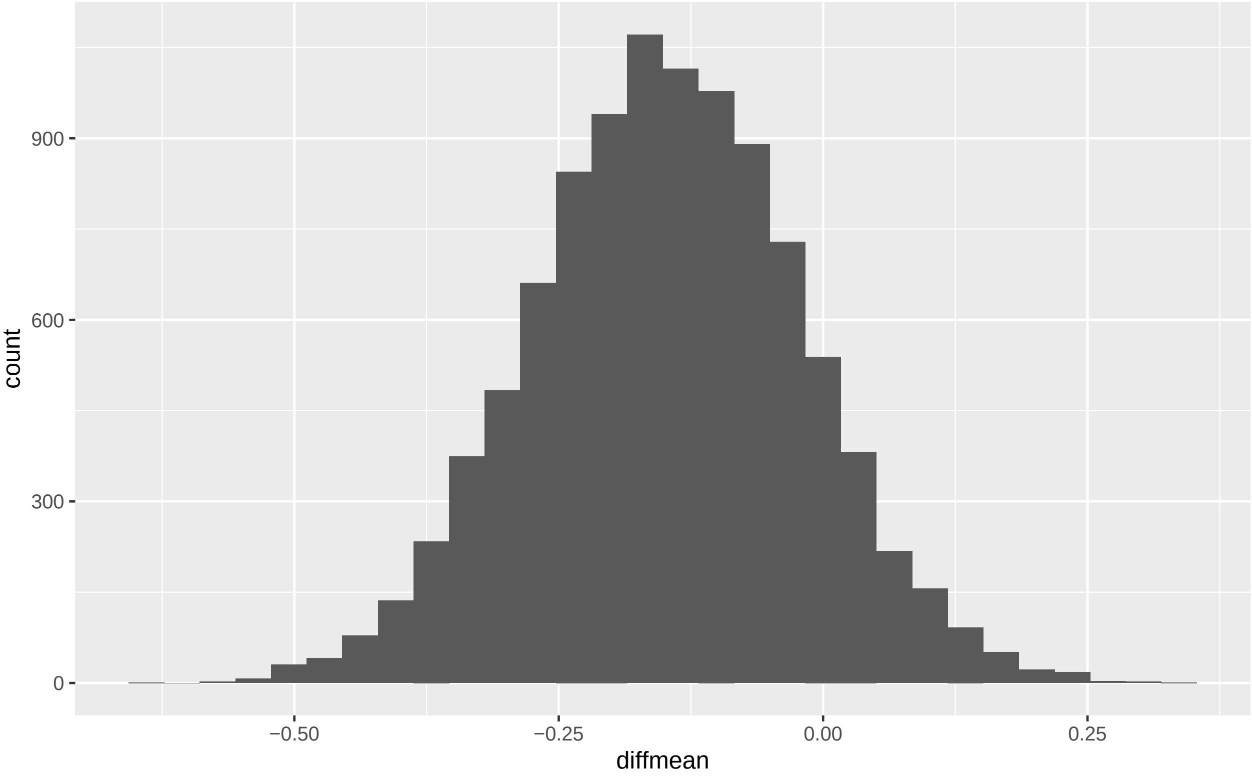
The average mileage of 2011 S-Class 63 AMGs that were hitting the used-car market when this data was collected is somewhere between 26245.82 and 31842 miles, with 95% confidence.



```
##        name     lower     upper level     method  estimate
## 1 prop_TRUE 0.4164071 0.4527518  0.95 percentile 0.4347525
```

The the proportion of all 2014 S-Class 550s that were painted black is somewhere between .417 and .453, with 95% confidence.

Is there evidence that one show consistently produces a higher mean Q1_Happy response among viewers?



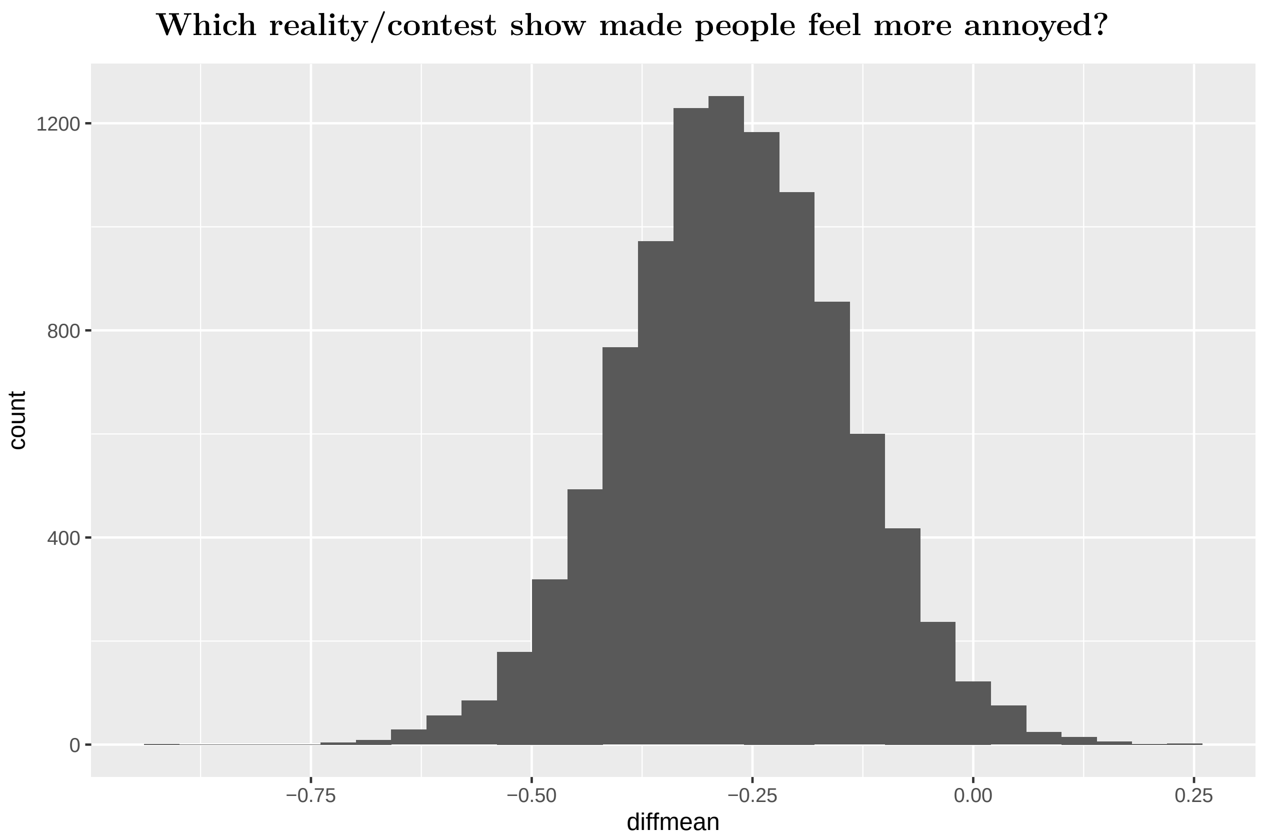
```
##       name      lower     upper level     method   estimate
## 1 diffmean -0.3976481 0.1028095  0.95 percentile -0.1490515
```

```
## # A tibble: 2 x 2
##   Show             mean
##   <chr>           <dbl>
## 1 Living with Ed   3.93
## 2 My Name is Earl  3.78
```

Approach: To answer this question, I filtered the nbc show data set to only data describing "Living with Ed" and "My Name is Earl". I then collected 10,000 bootstrapped samples and calculated the difference in means between the shows for the Q1_Happy variable for each sample. Finally, I constructed a histogram and confidence interval (using confint function) of all the bootstrapped samples.

Results: Evidence that this approach provided was a 95% confidence interval that estimates the population value of difference in mean Q1_Happy response between both shows. According to the data, the difference in mean Q1_Happy responses between these two shows is somewhere between -0.392 and 0.098 with 95% confidence.

Conclusion: Due to the fact that this confidence interval contains 0 (not statistically significant) and the estimated population difference is only -0.149, the claim that one show makes viewers happier over the other is not strongly supported by the data. With that said, there is no evidence that one show consistently produces a higher mean Q1_Happy response among viewers.



```
##       name      lower       upper level     method  estimate
## 1 diffmean -0.5169871 -0.02070027  0.95 percentile -0.270997
```

```
## # A tibble: 2 x 2
##   Show                        mean
##   <chr>                      <dbl>
## 1 The Apprentice: Los Angeles  2.31
## 2 The Biggest Loser            2.04
```

Approach: To answer this question, I filtered the nbc show data set to only data describing "The Biggest Loser" and "The Apprentice: Los Angeles". I then collected 10,000 bootstrapped samples and calculated the difference in means between the shows for the Q1_Annoyed variable for each sample. Finally, I constructed a histogram and confidence interval (using confint function) of all the bootstrapped samples.

Results: Evidence that this approach provided was a 95% confidence interval that estimates the population value for the difference in mean Q1_Annoyed response between both shows. According to the data, the difference in mean Q1_Annoyed responses between these two shows is somewhere between -0.512 and -0.020, with 95% confidence.

Conclusion: Due to the fact that this confidence interval does not contain 0 (statistically significant) and the lower bound of the mean is more than half a point, the result proves to be practically significant as well. The data-supported answer to the question "Which reality/contest show made people feel more annoyed?" is "The Apprentice: Los Angeles".

What proportion of American TV watchers would we expect to give a response of 4 or greater to the "Q2_Confusing" question?

```
##        name      lower     upper level     method   estimate
## 1 prop_TRUE 0.03867403 0.1160221  0.95 percentile 0.07734807
```

Approach: To answer this question, I filtered the nbc show data set to only data describing "Dancing With the Stars" . I then collected 10,000 bootstrapped samples and calculated the proportion of watchers that gave a response of 4 or greater to the "Q2_Confusing" question for each sample. Finally, I constructed a histogram and confidence interval (using confint function) of all the bootstrapped samples.

Results: Evidence that this approach provided was a 95% confidence interval that estimates the population value for the proportion of DWTS watchers that gave a response of 4 or greater to the "Q2_Confusing" question. According to the data, this proportion is somewhere between 0.039 and 0.116, with 95% confidence.

Conclusion: The data-supported answer to the question "What proportion of American TV watchers would we expect to give a response of 4 or greater to the"Q2_Confusing" question?" is estimated to be somewhere between .039 and 0.116, with 95% confidence. The mean value of all bootstrapped samples (p hat) is .077.

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

```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.09096022 -0.01362019  0.95 percentile -0.05228145
```

Approach: To answer this question, I collected 10,000 bootstrapped samples from the ebay.csv dataset and calculated the mean difference of revenue ratio between the control and treatment groups. Finally, I constructed a histogram and confidence interval (using confint function) of all the bootstrapped samples.

Results: Evidence that this approach provided was a 95% confidence interval that estimates the population value for the mean difference of revenue ratio between the adwords enabled / disabled groups. According to the data, this value is somewhere between -0.091 and -0.013, with 95% confidence.

Conclusion: Statistically speaking, because the confidence interval does not contain 0, the findings are statistically significant. Taking a more practical approach to make a conclusion, however, would be to inspect just how big this mean difference is, relative to the total revenue values. Differences within this interval, when looking at revenue in the tens of millions, could be worth hundreds of thousands to even millions of dollars. So in practical terms as well, paid search advertising on Google does create extra revenue for EBay. The estimated value from all bootstrapped samples is -0.052.