Final Project

Minh Tran

2021-06-15

Question of Interest

Is there a difference in completion rate (100% of expected time to completion) for first time, full time students at four-year institutions between Bachelor-degree students vs. Non-Bachelor (Associate or Graduate) degree students?

Preprocessing

```
#Selecting two variables, dropping NAs, and filtering out non degrees(0) and certificate degre
college <- college %>%
  select(HIGHDEG, C100_4) %>%
  filter(HIGHDEG != 0, HIGHDEG != 1)
```

```
#renaming columns
college <- college %>%
  rename(Completion_rate = C100_4)
```

```
#Making another column stating if the student is going for a Bachelor's degree or other
college <- college %>%
  mutate(Degree = if_else(HIGHDEG == 3,
                          'Bachelor',
                          'Other')) %>%
  select(Degree, Completion_rate)
```

```
#Convert completion rate to percentages
college <- college %>%
  mutate(Completion_rate = Completion_rate * 100)
```

```
glimpse(college)
```

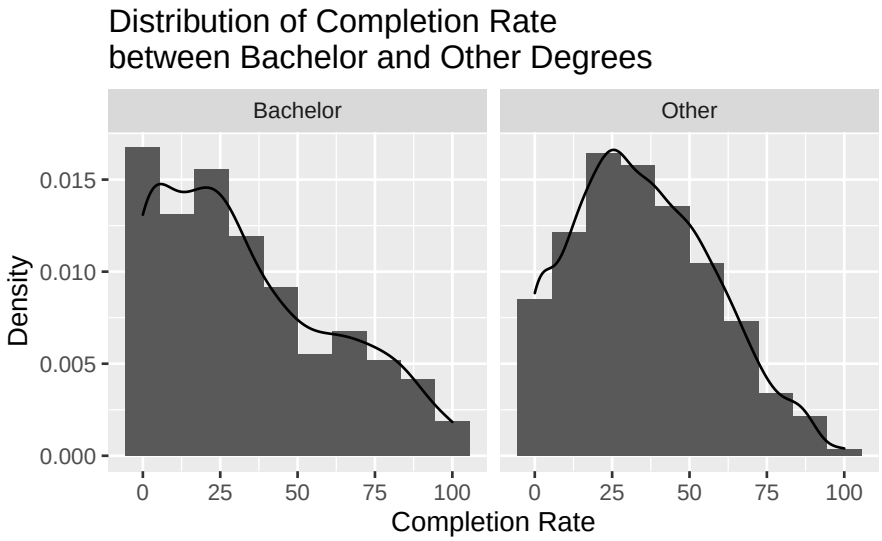
```
## Rows: 4,322
## Columns: 2
## $ Degree          <chr> "Other", "Other", "Other", "Other", "Other", "Other...
## $ Completion_rate <dbl> 3.81, 31.79, 0.00, 22.70, 11.29, 44.04, NA, NA, 8.8...
```

Visualization

```
#Creating a probability mass function plot
ggplot(data = college) +
geom_histogram(aes(x = Completion_rate, y = ..density..),bins = 10) +
geom_density(aes(x = Completion_rate)) +
facet_wrap(~Degree) +
labs(title = 'Distribution of Completion Rate
between Bachelor and Other Degrees',
x = 'Completion Rate',
y = 'Density')
```

```
## Warning: Removed 2219 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2219 rows containing non-finite values (stat_density).
```

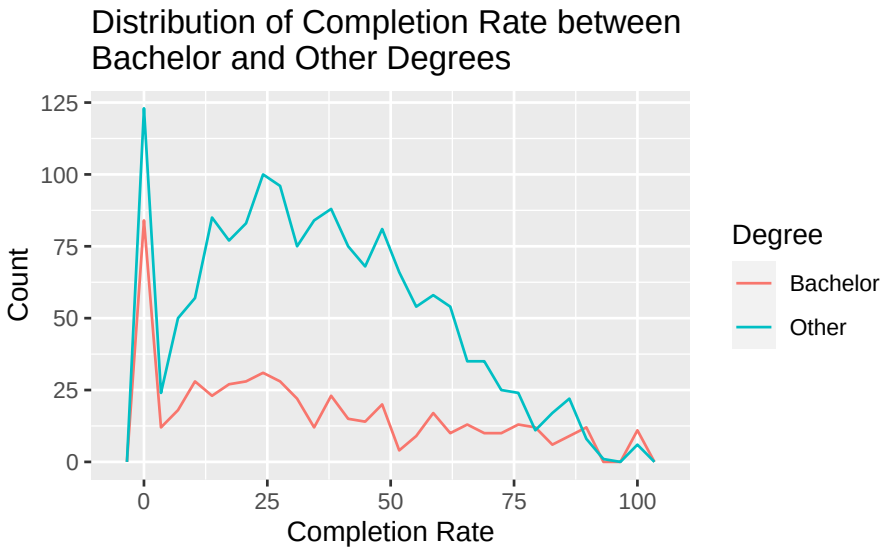


The probablity mass function plot show that the distribution between Bachelor and Other degrees are somewhat kind of similar to each other but they also differ especially in the first half of the plots.

```
#Creating a frequency polygon plot
ggplot(data = college) +
geom_freqpoly(aes(x = Completion_rate,color = Degree)) +
labs(title = 'Distribution of Completion Rate between
Bachelor and Other Degrees',
x = 'Completion Rate',
y = 'Count')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

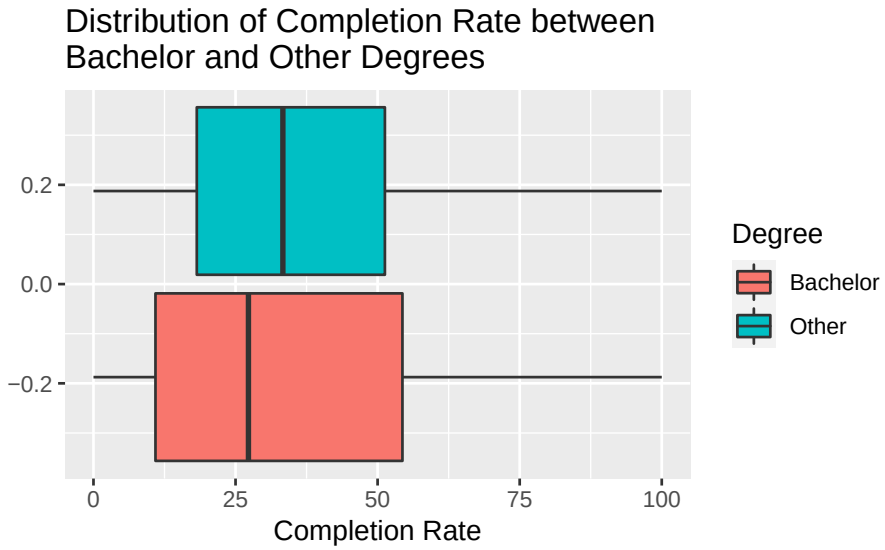
```
## Warning: Removed 2219 rows containing non-finite values (stat_bin).
```



The frequency polygon plot shows that there are more observed counts for Other degrees. Just like the probablity mass function plot, the distribution of Bachelor and Other Degrees are kind of similar but also differ from the other.

```
#Creating a boxplot
ggplot(data = college) +
geom_boxplot(aes(x = Completion_rate, fill = Degree)) +
labs(title = 'Distribution of Completion Rate between
Bachelor and Other Degrees',
x = 'Completion Rate')
```

```
## Warning: Removed 2219 rows containing non-finite values (stat_boxplot).
```



The boxplot shows that the 'Other' degrees have a higher median of completion rate compared to the Bachelor degrees. On the other hand, the Bachelor degree has a lower value of the 'lower quartile' and a higher value of the 'upper quartile'. In addition, the Bachelor degree has a wider inner quartile range of completion rates.

Summary Statistics

```
#Calculating summary statistics
college %>%
  group_by(Degree) %>%
  summarise(Count = n(),
            Mean = mean(Completion_rate, na.rm = TRUE),
            Median = median(Completion_rate, na.rm = TRUE),
            Sd = sd(Completion_rate, na.rm = TRUE),
            Iqr = IQR(Completion_rate, na.rm = TRUE),
            Minimum = min(Completion_rate, na.rm = TRUE),
            Maximum = max(Completion_rate, na.rm = TRUE))
```

Degree	Count	Mean	Median	Sd	Iqr	Minimum	Maximum
Bachelor	762	33.58879	27.27	27.91792	43.4600	0	100
Other	3560	35.43408	33.33	22.52260	33.1025	0	100

Data Analysis

$\alpha = 0.05$

Null Hypothesis: There is no difference in completion rate between Bachelor and Other degrees.

Alternative Hypothesis: There is a difference in completion rate between Bachelor and Other degrees.

A two-sided test will be used.

The test statistic that will be used is the difference of means in completion rate between Bachelor and Other degrees.

```
#null distribution
college_null <- college %>%
specify(Completion_rate ~ Degree) %>%
hypothesize(null = "independence") %>%
generate(reps = 10000, type = "permute") %>%
calculate(stat = "diff in means", order = c("Bachelor", "Other"))
```

```
## Warning: Removed 2219 rows containing missing values.
```

```
#obs stat
college_obs_stat <- college %>%
  specify(Completion_rate ~ Degree) %>%
  calculate(stat = "diff in means", order = c("Bachelor", "Other"))
```

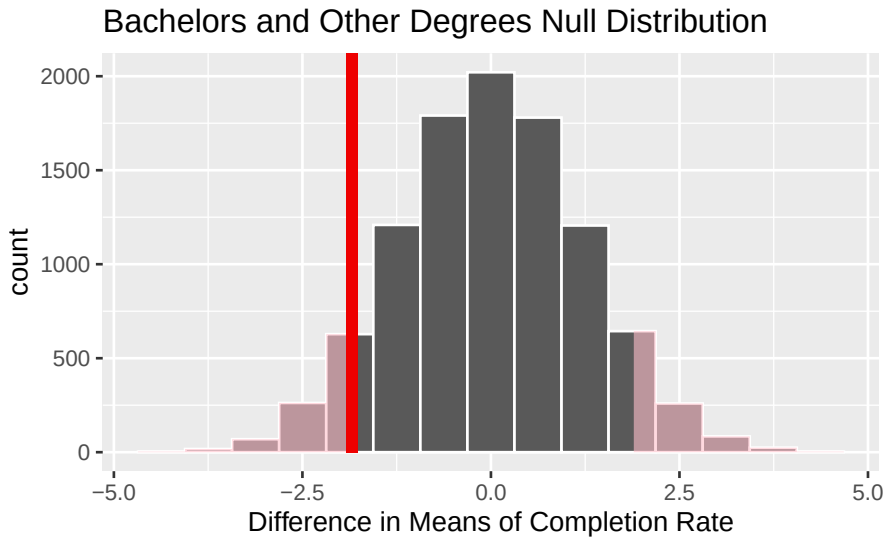
```
## Warning: Removed 2219 rows containing missing values.
```

```
#getting p-value
college_null %>%
```

```
  get_p_value(obs_stat = college_obs_stat, direction = "two_sided")
```

p_value
0.1246

```
#visualization
college_null %>%
  visualize() +
  shade_p_value(obs_stat = college_obs_stat, direction = "two_sided") +
  labs(title = 'Bachelors and Other Degrees Null Distribution',
  x = 'Difference in Means of Completion Rate')
```



We fail to reject the null hypothesis since the p-value is greater than $\alpha = 0.05$***. From this, we conclude that there is no significant difference in the completion rate between Bachelor and Other degrees.***

Conclusion

Based on the the analyses, the conclusion that has been reached shows that there is not a significant difference in the completion rate between Bachelor and Other degrees. As a result, the results indicate that the completion rate between Bachelor and Non-Bachelor degrees did not differ significantly. All the different sections support each other because the visualizations indicate that there are little differences, but they are not that significant. In addition, the data analysis agrees with this concept because the p-value is greater than $\alpha = 0.05$, so we fail to reject the null hypothesis and conclude that there is no significant difference.

For this specific question of interest, there were no other variables that I could of used since I just wanted to find if there is a difference in completion rate for students at four-year institutions (100% of expected time) among Bachelor and other degrees (Associate or Graduate). As a result, the only original columns that were used for this hypothesis testing was the 'C100_4' and the 'HIGHDEG' column. It was interesting to find out more about this because students are going after other degrees

other than a Bachelors at these four-year institutions. So, would the completion rate be different based on the type of degree? Based on the hypothesis testing, there is no significant difference in completion rate whether it be obtaining an Associate, Bachelor, or Graduate degree.