Final Project

Mallika Gupta

2020-12-09

Question of Interest

My question for the project is there a difference in means of engineering degrees and the location of the institution. This question is interesting as I want to see the different locations in the US that offer engineering degrees. It will also help me find out which location offers engineering degrees.

The variables that I will be using for this section are PCIP15 and LOCALE. The data in the variable PCIP15 is numerical, and for the variable locale is categorical.The two locations that I will calculate the difference between are city which has a large population of 250,000 or more and town which is distant(in urban cluster more than 10 miles and up to 35 miles from an urbanized area).

I will be using a hypothesis test for this question.

Preprocessing

```
college_reduced <- college
```

I am assigning the college dataset to a new datatframe which is called college_reduced.

```
college_wrangled <- college_reduced %>%
  select(PCIP15,LOCALE) %>%
  filter (LOCALE == '11' | LOCALE == '32')
```

In this code chunk, I am assigning the datatset to a new dataframe which is called college_wrangled. After that, I extracted the columns using the select function. The two columns that I selected are PCIP15 and LOCALE. Finally, I filtered the two values using the filter function. These two values were 11 and 32. The value 11 represents a city which has a large population of 250,000 or more. The value 32 represents a town which is distant(in urban cluster more than 10 miles and up to 35 miles from an urbanized area).

```
college_renamed <- college_wrangled %>%
rename(
  percent_engineering_degrees = `PCIP15`,
  locale = `LOCALE`
)
```

In this code chunk, I am assigning the dataset to a new dataframe which is called college_renamed. I am using the rename function to rename the columns more descriptively. I renamed the column PCIP15 to percent_engineering_degrees, and the column LOCALE to locale.

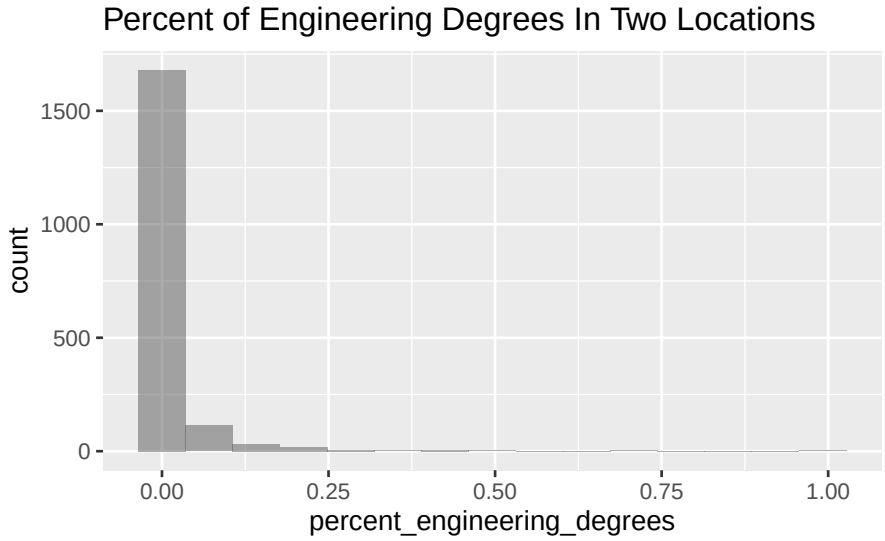
```
colleges_recoded <- college_renamed %>%
  mutate(
    college_locale = recode(
        locale,
        `11` = "city",
        `32` = "town"
    )
  )
```

In this code chunk, I am assigning the dataset to a new dataframe which is called colleges_recoded. I am using the mutate function. I created a new column which is called college_locale. I using the recode function to recode the integer values for the categorical variables that I am using. I recoded the value of 11 to city, and 32 to town.

Visualization

```
colleges_recoded %>%
 ggplot() +
 geom_histogram(mapping = aes(x = percent_engineering_degrees),
                position = "identity", alpha = 0.5, bins = 15) +
 labs (
   title = "Percent of Engineering Degrees In Two Locations"
 )
```

```
## Warning: Removed 128 rows containing non-finite values (stat_bin).
```



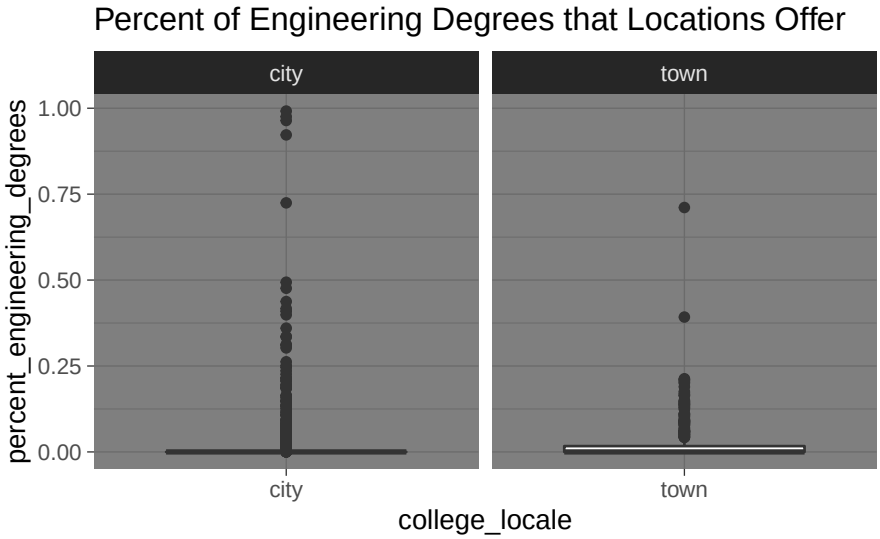
The first graph that I am using to visualize the data is a histogram. The reason I created this histogram is to show the percent of engineering degrees.

After looking at the histogram, I see that it is extremely skewed to the right. The spread of the histogram is from 0.0 to 0.25. Based on the histogram, I can interpret there are not many engineering degrees that are offered. This could be as not many colleges offer engineering degrees. Also, there are better colleges offering engineering degreees such as MIT, Georgia Tech and CalTech.

Maybe those colleges have a higher percent of engineering degrees as there are mainly focused on giving engineering degrees. Another reason could be that the locations I picked do not have a high percentage of engineering degrees.

```
colleges_recoded %>%
  ggplot () +
  geom_boxplot(mapping = aes(x = college_locale, y = percent_engineering_degrees)) +
  facet_wrap(~college_locale, scales = "free_x") +
    labs (
      title = "Percent of Engineering Degrees that Locations Offer"
    ) +
  theme_dark()
```

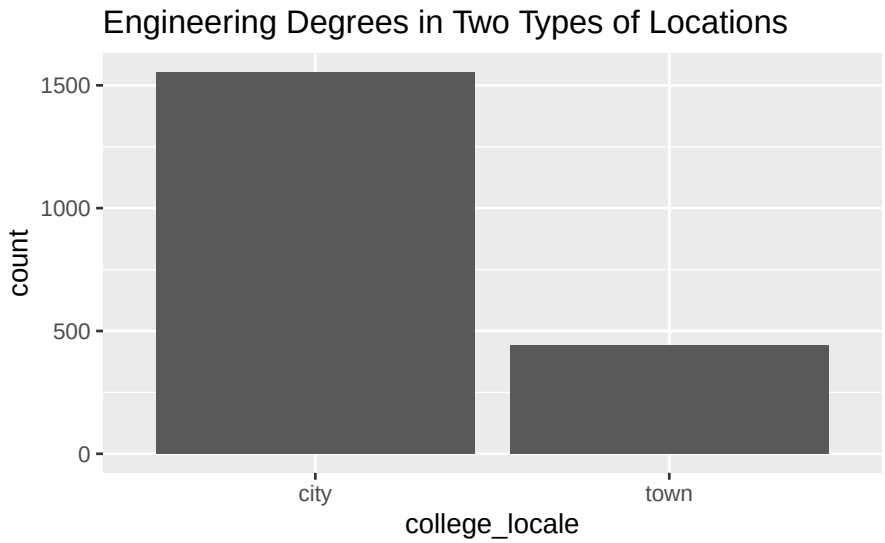
```
## Warning: Removed 128 rows containing non-finite values (stat_boxplot).
```



The next graph that I am using to visualize the data is a boxplot. The reason I created this boxplot was to show the percent of engineering degrees offered in two different locations which are a city which has a large population of 250,000 or more, and a town which is distant(in urban cluster more than 10 miles and up to 35 miles from an urbanized area).

For both locations, I can see that there are outliers present. For the location city, I can see there are about four outliers present. For the location town, I can see there are two outliers present. For the boxplot for location city, the spread of the data is from 0.0 to 1.00. For the boxplot for location town, the spread of the data is from 0.0 to 0.75. These boxplots shows that the percent of engineering degrees given at coleges in these two locations is not very much. Also, the colleges in these two locations may not be the best colleges for engineering. I used facet_wrap to have two separate boxplots show the percent of engineering degrees.

```
colleges_recoded %>%
  ggplot() +
  geom_bar(mapping = aes(x = college_locale)) +
  labs (
    title = "Engineering Degrees in Two Types of Locations"
  )
```



The last graph that I am using to visualize the data is a bar graph. The reason I created this bar plot was to see the was to see the number of engineering degrees offered in two locations which are which are a city which has a large population of 250,000 or more, and a town which is distant(in urban cluster more than 10 miles and up to 35 miles from an urbanized area).

After looking at the bar graph, I can see the count of engineering degrees in the location city is 1500. The count of engineering degrees in the location count is about 490. Based on the bar graph, more engineering degrees are given in the location city than the location town.

Summary Statistics

```
colleges_recoded %>%
  summarize(
    iqr = IQR(percent_engineering_degrees, na.rm = TRUE),
    mean = mean(percent_engineering_degrees, na.rm = TRUE),
    median = median(percent_engineering_degrees, na.rm = TRUE),
    range = range(percent_engineering_degrees, na.rm = TRUE),
    min = min(percent_engineering_degrees, na.rm = TRUE),
    max = max(percent_engineering_degrees, na.rm = TRUE),
    sd =  sd(percent_engineering_degrees, na.rm = TRUE),
    n()
  )
```

iqr	mean	median	range	min	max	sd	n()
0	0.0152221	0	0.0000	0	0.9917	0.066846	1991
0	0.0152221	0	0.9917	0	0.9917	0.066846	1991

In the previous code chunk for summary statistics, I found the interquartile range, mean, median, range, minimum value, maximum value, standard deviation, and count. After looking at the values, I noticed that the values were the same and given two times. I found these summary statistics for the variable percent_engineering_degrees.

```r
colleges_recoded %>%
  group_by(college_locale) %>%
  summarize(
    mean = mean(percent_engineering_degrees, na.rm = TRUE),
    median = mean(percent_engineering_degrees, na.rm = TRUE),
    iqr = IQR(percent_engineering_degrees, na.rm = TRUE),
    min = min(percent_engineering_degrees, na.rm = TRUE),
    max = max(percent_engineering_degrees, na.rm = TRUE),
    sd =  sd(percent_engineering_degrees, na.rm = TRUE),
    n()
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

college_locale	mean	median	iqr	min	max	sd	n()
city	0.0134278	0.0134278	0.0000	0	0.9917	0.0698632	1551
town	0.0211480	0.0211480	0.0169	0	0.7111	0.0553933	440

In this code chunk, I used the dataframe which is called colleges_recoded. I grouped by the continuous variable which is called college_locale. I found the summary statistics for the variable percent_engineering_degrees for the two locations which are city and town. The mean and median were exactly the same for the location city. The mean and median were exactly the same for the location town. The IQR was 0.0169 for the location town and 0 for the location city. The minimum value was same for both locations. The maximum value for the location city was higher than the maximum value for the location town.The standard deviation was higher for the location city than the maximum value for the locatio town. The percent of engineering degrees given in the location city was higher than the percent of engineering degrees given in the location town.

Data Analysis

The null hypothesis is there is no difference in means between percent of engineering degrees and the location of the institution. The alternative hypothesis is there is a difference in means between percent of engineering degrees and the location of the institution.

I am going to be using a two sided test.

The test statistic will be the difference in means between percent of engineering degrees and the location of the institution.

```r
college_recoded_null <- colleges_recoded %>%
specify(formula = percent_engineering_degrees ~ college_locale) %>%
hypothesize(null = "independence") %>%
generate(reps = 10000, type = "permute") %>%
calculate(stat = "diff in means", order = combine("city", "town"))
```

```
## Warning: Removed 128 rows containing missing values.
```

This is the first step in doing a hypothesis test. I am calculating the difference in means for percent of engineering degrees and location of the institution. I am conducting a hypothesis test for 10,000

observations.

```
college_recoded_obs_stat <- colleges_recoded %>%
specify(formula = percent_engineering_degrees ~ college_locale) %>%
calculate(stat = "diff in means", order = combine("city", "town"))
```

```
## Warning: Removed 128 rows containing missing values.
```

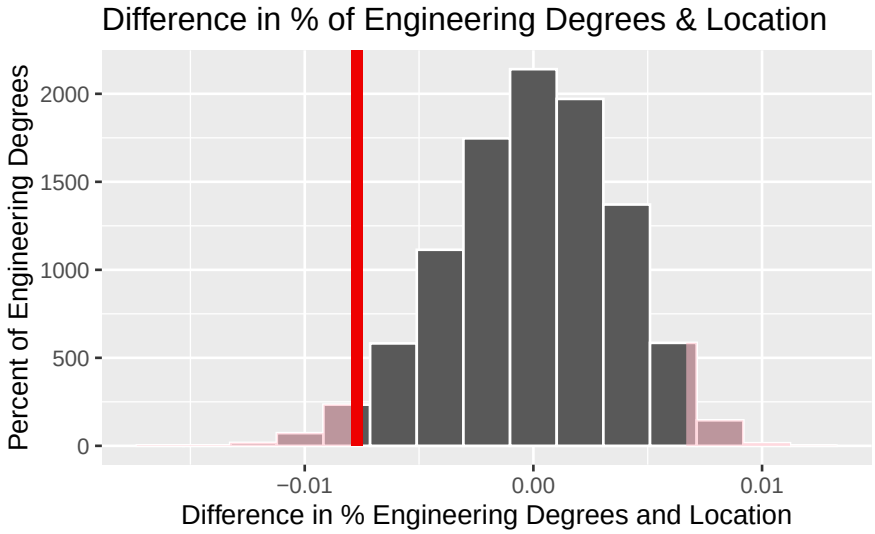
This is the second step in doing a hypothesis test.

```
college_recoded_null %>%
get_p_value(obs_stat = college_recoded_obs_stat, direction = "both")
```

p_value
0.047

The p - value is less than the alpha value of 0.05. I would reject the mull hypothesis. Based on this, I can reject the null hypothesis which is there is no difference in means between percent of engineering degrees and location of the insitution. The results are statistically significant as the p - value of 0.047 is less than the alpha value of 0.05.

```
college_recoded_null %>%
visualize() +
shade_p_value(obs_stat = college_recoded_obs_stat, direction = "both") +
labs (
title = "Difference in % of Engineering Degrees & Location",
x = "Difference in % Engineering Degrees and Location",
y = "Percent of Engineering Degrees"
)
```



In this code chunk, I am visualizing the p - value in the graph above.

```
college_bootstraps <- colleges_recoded %>%
specify(percent_engineering_degrees ~ college_locale) %>%
```

```
generate(10000, type = "bootstrap") %>%
calculate(stat = "diff in means", order = c("city", "town"))
```

```
## Warning: Removed 128 rows containing missing values.
```

This is the first step in doing a confidence interval. I am conducting a confidence interval for 10000 observations.

```
bootstrap_ci <- college_bootstraps %>%
get_confidence_interval()
```

```
## Using `level = 0.95` to compute confidence interval.
```

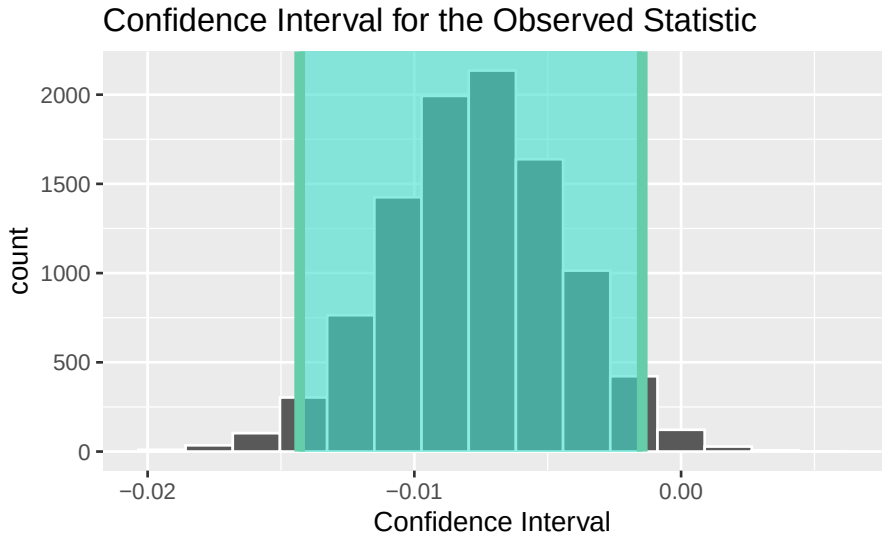
This is the second step of doing a confidence interval. The level of confidence that I am using for my confidence interval is 95%

```
bootstrap_ci
```

lower_ci	upper_ci
-0.0142997	-0.001456

I found the confidence interval which is (-0.014, -0.015)

```
college_bootstraps %>%
visualize() +
shade_confidence_interval(bootstrap_ci) +
labs (
title = "Confidence Interval for the Observed Statistic",
x = "Confidence Interval"
)
```



In this code chunk, I am visualizing the confidence interval.

The observed statistic or p - value is 0.047. The observed statistic does not fall within the range of the confidence interval. The confidence interval is (-0.014, -0.0015)

Conclusion

For my visualization section, I made three graphs which were a histogram, boxplot and bar graph. For the histogram, I could see that it was extremely skewed to the right. Based on the histogram, I could see that the percent of engineering degrees given is not that much. This could be that many colleges do not offer engineering degrees. For the boxplots, I could see that there were outliers for both locations. This shows that the percentage of engineering degrees is extremely low. For the bar graph, the count for percentage of engineering degrees is higher for the location city than the location town. The percentage of engineering degrees given in the city is higher than those given in the town.

When I did the summary statistics, the mean,median, IQR, standard deviation, minimum value, maximum value, count, and range were exactly the same. Ths was for the variable percentage of engineering degrees. I found the summary statistics for the variable percent_engineering_degrees for the locations which are city and town. The mean and median were exactly the same for the location city. The mean and median were exactly the same for the location town. The IQR was 0.0169 for the location town and 0 for the location city. The minimum value was same for both locations. The maximum value for the location city was higher than the maximum value for the location town.The standard deviation was higher for the location city than the maximum value for the locatio town. The percent of engineering degrees given in the location city was higher than the percent of engineering degrees given in the location town.

I conducted a hypothesis test. I got a p - value of 0.047 which is less than the alpha value of 0.05. Based on this, I can reject the null hypothesis which is there is no difference in means between percent of engineering degrees and location of the insitution. The results are statistically significant as the p - value of 0.047 is less than the alpha value of 0.05.

I conducted a confidence interval.The observed statistic or p - value is 0.047. The observed statistic does not fall within the range of the confidence interval. The confidence interval is (-0.014, -0.0015)