MA361 Project 1:

Mean Difference Between Gender Regarding College Completion

By: Alex Kuhn, Sophie Laposha, amd Corbin Couger

Table of Contents

Introduction

Completing college is one of the biggest accomplishments of life, it's the first steps into the real world. It's one of the most challenging and awakening periods throughout our lives. In college you learn a lot more than just the education you're there for, you begin to learn the different parts of how your life will pan out. This period of life can create a lot of stress and be overwhelming. In addition, there can be issues that make completing college hard. This is why we chose to dig deep and find out if you will be more likely to complete college depending on your sex among vastly different regions of the United States. So, we pulled data from datausa and found two separate datasets, one with all students who have recently completed college in New York City and then all the students who have recently completed college in Los Angeles. We combined and mined the data to help us forge our question and help us answer it using our selected statistical procedure. We want to know: Does the mean Women college graduates differ from the mean Men graduates? We are trying to see if this is a significant difference which can help us further find reasons as to why this might be a possibility.

Data Collection

As explained in the introduction, we collected our data from datausa and then put it into R Studio to check for outliers and to further mine the data. After our outlier analysis, we removed some non-essential columns and then subsetted the data into 2 different sets; “Men data” which is the men who completed college, “Women data” which is the women who completed college on the East Coast and West Coast combined. After doing this we were able to take the summary of each set (ex: summary(Mendata)) to find the mean completions, which we highlighted below. We also used R Studio to help us find the standard deviations to help us figure out what hypothesis test to use.

Summary of Men data:

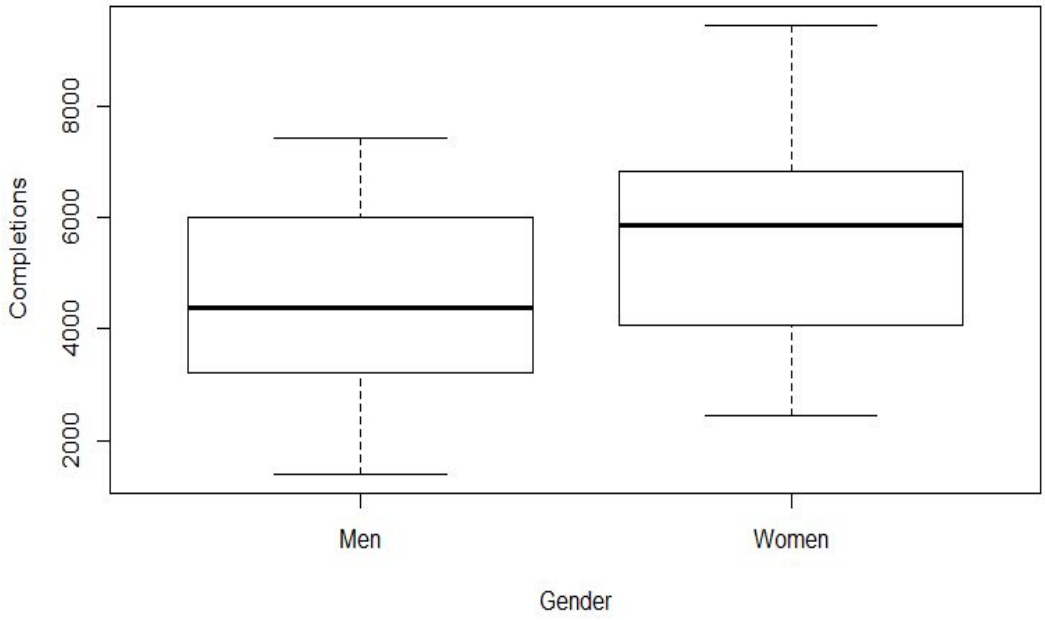
```
  Gender                                          University  Completions
 Men  :60   California State University-Long Beach     : 6   Min.   :1387
 Women: 0   California State University-Los Angeles    : 6   1st Qu.:3221
            California State University-Northridge     : 6   Median :4394
            Columbia University in the City of New York: 6   Mean   :4507
            CUNY Hunter College                        : 6   3rd Qu.:6000
            New York University                        : 6   Max.   :7437
            (Other)                                    :24
```

Summary of Women data:

```
 Gender                                          University  Completions
Men  : 0   California State University-Long Beach     : 6   Min.   :2446
Women:60   California State University-Los Angeles    : 6   1st Qu.:4076
           California State University-Northridge    : 6   Median :5857
           Columbia University in the City of New York: 6   Mean   :5847
           CUNY Hunter College                       : 6   3rd Qu.:6779
           New York University                       : 6   Max.   :9455
           (Other)                                   :24
```

Outlier Analysis

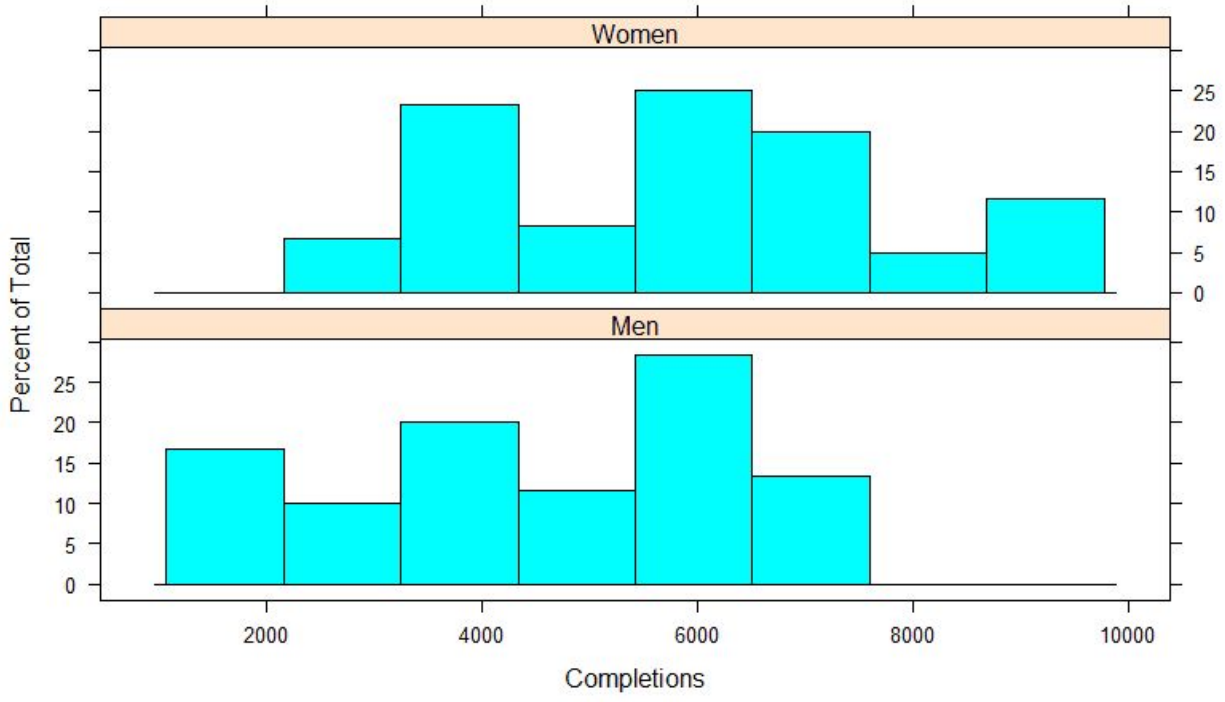
Our data provide the number of completions from men and women at individual colleges in the New York and Los Angeles regions. Before we perform our test proving the difference between genders, we need to see whether our individual observations contain outliers. When we performed a boxplot on our Completion observations both for our Men groups and our Women, we saw no outliers in either datasets and a very normal boxplot:



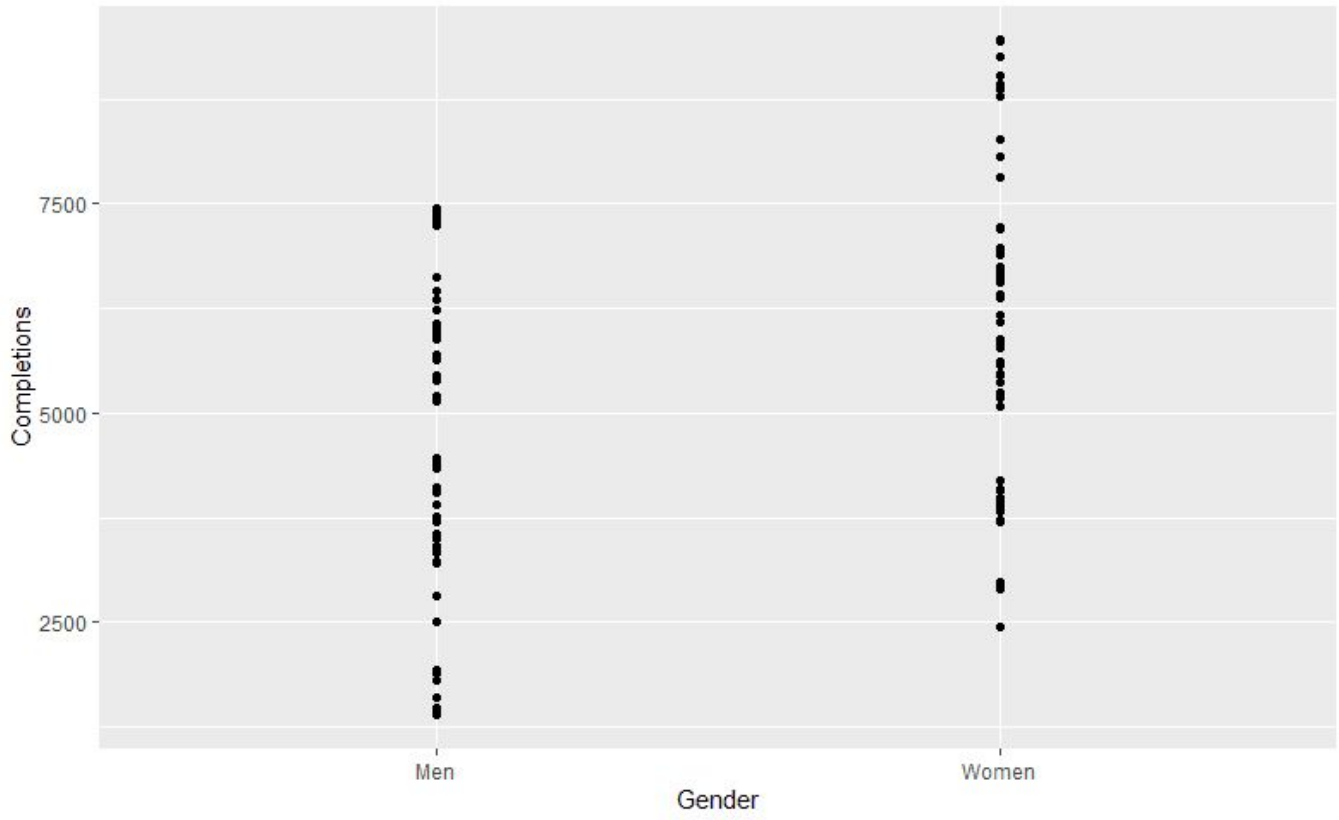
With the boxplot indicating that the mean Women completions might be different than the Men completions, we began to see the trend in our data and became interested in the statistical significance of this discovery.

Graphical Representation

To understand our data better we must visualize the data to help learn its trends. Below is a histogram of Completion data among Men and Women. The “x” axis is the number of completions recorded in the dataset, and they are divided by ranges of frequencies. Then, the “y” axis is the percentage of total completions that a certain range occupies in the dataset. This bar chart of frequencies (or histogram) shows us that the Women tend to produce completion ranges that appear more in higher amounts of completions, leading us to believe preliminarily that the amount of completions will differ among Women and Men. Of course, we will check that later on in our statistical tests, but for now, observe the histogram:



We also wanted to provide another graphical description of our data, so we used “ggplot2” in R to produce a Completion distribution. We used a “geom_point” command to display our observations for Completions among Men and Women. Each dot represents a college that contributed to our aggregate Completion means. Not only does this give an idea of the spread of the observations, but it can be used as another tool to identify outliers. This plot confirms that we do not see any outliers in our data.



Hypothesis Test

μ_W = True mean Women college graduates

μ_M = True mean Men college graduates

Null Hypothesis **(H_o): $\mu_W = \mu_M$;** The true mean Women college graduates is the same as the true mean Men college graduates.

Alternate Hypothesis (**H_a): $\mu_W \neq \mu_M$;** The true mean Women college graduates is different from the true mean Men college graduates.

We will use a significance level of 0.10 ($\alpha = 0.10$) for our hypothesis test.

Once we divide our data into subsets, we need to check the assumptions. We must ensure that the (1) samples are independent, (2) come from a simple random sample, and (3) are large enough (sample sizes both greater than 30). In addition, we must decide (4) whether we will use a pooled t-test or non-pooled t-test. First, though, we must check our assumptions:

Assumptions

1. The samples from men and the samples from women are obviously independent because they are two separate genders. For the purpose of this study, it is impossible to be both a man and a woman at the same time, so each individual case will fall into one of the two genders. Therefore, this assumption is met.
2. When we gathered our data from datausa, it was given that the data was collected from a simple random sample, therefore, this assumption is met.

3. The sample size for men is 60, and the sample size for women is also 60. 60 is greater than 30 (the minimum number that will satisfy this assumption), therefore, this assumption for "large samples" is met.
4. To determine whether we will use a pooled t-test or non-pooled t-test, we must divide the larger standard deviation of the two sets by the smaller standard deviation. If this result is greater than 2, we will assume that the population standard deviations *are not* approximately equal and we will use a non-pooled t-test. If this result is less than 2, we will assume that the population standard deviations *are* approximately equal and we will use a pooled t-test. So let's make this calculation.

    Std. Dev. of women = 1819.965

    Std. Dev. of men = 1875.657

1875.657 / 1819.965 = **1.031**, which is less than 2, therefore we can assume that the population standard deviations are approximately equal.

We will use a **pooled t-test.**

On the TI-84, we will perform a 2-SampTTest with the following inputs:

$\bar{X}_W$: 5847 $\bar{X}_M$: 4507

S_W: 1819.965 S_M: 1875.657

N_W: 60 N_M: 60

μ_W: $\neq \mu_M$

Pooled: Yes

Hypothesis Test Conclusion

Test statistic: 3.972

P-value: 0.000123

Decision: Since (P-value = 0.000123) < ($\alpha = 0.10$), we reject null hypothesis (or data do support alternative).

Conclusion: At 10% significance level, data provide sufficient evidence to conclude that the true mean Women college graduates is different from the true mean Men college graduates.

Confidence Interval

Since (above) we performed a "Two" tailed test, we need to use a confidence interval. For the hypothesis test, we used

"Null Hypothesis **(H_o):** $\mu_W = \mu_M$ **;** The true mean Women college graduates is the same as the true mean Men college graduates.

Alternate Hypothesis (**H_a):** $\mu_W \neq \mu_M$ **;** The true mean Women college graduates is different from the true mean Men college graduates."

We can rewrite this as

Null Hypothesis **(H_o):** $\mu_W - \mu_M = 0$; The true difference between Women college graduates and Men college graduates is zero (there is not a difference).

Alternate Hypothesis (**H_a):** $\mu_W - \mu_M \neq 0$; The true difference between Women college graduates and Men college graduates is not zero (there is a difference).

We used a significance level of 0.10 above, so we will find a 90% confidence interval.

To perform this, we must once again check that the assumptions are met. However, the assumptions are exactly the same as the assumptions for the hypothesis test (see p. 7-8), so we can proceed and find a **pooled t-interval**.

On the TI-84, we will perform a 2-SampTInt with the following inputs:

$\bar{X}_W$: 5847

$\bar{X}_M$: 4507

S_W: 1819.965

S_M: 1875.657

N_W: 60

N_M: 60

C-level: 0.90

Pooled: Yes

Confidence Interval Conclusion

Confidence Interval: (780.63, 1899.4)

Decision: Since 0 is outside this interval (780.63, 1899.4), we reject the null hypothesis (or data do support alternative).

Conclusion: At 10% significance level, data provide evidence to say that there is a difference between Women college graduates and Men college graduates. We are 90% confident that the true difference between Women college graduates and Men college graduates is between 780.63 and 1899.4. Since 0 is not in this interval, there is clearly a difference between the two.

Comparison of Hypothesis Test and Confidence Interval

We performed the hypothesis test and the confidence interval under the same conditions:

$\alpha = 0.10$ and $1-\alpha = 0.90$ (0.10 significance level and 90% confidence interval)

Because we did this, we can compare the results of the two performances easily.

With the same conditions, both the hypothesis test and the confidence interval resulted in the same decision to the given problem. We rejected the null hypothesis in both scenarios to say that data provide evidence to say that the true mean Women college graduates is different from the true mean Men college graduates.

Conclusion and Findings

From our Hypothesis Test and Confidence Interval we have proven with 10% significance that there is a difference between the amount of Women and the amount of Men that completed college in the year that this data was collected. With perspective to the world population, we know that the ratio of all humans among gender is not 50:50, so this conclusion makes sense. We loved being able to use R to graph our data and it gave us a great tool to put our data in perspective. We look forward to using R more in future statistics classes!