Homework 4 SDS315

Lucas Chiang

2024-02-17

Github Link:

Problem 1

Theory A

Claim

The claim here is that gas stations charge more if they lack direct competition in sight.

Evidence

Mean Comparison for Gas Prices on Whether Gas Stations Have Direct Competition in Sight

Competitors?	Mean Price
N	1.876
Y	1.852

"N" means no, they don't have direct competitors, and "Y" means yes, they do have direct competition.

Mean Difference: -.0234

Above, we see mean prices from the data-set that gas stations charge for regular unleaded gas. The means are split by groups on whether gas stations have direct competitors.

We can see that the mean for N is greater than Y, which means that there might be numerical evidence that supports the claim that "gas stations charge more if they lack direct competition in sight." In the data-set given, the gas stations that had no competitors had a mean price .0234 greater than those that do have competitors.

By creating a bootstrapped sampling distribution of the gas stations we have and finding the difference in means repeatedly, we can have an idea of what the real difference between gas stations with no competitors and those with competitors. Below is the confidence interval obtained of the mean differences between these subgroups of gas stations using 10000 bootstrapping re-samples.

Confidence Interval

name	lower	upper	level	method	estimate
diffmean	-0.0558823	0.0082185	0.95	percentile	-0.0166903

Conclusion

The data does support the claim. Since we obtained numerical evidence that the gas stations with no competitors having a $.0234 greater mean price than those with competitors, this our best guess from the immediate data which supports the claim. By the confidence interval calculated to find the effect size, we are 95% confident that the real difference between the mean gas prices for gas stations with and without direct competitors is somewhere between $-.056 and $.0082. Since this confidence interval contains 0, the difference between the mean values for gas stations with competitors and no competitors is not statistically significant. Even though the case is that we are "statistically uncertain" with our estimate since the difference could sway in either direction, $.056 is a much larger value than $.0082 indicating that we are we can be more confident that it falls in the negative difference. The negative difference favors that of gas stations having no competitors having greater prices. Thus, we can say that the evidence from the data does support the claim.

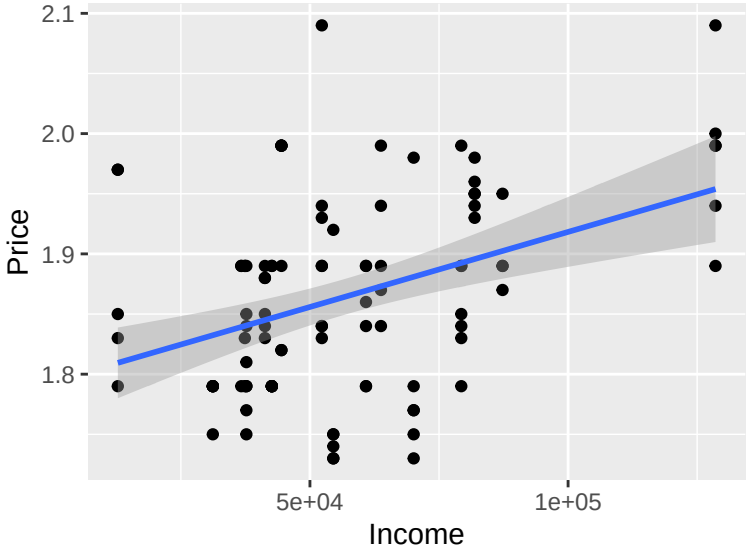
Theory B

Claim

The claim here is that the richer the area, the higher the gas prices.

Evidence

Scatterplot of Income vs. Gas Price For Regular Unleaded



	Linear Model for Price vs. Income
(Intercept)	1.7934425
Income	0.0000012

$$Price = .0000012(Income) + 1.79$$

The **R-value** for the correlation between income and the price is **0.396** which is evidence pointing towards a positive relationship between the two variables suggesting. This may suggest that gas prices could be higher in higher income areas. We also see that the linear model that results from this model shows that for every dollar increase in income, the price tends to increase by $.0000012. However, we also must account for statistical uncertainty, and on the next page is a 95% confidence interval of 10000 Monte Carlo simulations using bootstrapping to show an idea of what the true intercept and slope is.

Confidence Interval

name	lower	upper	level	method	estimate
Intercept	1.7590391	1.8293386	0.95	percentile	1.8027941
Income	0.0000007	0.0000018	0.95	percentile	0.0000010
sigma	0.0638650	0.0847947	0.95	percentile	0.0803473
r.squared	0.0392350	0.3204407	0.95	percentile	0.0656973
F	4.0428925	46.6826438	0.95	percentile	6.9613716

Conclusion

The data supports the claim. The linear model having a positive slope of $.0000012 supports the claim that the richer the area, the higher the gas prices. This is our best guess from the immediate data. More specifically, from the confidence interval, we can be 95% confident that the size of the association between gas price and income is between $.0000007 and $.0000018 dollars per increase in a dollar of income. The interval does not contain 0 so we know this difference is statistically significant and thus has a clear direction that gas prices in higher income areas tend to be higher.

Theory C

Claim

The claim here is that stations at stoplights charge more.

Evidence

Mean Price Comparison for Gas Prices on Whether Gas Stations are at Stoplights

Stoplights?	Mean Price
N	1.866
Y	1.863

"N" means no, the station is not at a stoplight, and "Y" means yes, the station is at a stoplight

Mean Difference: -0.00330

Here, from the data, gas stations that are at stoplights actually have a mean gas price for regular unleaded lower than gas stations not at stop lights. This contradicts the claim. However, we will also consider statistical uncertainty as well.

The confidence interval below was constructed by bootstrapping with 10000 samples and each time calculating the mean gas price difference between gas stations at stoplights and those that are not.

Confidence Interval

name	lower	upper	level	method	estimate
diffmean	-0.0375626	0.0304093	0.95	percentile	0.0027459

Conclusion

The data does not support the claim. In the confidence interval, we are 95% confident that the true mean difference in gas prices between gas stations at stoplights and those not at stoplights is between $-.0376 and $.0304. Since the confidence interval contains 0, the mean difference between gas stations at stoplights or not isn't statistically significant at the 5% level. The interval suggests statistical insignificance and thus suggests uncertainty of our estimate, so the data overall does not support the original claim. The real effect size could be in either direction. This time, there is not one positive or negative bound that is clearly more sizable than the other. Additionally, even though we found a mean price difference of $0.00330 in favor of gas stations not at stop lights, the confidence interval tells us that the mean price difference still could likely go either way and hence does not tell us that gas stations not at stoplights charge more either. Neither the original claim or an opposite claim can be supported by the data.

Theory D

Claim

The theory here is that gas stations with direct highway access charge more.

Evidence

Mean Price Comparison for Gas Prices on Whether Gas Stations Have Direct Access to the Highway

Highway Access?	Mean Price
N	1.854
Y	1.900

"N" means no, the station does not have direct access to the highway, and "Y" means yes, the station does have direct access to the highway.

Mean Difference: 0.0457

When observing the difference between the means, we see that gas stations with direct access to the highway have a mean price that is $0.0457 higher than those that don't have direct access, so it does seem like the evidence point towards the claim being true. Now, we must test for statistical uncertainty to determine the effect size.

Confidence Interval

name	lower	upper	level	method	estimate
diffmean	0.0085888	0.0809252	0.95	percentile	0.0284307

Conclusion

The theory is supported by the data. Our best guess for the mean difference from the immediate data set was $0.0457. Furthermore, based on the confidence interval, we are 95% confident that the true mean price difference between gas stations with direct highway access and those that don't is between $.00859 and $.0809. Since the interval doesn't contain 0, we know this difference is statistically significant and we know a clear direction in difference that the idea gas stations with direct highway access charge more. Hence, the theory is supported by the data.

Theory E

Claim

The claim here is that Shell charges more than all other non-Shell brands.

Evidence

Mean Price Comparison for Gas Prices for the Shell brand and Non-Shell Brands

Company	Mean Price
Not Shell	1.856
Shell	1.884

Mean Difference: 0.0274

Here, the mean difference in prices between gas stations under Shell and those that are not under Shell is $.0274 which may be evidence that the Shell brand tends to charge higher. Now, lets construct a confidence interval in order to account for statistical uncertainty.

Confidence Interval

name	lower	upper	level	method	estimate
diffmean	-0.0096966	0.0655659	0.95	percentile	0.0357534

Conclusion

The data does support the claim. Our best guess for the mean difference from the immediate data is $.0274 in favor of the Shell brand. Although the confidence interval containing 0 (-0.00979 to 0.0656 dollars) for the mean price difference between Shell and non-Shell brands indicates that our estimate is statistically insignificant, we can observe that there is still a clear direction of which brand has the higher price. This is because $.0656 is clearly higher than the negative bound of -$.00979 which goes to show that the data still supports the idea that there is still evidence of the effect of brand on the mean gas price. Therefore, the data does support the claim.

Problem 2

Part A

Confidence Interval for Average Mileage of 2011 S-Class 63 AMGs

name	lower	upper	level	method	estimate
mean	26322.09	31862.66	0.95	percentile	28609.7

Based on the confidence interval above, we are 95% confident that the average mileage of 2011 S-Class 63 AMGs is between 26,322 and 31,862 miles.

Part B

Confidence Interval for Proportion of 2014 550 S-Class Cars Painted Black

name	lower	upper	level	method	estimate
prop_TRUE	0.4170993	0.4527518	0.95	percentile	0.4461751

From the confidence interval above, we are 95% confident that the proportion of all 2014 550 S-Class cars painted black is somewhere between 41.7% and 45.3%.

Problem 3

Part A

Question

The question we are trying to answer is if there is evidence that one show consistently produces a higher mean happiness response among viewers.

Approach

The approach I used to answer the question is creating a 95% confidence interval to have an idea of where the true difference in mean rating for happiness lies. This would give us an idea if one show consistently has higher happiness ratings than the other. I used the bootstrapping technique by re-sampling the sample using the mosaic::resample() function 10000 times, calculating the difference between the two show's means by diffmean() each time, and finally calculated a 95% confidence interval with the confint() function that can report large sample confidence intervals.

Results

Confidence Interval for the Mean Difference Between Happiness Ratings for Living with Ed and My Name is Earl

name	lower	upper	level	method	estimate
diffmean	-0.3983935	0.0992348	0.95	percentile	-0.1047777

Conclusion

From the confidence interval shown above, the result obtain shows that we are 95% confident that the true mean difference between the happiness rating for the shows is between -0.398 points and .0992 points. Note that the negative difference favors the show *Living with Ed* and the positive favors *My Name is Earl*. Even though the true value could sway either way, notice that it is more plausible that the true value is a negative difference since .398 is much larger than .0992. Thus, as a conclusion, there is evidence that *Living with Ed* consistently produces a higher mean rating in happiness than *My Name is Earl*.

Part B

Question

The question we are answering here is whether the show *The Biggest Loser* or the show *The Apprentice: Los Angeles* consistently produces a higher mean in the "annoyed" response among viewers.

Approach

The approach I used was creating a 95% confidence interval to find a range where the true mean rating difference between the two shows reside in the annoyed response category. The tools I used included the mosaic::resampling function done 10000 to simulate bootstrapping, the diffmean() function to calculate the mean difference for each resample, and finally the confint() function to construct a 95% confidence interval.

Results

Confidence Interval for the Mean Difference Between Annoyed Ratings for The Biggest Loser and The Apprentice: Los Angeles

name	lower	upper	level	method	estimate
diffmean	-0.5263142	-0.020239	0.95	percentile	-0.3511905

Conclusion

The confidence interval tells us we can say with 95% confidence that the true mean difference between the annoyance ratings for both of the shows is between -.526 and -.020. The negative difference favors the show *The Biggest Loser* while the positive difference favors the show *The Apprentice: Los Angeles*. As a result, we are very confident that the true mean rating difference is a negative one that favors the show *The Biggest Loser*. Hence, there is evidence that the show *The Biggest Loser* consistently produces a higher mean in the annoyed response among viewers.

Part C

Question

The question we are answering is what proportion of American TV watchers would we expect to give a response of 4 or greater on "confusing" question for the show *Dancing with the Stars*, which indicates if the show confuses them or not.

Approach

Like the another parts, I constructed a 95% confidence interval to find a range where the true proportion of 4 or greater responses might lie. I used the mosaic::resample() function 10000 times to mimic a the bootstrapping procedure (resampling the sample with replacement), the prop() function to calculate the proportion of 4 or greater rating responses for the confused category, and finally the confint() function to construct the 95% confidence interval.

Results

Confidence Interval for the Proportion of Responses of Americans Rating 4 or Greater on "Confusing" for the show Dancing with the Stars

name	lower	upper	level	method	estimate
prop_TRUE	0.038674	0.1160221	0.95	percentile	0.0828729

Conclusion

With the 95% confidence interval above, we can interpret the results as we are 95% confident that the true proportion of American TV watchers to expect to give a 4 or greater on the confusing category is between

4% and 12%. To reiterate as a general conclusion, you can be 95% confident that 4-12% of Americans would find the show confusing.

Problem 4

Question

The question we are answering here is whether paid search advertising on Google Adwords is driving extra revenue or not based on an experiment by Ebay in where they did a controlled experiment by choosing a treatment group of DMAs (designated market areas) to shut down the advertising on Google Adwords for a month.

Approach

The approach I used was first finding the difference in the mean revenue ratio of the treatment and control groups by using diffmean(). This gives me my best guess as to the difference between them. However, to account for statistical uncertainty to obtain a degree of certainty to generalize the findings, I created a confidence interval. The confidence interval was constructed using 10000 Monte Carlo simulations using the bootstrap method by using the mosaic::resample() function to re-sample and using diffmean() to calculate each sample's difference in mean revenue ratio. The 95% confidence interval shown below was constructed using the confint() function.

Results

Comparison of the Mean Revenue Ratio for the Treatment Group and the Control Group

Company	Mean Revenue Ratio
0	0.949
1	0.897

1 indicates the treatment group, the group where the paid advertising on Google AdWords was paused for a month. 0 is the control group where the advertising on Google Adwords persisted.

Mean Difference: -0.0523

Confidence Interval for the Difference of Mean Revenue Ratio for the Treatment Group and the Control Group

name	lower	upper	level	method	estimate
diffmean	-0.0917418	-0.0134863	0.95	percentile	-0.0486815

Conclusion

For the given data set on the experiment, the our best guess of the mean difference in revenue ratio between the treatment and control group is -0.0523 in favor of the control group having a higher revenue ratio. The confidence interval tells us that we can be 95% confident that the true mean difference of revenue ratio between the control group and the treatment group is between -.0917 and -.0134 where the negative difference favors that of the control group having a higher revenue ratio. Even though this difference may seem small, 0 is not included in the interval. This tells us that the mean difference in revenue ratio is statistically significant and it gives us a clear sense of which direction the difference favors. Since this interval gives us a degree of certainty of the direction of difference favoring the control group, we can say that the evidence does support the idea that Google search advertising creates extra revenue for Ebay.