Final Project

Avantika Prativadhi

2021-12-05

Question of Interest

As an interational student myself, I remember having a difficult time finding universities to apply to. With different universities offering up to different levels of degrees, it's difficult to navigate through as a student that is not used to the American education system. Hence, for this final project, my question of interest is to see if there a difference in means of the alien (international)student enrollment rate between universities that grant only Bachelor's degree and universities that grant both Bachelors degree and Masters degree. I will be using statistical inference to explore this question with my categorical variable being the HIGH degree column which contains the highest degree that is offered at the university and my continuous response variable will be the alien students enrollment rate which has the total share of international student enrollment. I think that this analysis will be interesting as it will give an insight on international student enrollment in terms of the type of college they choose. This analysis can be helpful for universities as well as other international students.

Preprocessing

1.

```
college_reduced <- college %>%
  select (HIGHDEG, UGDS_NRA)
```

In this code block, I used the select function to select the two columns I will be working with. The first is the HIGHDEG column which will be my categorical variable and the UGDS_NRA which will be my continuous response variable.

2.

```
college_reducednew <-college_reduced %>%
  rename(
    "degrees_awarded" = HIGHDEG,
    "enrollment_rate" = UGDS_NRA
    )
```

3.

```
college_data <-college_reducednew  %>%
  mutate(
    degree_type = recode(
        degrees_awarded,
```

```
        `0` = "Non-degree",
        `1` = "Certificate Degree",
        `2` = "Associates Degree",
        `3` = "Bachelor's Degree",
        `4` = "Master's Degree"
    )
  ) %>%
  filter(degree_type == "Bachelor's Degree"| degree_type == "Master's Degree")
```

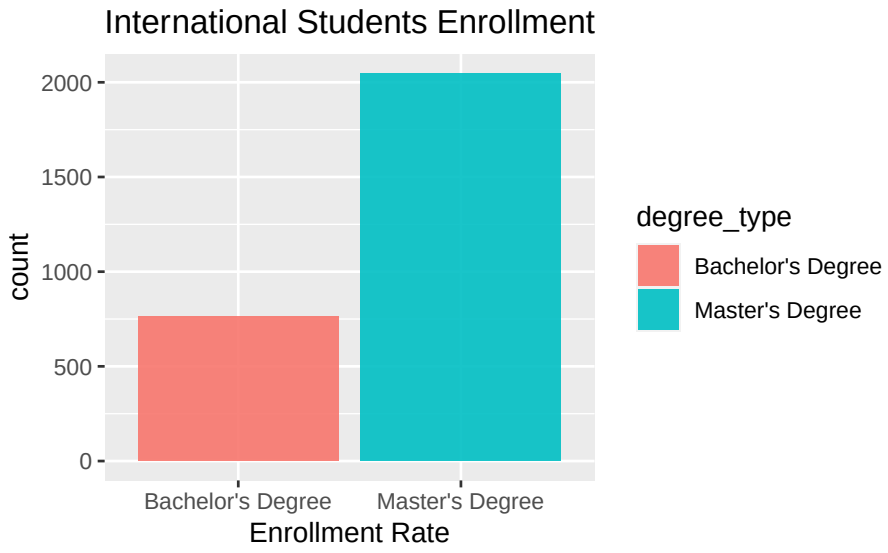
In the code chunk from question 2, I used the rename function to rename the two columns.For question 3, I used the mutate function as well as the recode function to rename the integer values in the categorical variables.The HIGHDEG column contains the highest degree level awarded by the universities where 0 represents Non-degree granting, 1 represents a certificate degree, 2 represents an Associate degree, 3 represents a Bachelor's degree, and 4 represents a Graduate degree. The UGDS_NRA represents the total share of enrollement of undergraduate degree-seeking students that are considered non-resident aliens (also known as international students). Using the filter function, I filtered the dataset so that only the rows with Bachelor's or Master's degrees are shown.

Visualization

In this section of the project, I will be visualizing my data through different plot functions using ggplot. My main goal in this section is to conduct Exploratory Data Analysis and answer two general goals.

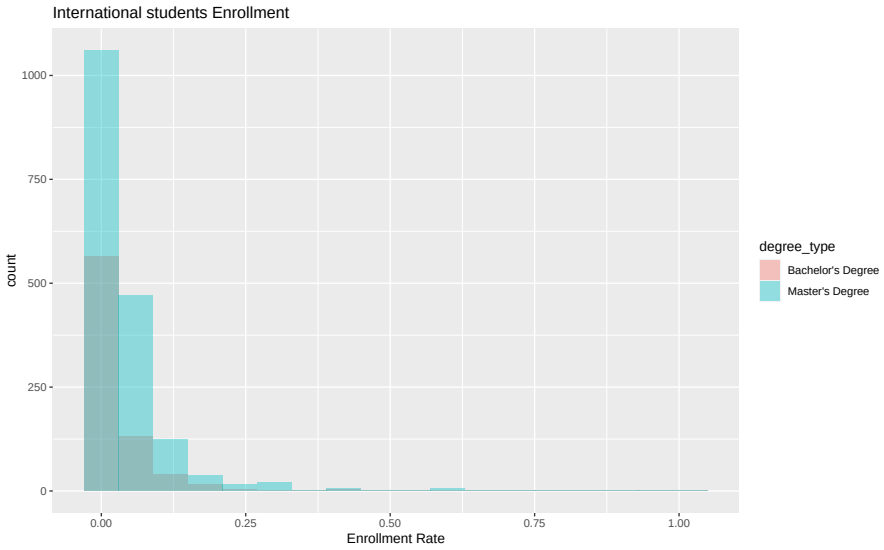
1. Using the bar plot and histogram below, I explored the data to answer the first goal which is: What type of variation occurs within my variables?

```
college_data %>%
  ggplot() +
  geom_bar(aes(x = degree_type, fill = degree_type),position = "identity",
           alpha = 0.9)+
  labs(title='International Students Enrollment', x = "Enrollment Rate")
```



In the above bar plot, I used the fill parameter to differentiate between the types of degrees to see where the enrollment of students is higher.The blue bar is the colleges that offer upto Master's Degree and the red bar is colleges that offer only up to Bachelor's degree. The height of the bars display the number of observations that occured with each enrollment rate. Based on the plot, the count of the internation enrollment for Bachelor's degree is between 500 and 1000 while the count for Master's degree is a little over 2000. It can be concluded that the international enrollment is higher for the universities that offer upto Master's Degree.

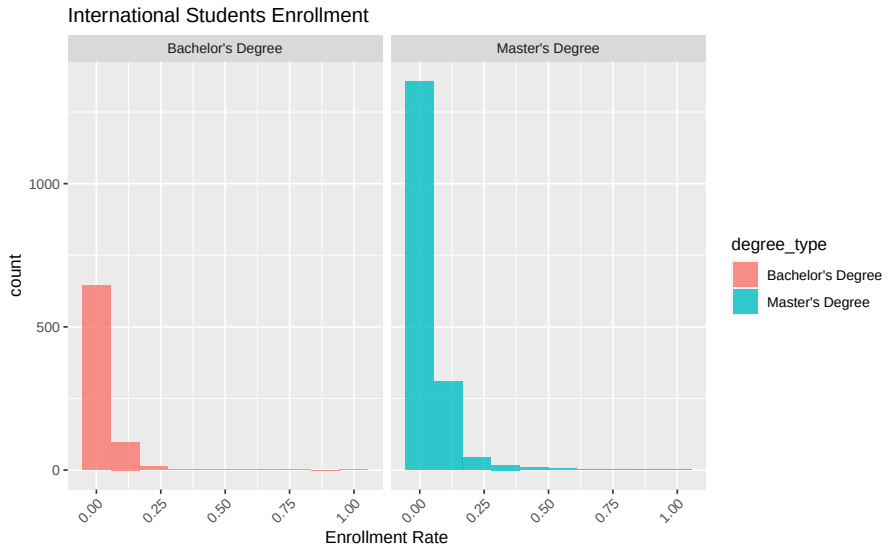
```
ggplot(college_data)+
  geom_histogram(aes(x = enrollment_rate,
                     fill = degree_type),
                 binwidth = 0.06,
                 position = "identity",
                 alpha = 0.4)+
  labs(title='International students Enrollment', x = "Enrollment Rate")
```



The above histogram helps futher compare the enrollment rate with the type of universities that international student enroll into.Using the bindwidth argument, I was able to set the range of the values of each bar and display a trend in the data. With the alpha parameter, I made the bars lighter so that both the degree types can be visible.The histogram is skewed right as the tail of the distribution is towards the right. An interesting observation in this histogram is that the count for Master's degrees is a lot higher than the count for Bachelor's Degree at the 0 enrollment rate. However, when looked at the overall trend in the graph, it shows that the number of enrollment in universities with a Master's degree is higher than universities with only up to Bachelor's degree

```
college_data%>%
  pivot_longer(cols = enrollment_rate,
               names_to="Names",
               values_to = "Values") %>%
  ggplot() +
  geom_histogram(aes(x = Values, fill =degree_type),
                 bins = 10,
                 position = "identity",
```

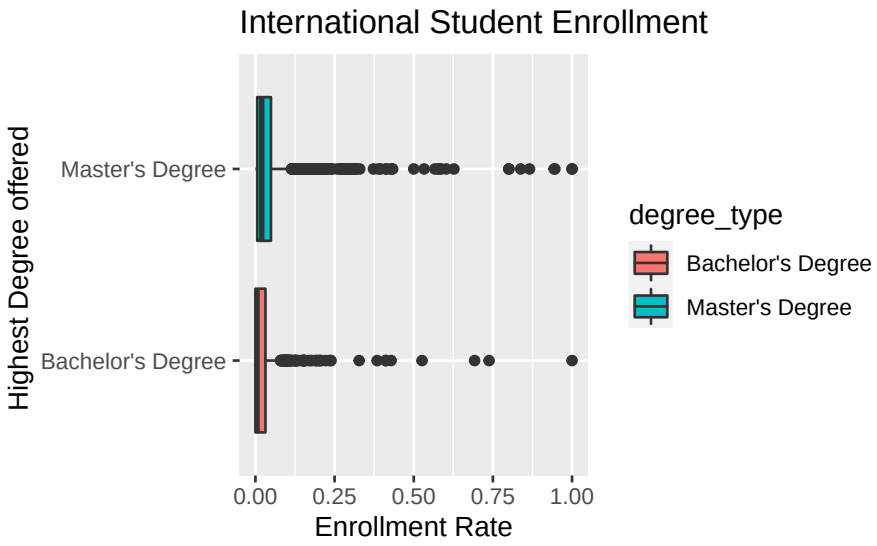
```
                  alpha = 0.8) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  facet_grid(~degree_type,scales="free")+
  labs(
  title ='International Students Enrollment',
    x='Enrollment Rate'
)
```



In the above visualization, I used pivot_longer and facet_grid, to break down the two degree types to see where the enrollment rate is higher. Similar to the previous histogram, this distribution is skewed right. When zoomed into the distrubution using the fig.width parameter, the outliers for both degree types are visible towards the tail end. Based on the distribution, the enrollment rate in universities that offer both Bachelor's degree and Master's Degree is high.

2. Using the boxplot below, I explored the second question: What type of covariation occurs between my variables?

```
ggplot(data = college_data)+
  geom_boxplot(aes(x = reorder(degree_type,enrollment_rate,FUN=median),
                   y = enrollment_rate,
                   fill = degree_type))+
  coord_flip()+
  labs(
    title = "International Student Enrollment",
    x = "Highest Degree offered ",
    y = "Enrollment Rate"
  )
```



The boxplot above displays the distribution of the enrollment rate broken down by the degree type.Using the fill variable, I was able to differentiate the type degree types based on color. To make the trend easier to see, I reordered the degree_type variable based on the median value of enrollment rate. I also used coord_flip function to flip the boxplot. There are a lot of outliers in the plot and the iqr for the Master's degree is slightly higher than the Bachelor's degree.This plot supports the findings in the previous plots that the colleges with Master's degrees have a higher enrollment overall.

Summary Statistics

1. In this section of the project, I will be calculating the summary statistics for each of my variables.

```
college_data%>%
  group_by(degree_type)%>%
  summarise(
    count = n()
  )
```

degree_type	count
Bachelor's Degree	762
Master's Degree	2047

The code chuck above is the summart statistics of the degree type variable. Using the summarize variable, the count of both of the two categories was calculated. While the count for the Master's degree is a lot higher than the count for the Bachelor's degree, this does not affect the analysis as the mean that was calculated is independent from the count of the variables.

```
college_data%>%
  group_by(degree_type)%>%
  summarise(
    count = n(),
```

```
    mean = mean(enrollment_rate, na.rm = TRUE),
    median = median (enrollment_rate,na.rm = TRUE),
    std.dev = sd( enrollment_rate,na.rm = TRUE),
    iqr = IQR (enrollment_rate,na.rm = TRUE),
    min = min (enrollment_rate,na.rm = TRUE),
    max = max(enrollment_rate,na.rm = TRUE)
)
```

degree_type	count	mean	median	std.dev	iqr	min	max
Bachelor's Degree	762	0.0296115	0.00515	0.0724745	0.03165	0	1
Master's Degree	2047	0.0458733	0.02000	0.0889476	0.04345	0	1

The above code chunk is the summary statistics for the enrollment rate. The summarize function was used to find the count, mean, median, standard deviation, iqr, min, and max of the variable. I used the na.rm parameter to ignore all the rows with missing values. Based on the table, the mean,iqr, and standard deviation for the Master's degree is slightly higher than Bachelor's degree. This supports the plots from the previous section with the Master's degree having a higher rate of enrollment. While the difference in the means is small, it is still notable enough to test it and find the test statistic.

Data Analysis

1. In this section of the project, I will be using inference to answer my question of interest. I will be running a difference of means test to either reject or accept my null hypothesis.

The null hypothesis is that there is no difference in mean of the alien enrollment rate between universities that grant only Bachelor's degree and universities that grant both Bachelors degree and Masters degree.The alternate hypothesis is that there is a difference in mean of alien enrollment rate between universities that grant only Bachelor's degree and universities that grant both Bachelors degree and Masters degree.For my data analysis, I will be using a two-sided hypothesis test and the test statistic is the difference between both the means.

```
college_null <- college_data%>%
  specify(enrollment_rate~degree_type) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in means",
            order = c("Bachelor's Degree", "Master's Degree"))
```

In this code chunk, I created a null distribution by running 10,000 permutations of the original data. Since my test statistic is difference in means, I added that to the stat parameter in the calculate function.

```
college_observed_stat <- college_data%>%
  specify(enrollment_rate~degree_type) %>%
  calculate(stat = "diff in means",
            order = c("Bachelor's Degree", "Master's Degree"))
```

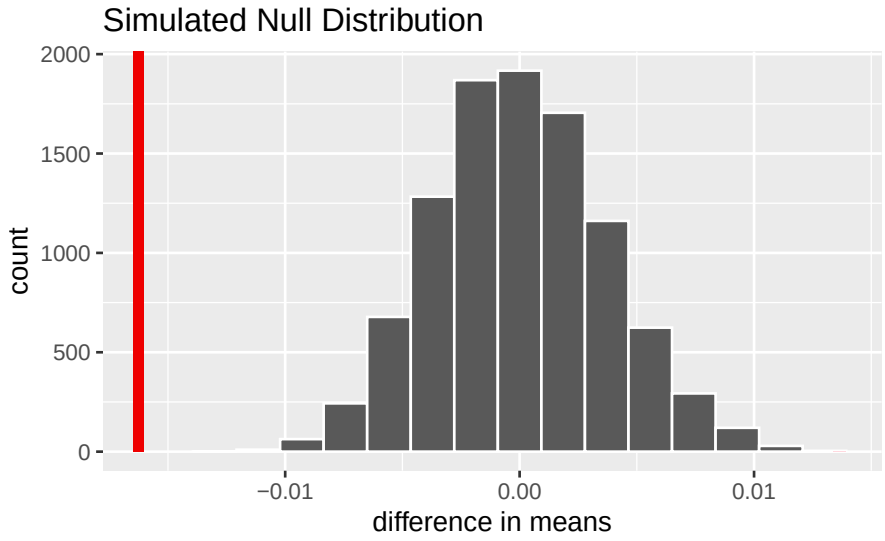
With the null distribution calculated in the previous code chunk, I calculated the observed statistic which I will use to calculate the p value.

```
college_null%>%
get_p_value(obs_stat = college_observed_stat, direction = "two sided")
```

p_value
0

In the above code chunk, I calculated the p value of the data using the get_p_value function. I will use this value to either accept or reject the null hypothesis.

```
visualise(data = college_null)+
  shade_p_value(college_observed_stat, direction = "two sided")+
  labs(
    title = 'Simulated Null Distribution ',
    x = 'difference in means'
  )
```



I used the visualize and shade_p_value to visualize the p value that was calculated previously.

The p value of 0 is less than the alpha value of 0.05. This means that there is strong evidence against the null hypothesis and it is rejected.This difference of means test is statistically significant. Based on this test, there is a difference in mean of alien enrollment rate between universities that grant only Bachelor's degree and universities that grant both Bachelors degree and Masters degree. Which means that international students are favoring universities that offer upto Master's Degrees over universities that have Bachelor's degrees as their highest level of degree awarded.

Conclusion

Based on all the findings from the previous sections, it can be concluded that international (alien) students tend to attend colleges that offer both, a Bachelor's degree as well a Master's degree. To answer the original question of interest, yes, there is a difference in mean of international student

enrollment between univerisities that grant only until Bachelor's degree and universities that grant both Bachelors degree and Master's degree. The analysis from each of the section support each other. The graphs in the visualization section were a visual representation for the difference in the enrollment between both kinds of universities. In all these graphs, the count of the universities with Master's degrees were higher than the count of the universities with only upto Bachelor's degree.The summary statistics secton, further supported the graphs through calculations as the mean,standard deviation, and iqr of the Master's degree was higher than the Bachelor's degree. Finally, the data analysis section, which ran the difference of means test established that there was a strong statistical difference between each of the type of degree offered. My findings can help other international students decide the which kind of university to apply to. It can also help universities when they want to send target ads to a specific demographic. Universities which offer upto Master's degrees can use the data to persuade more international students to join. Additionally, universities that only offer until Bachelor's degrees can start adding more programs and granting Master's degrees to increase their overall student enrollment.