Deliverable Two

Nathaniel Jones

#### **Background Information and Variable Description**

Does the proportion of federal loan borrowers differ at institutions where the majority of undergraduate students are Pell recipients? The data used to test this hypothesis were two columns of data described in the table below. The variable PCTPELL is the percent of undergraduate students receiving a Pell grant. This variable was used to create a categorical variable with two levels separating institutions into the following groups:

* Institutions with more than 50% of their undergraduate population receiving a Pell grant will be referred to as having a majority Pell population and given a value of 1.
* Institutions with 50% or less of their undergraduate population receiving a Pell grant will be referred to as having a minority Pell population and given a value of 0.

**Variable Definitions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Definitions** | **Data type** | **n** |
| *PCTPELL* | The percent of undergraduate students receiving a Pell grant | Quantitative | 4423 |
| *PCTFLOAN* | The percent of undergraduate students borrowing a federal loan | Quantitative | 4423 |
| *PCTPELL\_CAT* | Binary indicator for the institutions that have more than 50% of their undergraduate population receiving a Pell grant. | 0 – <=50%  1 – >50% | 2807  1616 |

This report compared the percent of undergraduate students borrowing a federal loan by the grouping of institutions created above. In this report, seven tests were performed to determine if the two groups of institutions differ in the percentage of federal loan borrowers. These tests were the following: Kolmogorov-Smirnov, Ratio Mean Difference, Wilcoxon Rank-sum, t-test, permutation test on the difference in two population means, permutation test on the difference in two population medians, and permutation test on the difference in two population trimmed means.

**Hypothesis**

|  |  |
| --- | --- |
| *Claim:* | Institutions with a minority Pell population differ in the percentage of federal loan borrowers than institutions with a majority. |
| *Variables:* | PCTPELL\_CAT and PCTFLOAN |

#### **Data Visualization and Check for Normality**

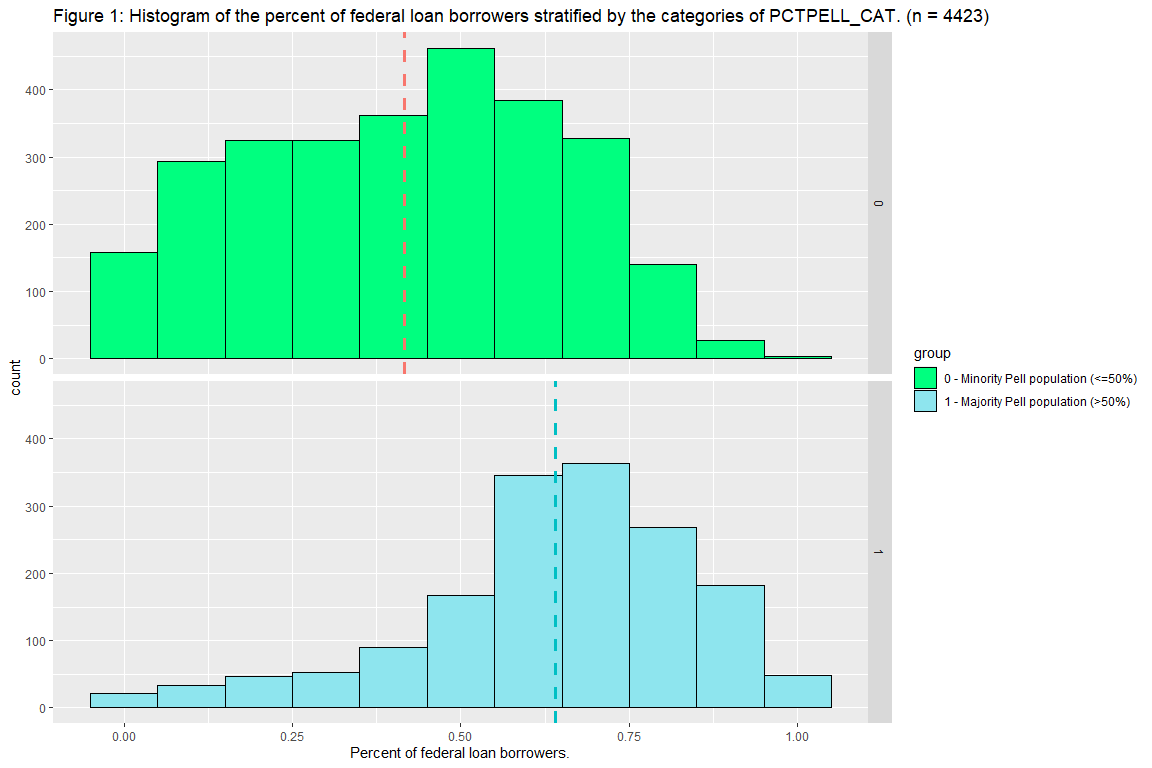
To begin, a categorical variable with two levels was created from the variable PCTPELL. This variable, labeled PCTPELL\_CAT, was created by assigning the value of 1 to all institutions with more than 50% of the undergraduate population receiving a Pell grant. Inversely, institutions with 50% or less of their undergraduate population receiving a Pell grant were assigned a value of 0. It was found that 642 observations for the variable PCTFLOAN had a value of zero. After further research, a conclusion was made to remove these observations. This conclusion was made after information from this [URL: http://www.collegescholarships.org/loans/community.htm](http://www.collegescholarships.org/loans/community.htm) was found. It was found that many community college students mistakenly believe that they are ineligible to receive financial aid and fail to fill out the Free Application for Federal Student Aid (FASFA).

**DESCRIPTIVE STATISTICS:**

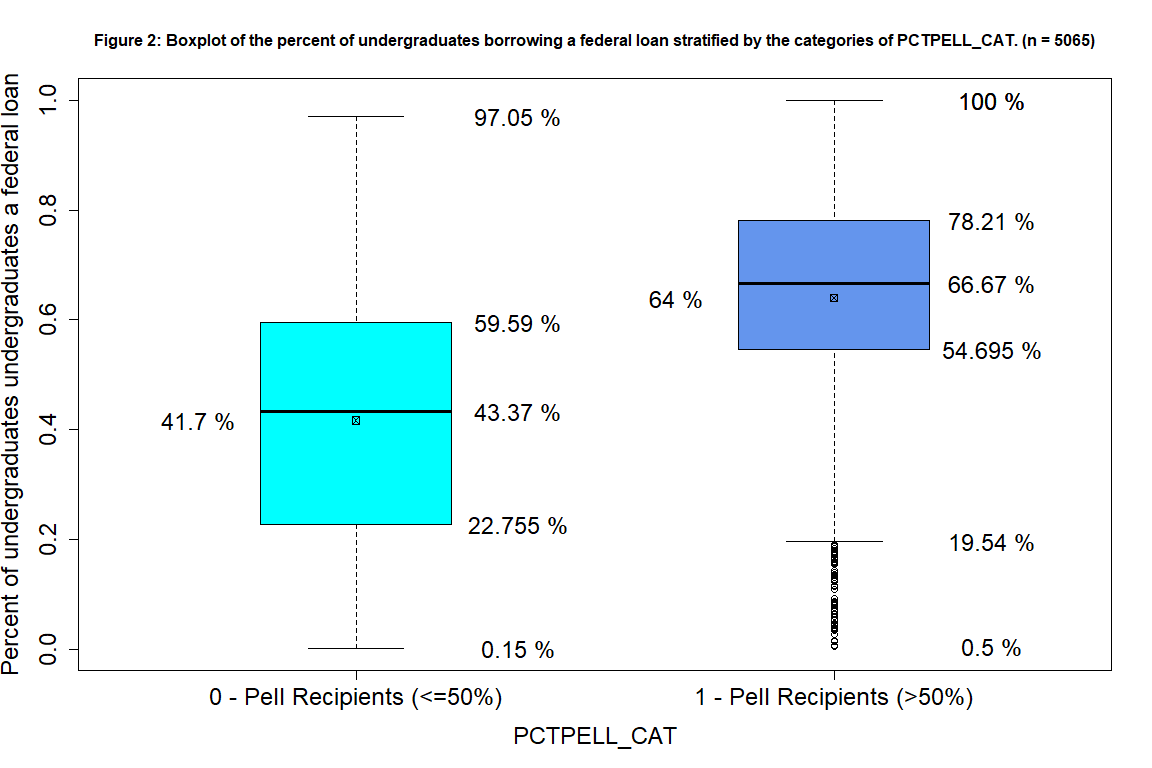
Descriptive Statistics

|  |  |  |  |
| --- | --- | --- | --- |
| Group 0 | PCTFLOAN | Group 1 | PCTFLOAN |
| Min | 0.0015 | Min | 0.0048 |
| Median | 0.4337 | Median | 0.6667 |
| Mean | 0.4168 | Mean | 0.6404 |
| Max | 0.9705 | Max | 1 |
| Missing | 0 | Missing | 0 |
| n | 2807 | n | 1616 |

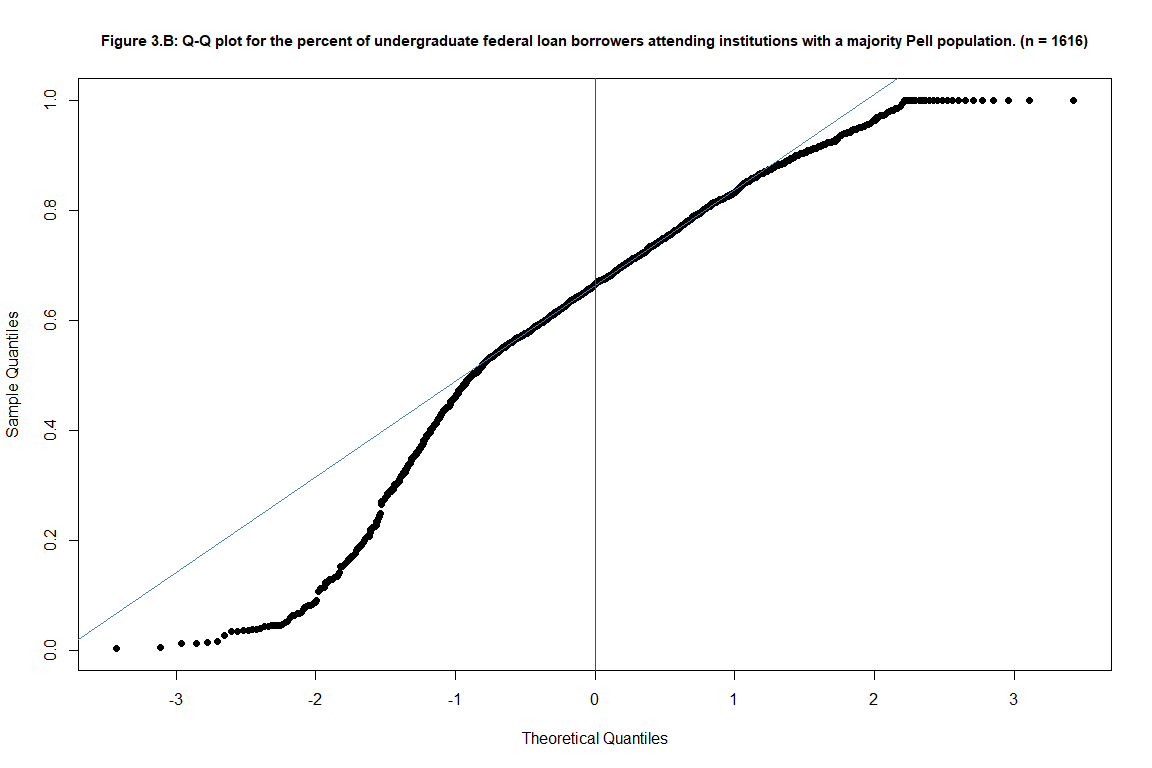
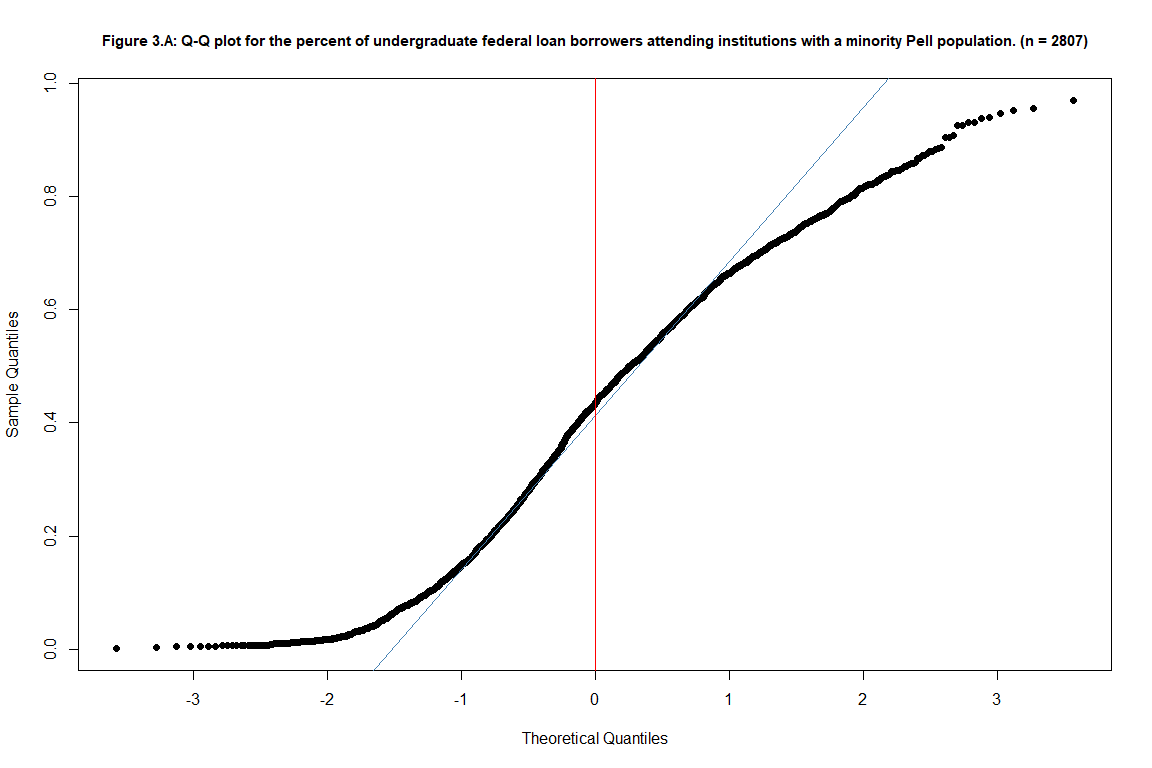
**HISTOGRAM:** A histogram for the percent of federal loan borrowers was created for the different PCTPELL\_CAT categories. Figure 1 displays this histogram. To help visualize the results, two lines were placed on these graphs to represent the mean of each group. The red line represents the mean percent of federal loan borrowers at an institution with a minority Pell population while the blue line represents the mean percent of federal loan borrowers at an institution with a majority Pell population. One can see a visual shift in the mean between the two categories.



**BOXPLOT:** Figure 2 displays the stratified boxplot of the percent of federal loan borrowers for the two categories of PCTPELL\_CAT. From this plot, one can visually see a difference in the percent of federal loan borrowers at institutions with a majority Pell population. The mean percent of federal loan borrowers at an institution with a majority was 55.2% while at institutions with a minority had 36.7%. Notably, the lower bound of the interquartile range for institutions with a majority Pell population is more than double the lower bound at institutions with a minority.



**NORMALITY CHECK:** With no clear visual evidence of normality coming from the histogram or boxplot, a Q-Q plot for each category was created and a Shapiro-Wilk’s normality test was performed. Figure 3.A corresponds to the percent of undergraduate students borrowing a federal loan at institutions with a minority Pell population while figure 3.B corresponds to institutions with a majority. Both figures show strong deviation from the normal line at the tails, suggesting that neither group came from a normal population. The results of the Shapiro-Wilk normality test gave further evidence that the data is skewed. Since both p-values (0 for category zero and 0 for category one) were less than the chosen level of confidence (5%), one can conclude that the data may not have come from a normally distributed population.



**Notation used in the hypotheses**

|  |  |
| --- | --- |
| **Label** | **Definition** |
|  | The mean percent of federal loan borrowers at institutions with a minority Pell population. |
|  | The mean percent of federal loan borrowers at institutions with a majority Pell population. |
|  | The CDF of percent of federal loan borrowers at institutions with a minority Pell population. |
|  | The CDF of percent of federal loan borrowers at institutions with a majority Pell population. |
|  | The variation in the percent of federal loan borrowers at institutions with a minority Pell population. |
|  | The variation in the percent of federal loan borrowers at institutions with a majority Pell population. |

#### **Results and findings**

      A difference in the percent of federal loan borrowers at institutions with a majority Pell population and a minority Pell population was found. Seven tests were performed to test for a difference in distribution, variation, and center. The notable results are:

* The permutation test on the difference in medians showed that there is a 0.0546% chance of receiving a difference that is equal to or greater than the observed difference of 23.3%.
* The data showed from the Wilcoxon Rank-sum test that the true difference in the center of the two distributions is between 0.2143 and 0.2425.
* The results of the ratio mean difference test found that institutions with a minority Pell population are 1.2371 times more variable than institutions with majority.

Notably the t-test, the permutation test on the difference in two population means, and the permutation test on the difference in two population trimmed means found very similar results to the permutation test on the difference in two population medians. The test on the medians was chosen due to the skew seen in the Q-Q plot in a previous section.

#### **Conclusion from the Kolmogorov-Smirnov test on the difference of shape, center, and spread.**

Hypothesis

|  |  |
| --- | --- |
| H0: | The shape, center, and spread of the two distributions are not different. |
|  |  |
| Ha: | The shape, center, and spread of the two distributions are different. |
|  |  |

      The Kolmogorov-Smirnov test showed that there is a 0% chance of receiving a test statistic that is 0.4335 or more. Such a small chance suggests that there are very few permutations that are greater than the observed test statistic. With this information there is enough evidence to conclude that at least one of the properties, either the shape, the center, or the spread, of the distribution is different.

#### **Conclusion from the RMD (Ratio Mean Difference) test**

Hypothesis

|  |  |
| --- | --- |
| H0: | The variation in the distribution of the two groups is not different. |
|  |  |
| Ha: | The variation in the distribution of the two groups is different. |
|  |  |

      The results of the ratio mean difference test showed, after 1e+06 iterations, a test statistic of 1.2371. This test statistic means that institutions with a minority Pell population are 1.2371 times more variable than institutions with a majority. There is a 0% chance of receiving a test statistic of 1.2371 or more in two-tails. With this information, we can conclude that the variation in the percent of federal loan borrowers at institutions with more than 50% of their undergraduate population receiving a Pell grant is different from institutions with 50% or less of their undergraduate population receiving a Pell grant.

#### **Conclusion from the Wilcoxon Rank Sum Test on the Difference of Two Population Centers**

Hypothesis

|  |  |
| --- | --- |
| H0: | The distributions of the ranked data do not have different centers. |
|  |  |
| Ha: | The distributions of the ranked data have different centers. |
|  |  |

      The data showed a Wilcoxon test statistic (3500067) with a very small p-value (0). This suggests that there was a very small number of rank sums that were larger than the observed rank sum of 3500067. With this information, there is enough evidence to conclude that the distributions of the ranked data have different centers. At a confidence level of 95%, the data showed that the true difference in the center of the two distributions is between 0.2143 and 0.2425. This test shows that there is a 0% chance observing a difference in the center of the two groups’ distributions that is greater than 22.8328% or more.

#### **Conclusions from the Parametric t-test on the difference of two population centers**

Hypothesis

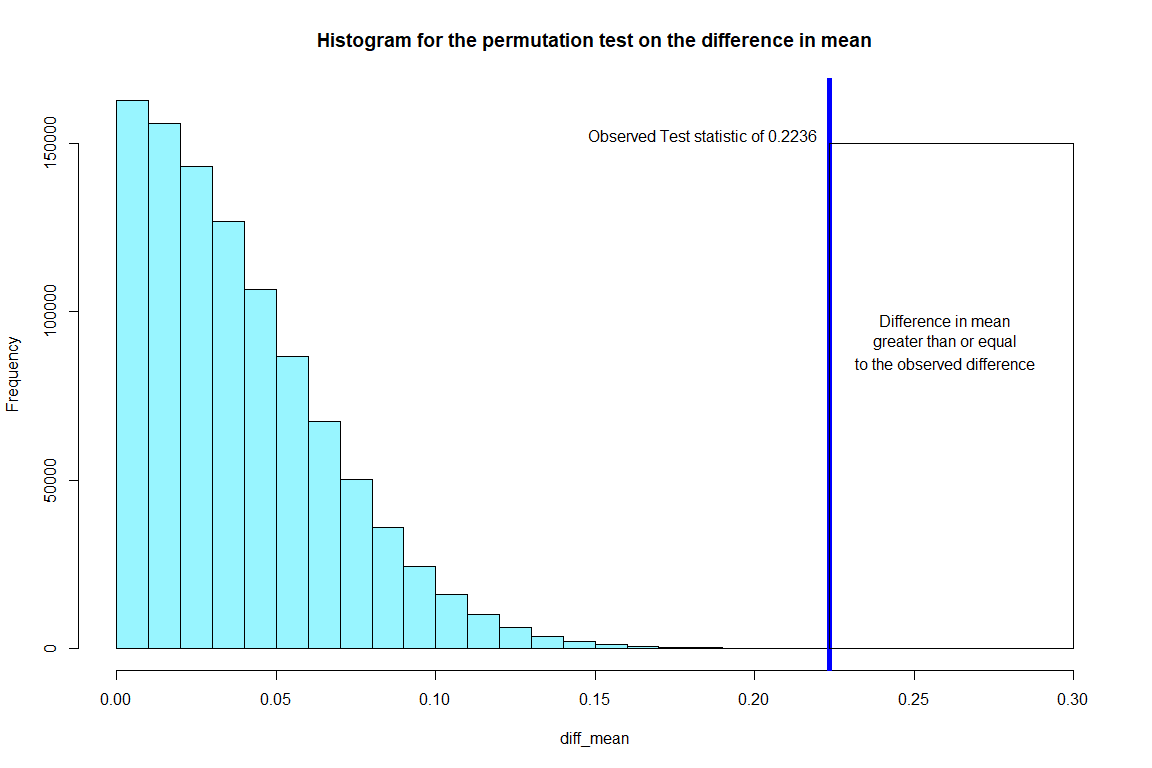
|  |  |
| --- | --- |
| H0: | There is not a difference between the means of the two groups. |
|  |  |
| Ha: | There is a difference between the means of the two groups. |
|  |  |

      The results of this parametric t-test showed that the consistency of the data with the null hypothesis was 0%. It was observed that institutions with a minority Pell population had a mean percent of 0.4168 federal loan borrowers while institutions with a majority had a mean percent of 0.6404 federal loan borrowers. Thus, there is a 0% chance of observing the difference in the means of the two groups to be 0.2236 or more when the two groups have the same mean test score. Since this p-value shows that it is very unlikely to receive a difference in the mean larger than the one observed in this sample, there is enough evidence to conclude that there is a difference between the mean of the two groups. In fact, at a confidence level of 95%, the data shows that institutions with a majority Pell population have between 21.06% and 23.66% more undergraduate federal loan borrowers than institutions with a minority.

#### **Conclusions from the Permutation Test on the Difference of Two Population Means**

Hypothesis

|  |  |
| --- | --- |
| H0: | The distributions of the two groups are not different. |
|  |  |
| Ha: | The distributions of the two groups are different. |
|  |  |



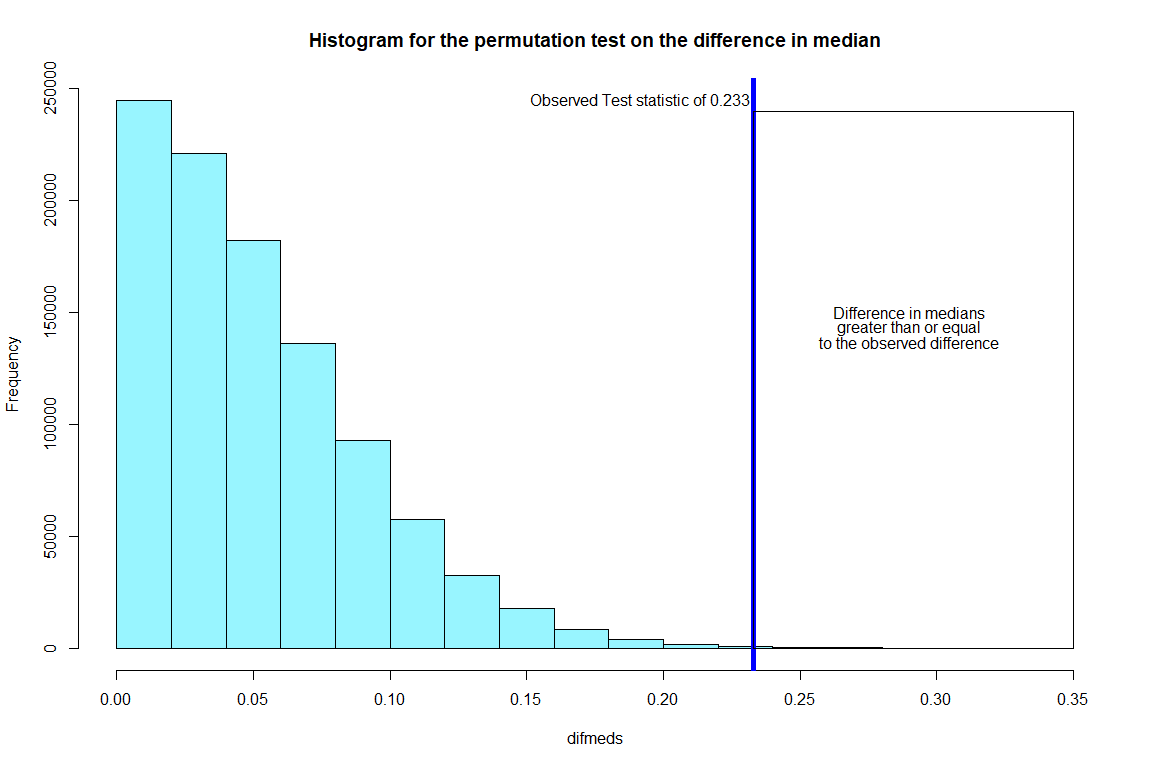
      After 1e+06 simulations at a confidence level of 99%, the data showed that the true p-value is anywhere from 0% to 0.0011%. Using the estimate from the sample, there is a 5e-04% chance that one will observe that the difference in the mean percent of federal loan borrowers at institutions with a majority of their undergraduates receiving a Pell grant is 22.3619% more than institutions with a minority of their undergraduates receiving a Pell grant when the two groups have the same mean test score. This test suggests that the distributions of the two groups are different and the values that are found most typically are group specific. Specifically, institutions with a majority Pell population typically have a higher percentage of federal loan borrowers than institutions with a minority.

The histogram above displays the results of the permutation test on the difference in means. The blue line is placed on the observed test statistic which is the difference between the means of the two groups. The box to the right of the blue line contains all the permuted test statistics that were greater than or equal to the observed. As one can see, the box is nearly empty. This visually shows how extreme the observed sample is in comparison to when the two groups have the same mean test score.

#### **Conclusions from the Permutation Test on the Difference of Two Population Medians**

Hypothesis

|  |  |
| --- | --- |
| H0: | The distributions of the two groups are not different. |
|  |  |
| Ha: | The distributions of the two groups are different. |
|  |  |



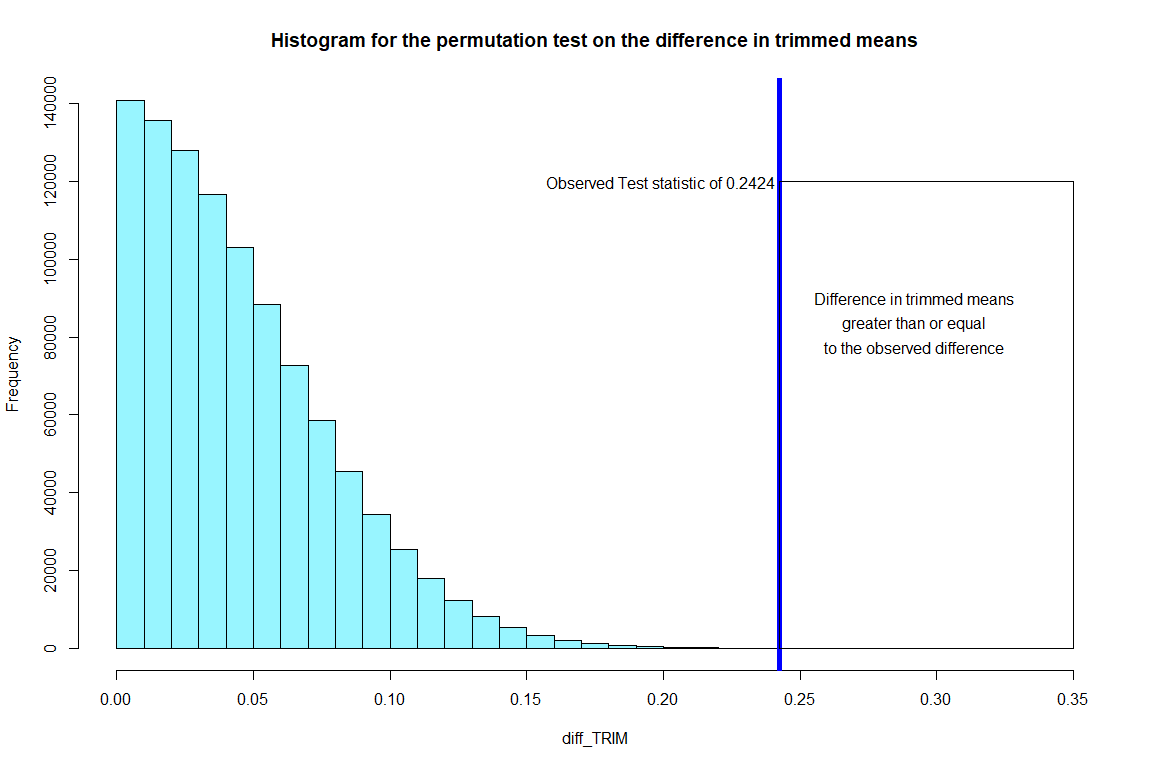
      After 1e+06 simulations, the data showed that the true p-value is 0.000589. This suggests that there is a 0.0589% of chance of observing a difference in the medians of the two groups that is equal to 23.3% or greater. With this information, there is evidence to conclude that the distribution of the percent of federal loan borrowers at institutions with a majority Pell population is different from the distribution of the percent of federal loan borrowers at institutions with a minority population.

The histogram above is of similar form to the one under the permutation test for the mean. This plot has the blue line placed in the position of the observed test statistic with a box outlining the region of permuted test statistics that are greater than or equal to the observed. Since very few permuted test statistics were more extreme than the observed test statistic, there is a difference between the distributions of the two groups.

#### **Conclusion from Permutation test on the difference of two population trimmed means**

Hypothesis

|  |  |
| --- | --- |
| H0: | The distributions of the two groups are not different. |
|  |  |
| Ha: | The distributions of the two groups are different. |
|  |  |



      This test is similar to the permutation test for the difference in group means except 15% of the extreme observations are ‘trimmed’ before the test. This test found that there is a 0.0021% of chance of observing a difference in the trimmed means of the two groups that is equal to 24.24% or greater. This suggests that the distributions of the two groups are different even if we trim 15% of the most extreme observations before running the test.

The histogram displayed above is in the same format as the last two permutation tests. The blue line represents the observed test statistic and the box to the right of this line contains the region of permuted test statistics that are greater than or equal to the observed test statistic. One can see that the observed statistic is very extreme compared to the majority of permuted statistics, suggesting that the distributions are very different.