Shawn Koohy

4 March 2022

Dr. Davis

MTH332

Test Scores and Teaching Styles, Who Will Benefit More, Does It Matter?

Abstract

In the education system pre-tests are an evaluation on how much a student knows about a topic before they have taken the class. These pre-test score can help a teacher or professor focus more on issues or topics related to the class material that less people know or understand. Post-tests on the other hand can tell us how much a student has learned from the class they've just taken. In this report we are going to be investigating the gains scores of different data sets of students. What this means is that we will be looking at how much students scores increased from when they took their post-test vs when they took their pre-test. In our first data set we will be looking to see if students who score below the median pre-test score tend to have a larger gain score versus those who score at or above the median pre-test score. Our goal is to determine if there is a statistically significant difference between the mean gain scores. In our second data set we will be understanding whether a traditional or a pilot style of teaching will lead to better overall gain scores. For this data set our first group will be the traditional set of scores and the second group being the set of scores for the pilot style of teaching. The null hypothesis for both data sets are that there is no difference between the groups. In both data sets we will be finding descriptive statistics, analyzing plots and histograms along with implementing and conducting the concept of bootstrapping to find the corresponding p-values, this will all be explained in the "Methods" section of the report. From all of this we will be able to conclude that students who score lower than the median on a pre-test tend to have larger gains scores than those who scored above or at the median on pre-tests. We will also see that the pilot style of teaching is, from the evidence we have, a slightly more efficient and effective style of teaching than using traditional methods.

Introduction & Background

In our first data set we are going to be looking at the pre-test and post-test percentage grades for a group of 155 pre-service teachers taking a college mathematics course. With this group we are going to see if there are statistically significant differences between two groups of students taking the course. This first group, Group A, will contain all who scored less than the median pre-test score and Group B will contain all who either scored the same or greater than the pre-test median. Our end goal is to see if there is a statistical difference in the gains for Group A (from the pre-test to post-test) versus the gains in Group B. Is the average gain for Group A statistically different than the average gain for Group B? Theoretically the assertion by most is that those who scored equal to or greater than the median (Group B) will not have as much gain as they are already closer to the max score, 1, than Group A is.

In our second data set we have a similar situation. Here we have pre-test and post-test scores of students coming from two different lecture styles, traditional and pilot, at a mid-west university. The traditional group was taught by a lecture style method using existing texts. The pilot group was taught using modified texts and student engagement was much more emphasized. There was a total of 104 students taught by the pilot style and a total of 93 student taught in the traditional style. Our issue to examine is whether there is a statistical difference in the mean gain scores between the traditional teaching style and the pilot teaching style. We will be calling the traditional style “Group A”, and the pilot style “Group B” for this data set.

Methods

Let’s start with our first data set. By using Mathematica, we will find the descriptive statistics, mean, median, Q1, Q3, minimum, maximum, kurtosis, skewness, and standard deviation of the total pre-test and post-test scores. After that we will split the data into 2 main groups, Group A containing all students who scored less than the median pre-test score (0.43) and Group B, those who scored equal to or greater than the median pre-test score. Group A now has its pre-test scores and the corresponding post-test score and same with Group B. Now that we have our two groups split, we will find the descriptive statistics of Group A pre-test, Group A post-test, Group B pre-test and Group B post-test. All of this will give a good intuition and an

idea on what our data looks like, its behavior and how it varies from group to group. Once we have our descriptive statistics out of the way, we need to find one of the most important aspects of this report, the gain scores. We define the gain of each group to be given by the following:

$$Gain = \frac{[post - test\ score] - [pre - test\ score]}{1 - [pre - test\ score]}$$

We will calculate the gain of each set of pre-test and post-test score for Group A and Group B then plot a Histogram, Smooth Histogram, overlap them and compare. After comparing the groups gains, we will start to do a process called bootstrapping. In this process we are going to examine the difference in means from what we will call PseudoGroupA and PseudoGroupB. First, we are going to combine the gains score for Group A and Group B into one set named TotalGains, we can easily calculate the mean of the set since we know all the data points. The idea here is to, at random, choose 73 (the number of students who scored below the median pre-test score) gain scores, called PseudoGroupA, find the mean of that and then we can find the mean of the remaining 82 gain scores called PseudoGroupB by the following:

- $PS_A =$ PseudoGroupA and $PS_B =$ PseudoGroupB
- $\|PS_A\| =$ Size of PseudoGroupA and $\|PS_B\| =$ Size of PseudoGroupB
- $n =$ Size of TotalGains

$$Mean(PS_A)\left(\frac{\|PS_A\|}{n}\right) + Mean(PS_B)\left(\frac{\|PS_B\|}{n}\right) = Mean(TotalGains)$$

Then, solving for $Mean(PS_B)$

$$\left(\frac{n}{\|PS_B\|}\right) Mean(TotalGain) - \left(\frac{\|PS_A\|}{\|PS_B\|}\right) Mean(PS_A) = Mean(PS_B)$$

What we are going to do with this is in Mathematica we will repeat this process 1,000,000 times and append all those value into a set such that $L = \{Mean(PS_A) - Mean(PS_B)\}$. From this set we are going to see how many times out of a 1,000,000 this difference, $Mean(PS_A) - Mean(PS_B)$, is greater than $Mean(GG_A) - Mean(GG_B)$. Where $GG_A =$ Gains in Group A and $GG_B =$ Gains in Group B. This fractional difference will give us our p-value. A p-value of 0.05 or below means that there is a statistical significance in our data and the null hypothesis should either be rejected (for very low p-values) or considered suspicious/inconclusive for values under but close to 0.05. A p-value above 0.05 would mean that we wouldn't consider our distribution/data set to be suspicious compared to our null hypothesis or we could say our observations are possibly in-conclusive.

When it comes to our second data set of traditional and pilot teaching styles, we will be doing a similar process. The only difference will be that the entire traditional pre-test data set will be considered our Group A and the entire pilot pre-test data set will be considered our Group B. The same process of bootstrapping will be applied in this data set along with the creation of pseudo groups and a calculation of the p-value after 1,000,000 statistical computations have been made.

Results

First Data Set		
Measure \ Group	Pre-test Only	Post-test Only
Mean	0.436645	0.745548
Standard Deviation	0.176257	0.105892
Skewness	0.195122	-0.182917
Kurtosis	2.3225	3.23162
Minimum	0.06	0.41
Maximum	0.87	0.99
Median	0.43	0.75
Q1	0.3	0.68
Q3	0.57	0.81

First Data Set		
Measure \ Group	Group A Pre-test	Group A Post-test
Mean	0.28274	0.716438
Standard Deviation	0.0843975	0.102583
Skewness	-0.431088	-0.273416
Kurtosis	2.34841	3.60671
Minimum	0.06	0.41
Maximum	0.4	0.95
Median	0.3	0.73
Q1	0.2225	0.64
Q3	0.37	0.7825

Figure 1. Pre and Post Tests Only (Left) and Group A Pre and Post Tests (Right) (First Data Set)

First Data Set		
Measure \ Group	Group B Pre-test	Group B Post-test
Mean	0.573659	0.771463
Standard Deviation	0.111272	0.102573
Skewness	0.56178	-0.137542
Kurtosis	2.71024	2.84665
Minimum	0.43	0.47
Maximum	0.87	0.99
Median	0.57	0.77
Q1	0.47	0.7
Q3	0.63	0.85

First Data Set		
Measure \ Group	Group A Gains	Group B Gains
Mean	0.603735	0.4434
Standard Deviation	0.136665	0.262181
Skewness	0.0668998	-0.538629
Kurtosis	2.7296	3.12355
Minimum	0.325	-0.25
Maximum	0.928571	0.972973
Median	0.61194	0.462113
Q1	0.506849	0.26
Q3	0.690283	0.625

Figure 2. Group B Pre and Post Tests (Left) and Group A and B Gains (Right) (First Data Set)

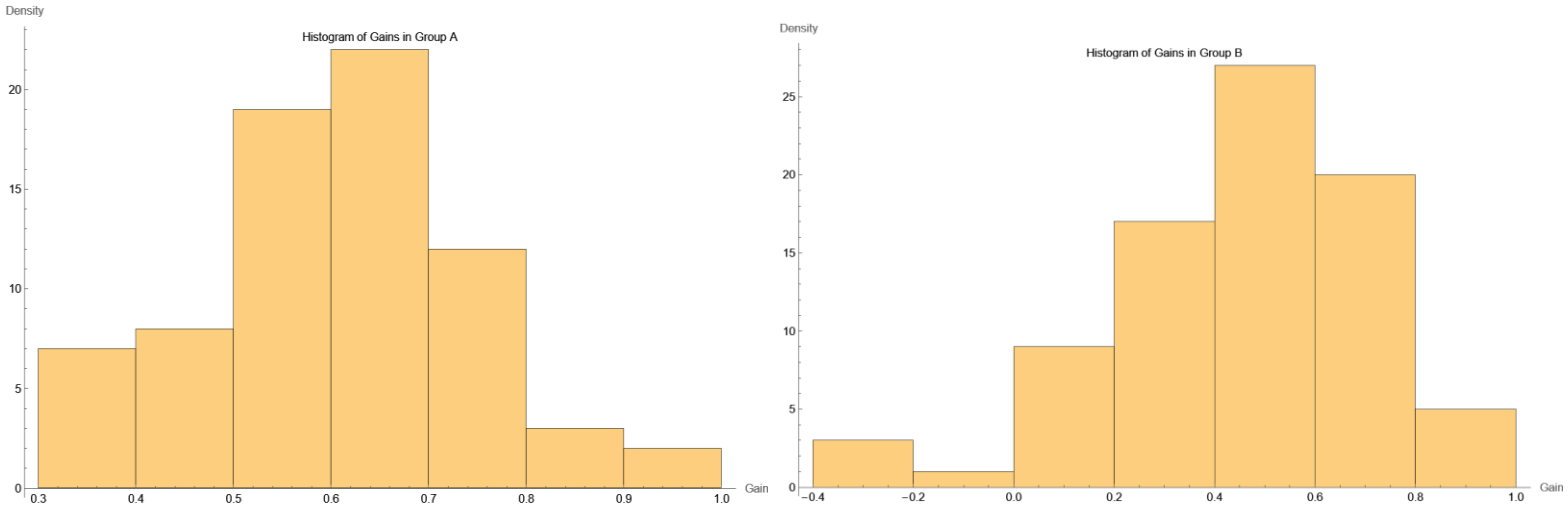

Figure 3. Histogram of Gains in Group A (Left) and Histogram of Gains in Group B (Right) (From First Data Set)

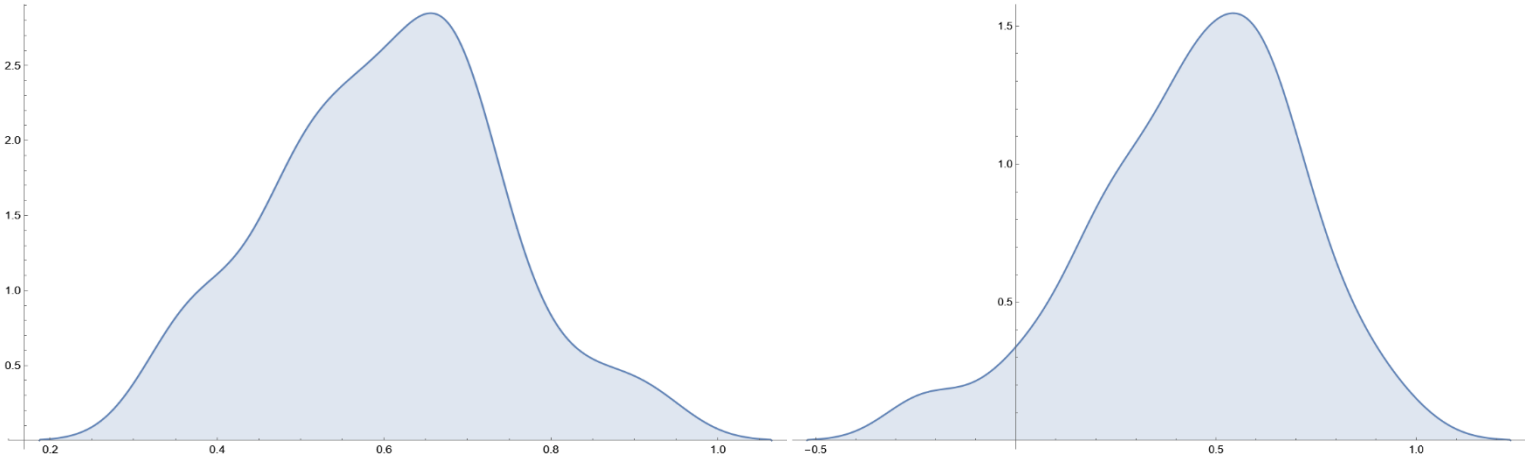

Figure 4. Smooth Histogram of Gains in Group A (Left) and Smooth Histogram of Gains in Group B (Right) (From First Data Set)

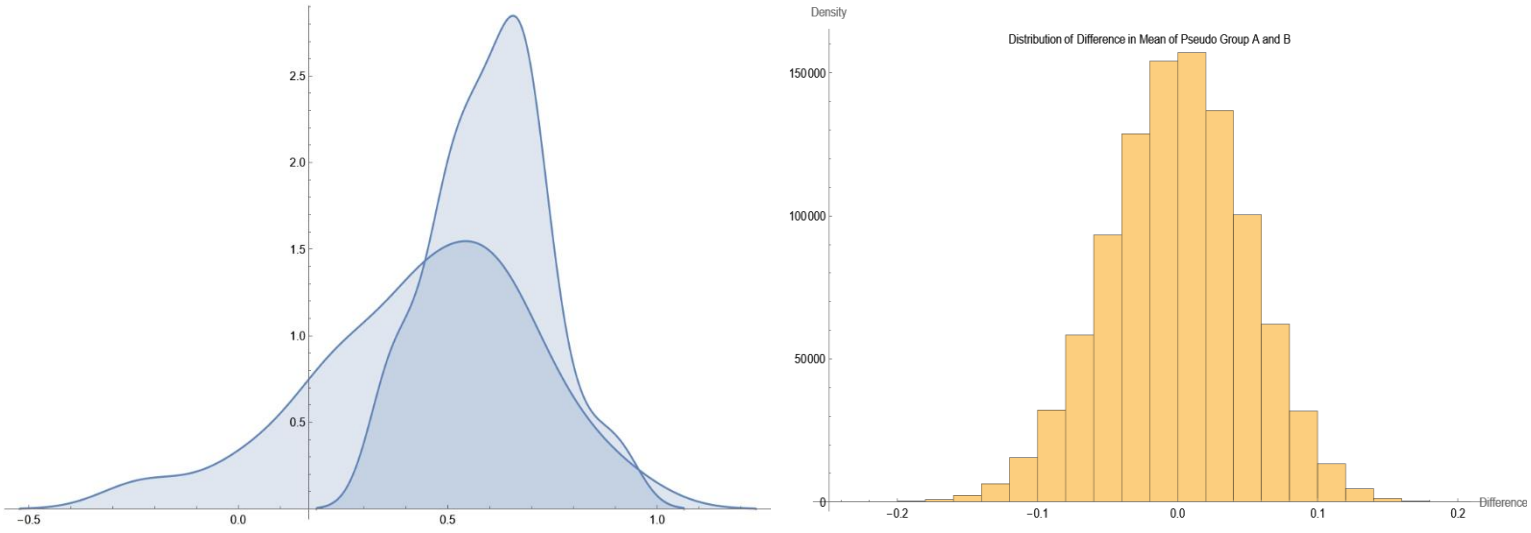

Figure 5. Overlap of Figure 8. (Left) and Histogram of Distribution Between the Difference in Means in Pseudo Group A and Pseudo Group B (Right) (From First Data Set)

Second Data Set		
Measure \ Group	Traditonal Gains	Pilot Gains
Mean	0.204277	0.246583
Standard Deviation	0.177572	0.164145
Skewness	0.235523	0.121544
Kurtosis	3.22476	3.43953
Minimum	-0.194444	-0.206897
Maximum	0.662338	0.648649
Median	0.194444	0.243715
Q1	0.0886076	0.142857
Q3	0.313624	0.37037

Second Data Set		
Measure \ Group	Traditonal Pre	Traditional Post
Mean	0.182903	0.349247
Standard Deviation	0.092614	0.162168
Skewness	0.468815	0.225482
Kurtosis	3.02651	2.77181
Minimum	0	0.05
Maximum	0.44	0.74
Median	0.19	0.35
Q1	0.12	0.23
Q3	0.23	0.4475

Figure 6. Traditional and Pilot Gains (Left) Traditional Pre and Post Tests (Right) (Second Data Set)

Second Data Set		
Measure \ Group	Pilot Pre	Pilot Post
Mean	0.207019	0.40125
Standard Deviation	0.106715	0.160525
Skewness	0.191962	-0.171716
Kurtosis	2.1078	2.1078
Minimum	0.02	0.09
Maximum	0.42	0.74
Median	0.19	0.40125
Q1	0.12	0.28
Q3	0.3	0.51

Figure 7. Pilot Pre and Post Test (Second Data Set)

P-Values	
First Data Set	0.000354
Second Data Set	0

Figure 8. Data Sets P-Values

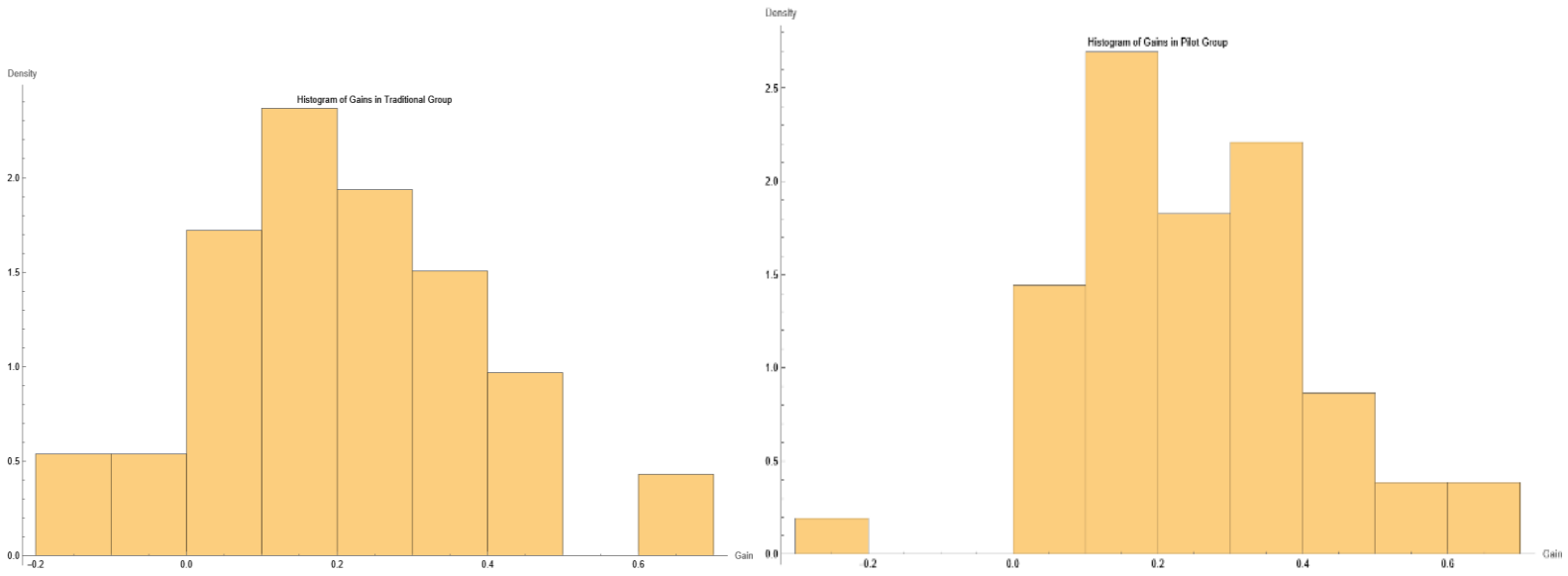

Figure 8. Histogram of Traditional Gains (Left) and Histogram of Pilot Gains (Right) (Second Data Set)

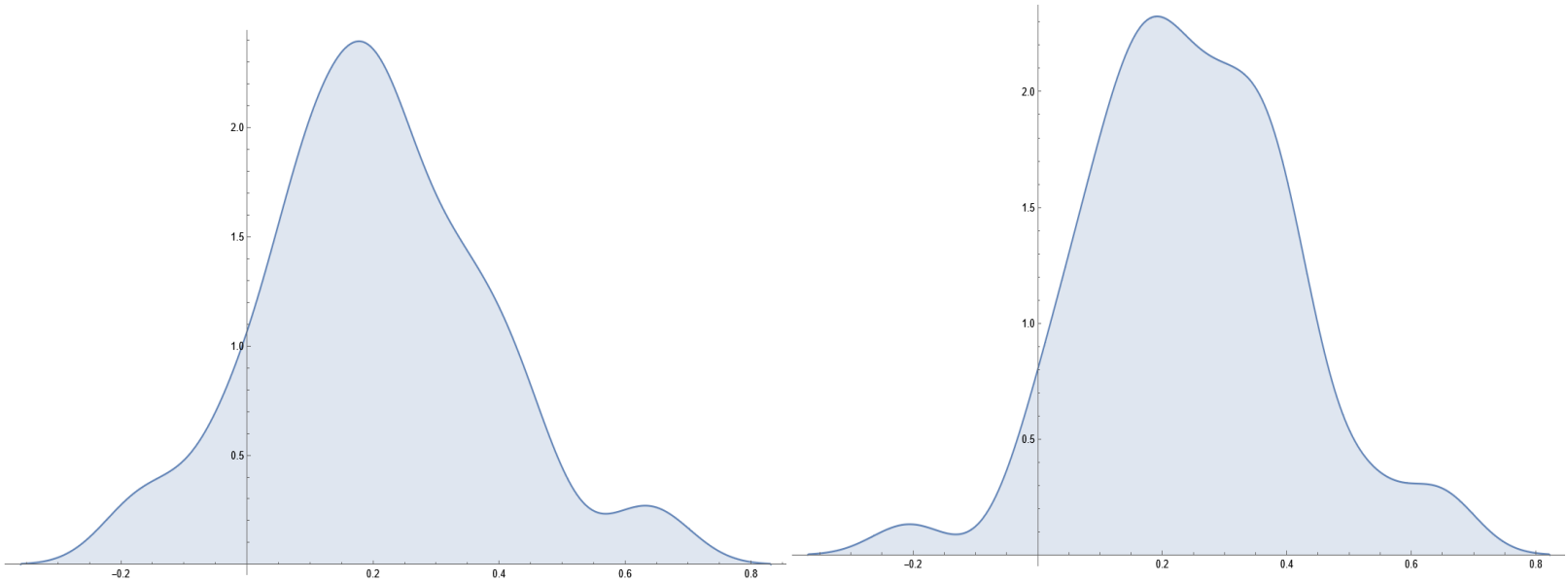

Figure 9. Smooth Histogram of Traditional Gains (Left) and Smooth Histogram of Pilot Gains (Right) (Second Data Set)

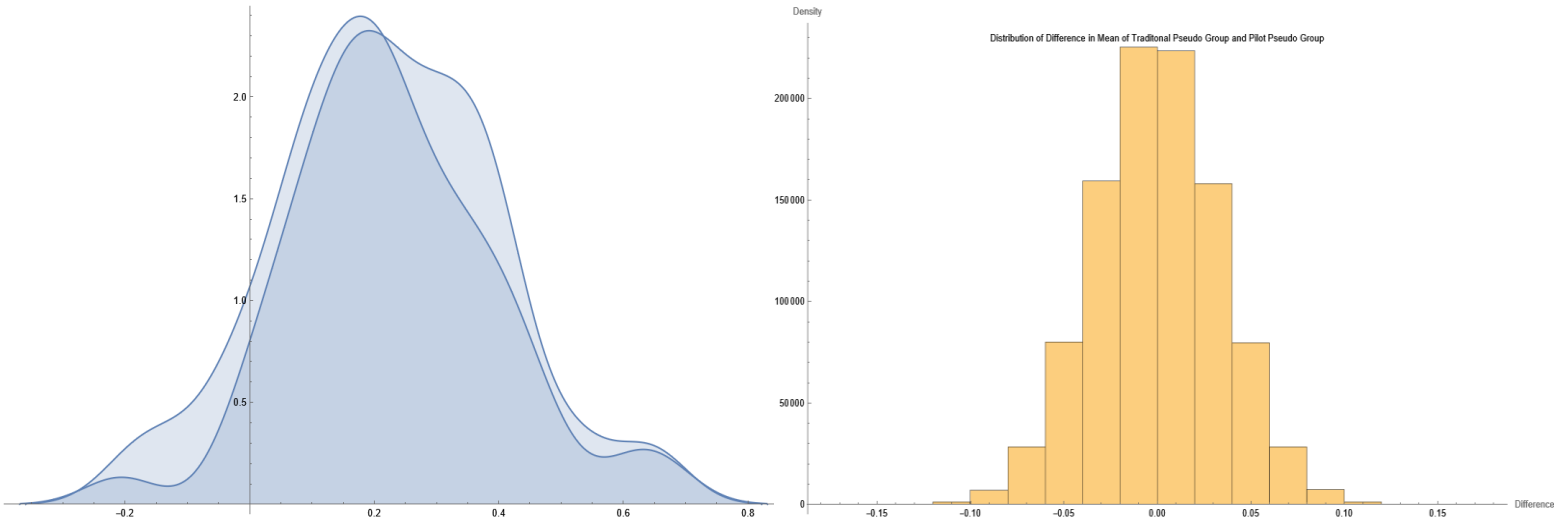

Figure 10. Overlap of Figure 10. (Left) and Histogram of Distribution Between the Difference in Means in Pseudo Traditional Group and Pseudo Pilot Group (Right) (Second Data Set)

Discussion & Conclusions

One of the most important factors in determining the answer to the question about both data sets, that is, is there a statistically significant difference between the groups is the p-value. The p-value for the first data set was 0.000354. Typical any p-value below or at 0.05 means we should be conscious and suspicion in our results and possibly reject the null hypothesis. The null hypothesis was that there is no statistically significant difference between the groups. With our method and what we got for a p-value we have enough evidence to strongly suggest that this null hypothesis is false and that there is a statistical difference between the groups. We can see in Figure 2. the specifics of our Group A and Group B gain scores. Group A has a larger mean, minimum, median, Q1 and Q3. When we look at the distribution of these gains in either Figure 3. or Figure 4. Group A tends to have a higher density of larger gains than in Group B. We can describe this analytically by saying that on average Group A will have larger gain scores than those in Group B.

When looking at our second data set, we want to assume the same null hypothesis, that there is no statistically significant difference between the traditional and pilot group. Clearly with a p-value of 0.0 we can make a strong conclusion to reject the null hypothesis. This means that there is a statistically significant difference between the Traditional and Pilot groups, what is this

difference? Examining the gains scores of the Pilot and Traditional Groups in Figure 6. we see that the Pilot Group has a larger, mean, Q1 and Q3 where both groups have a similar standard deviation, skewness, and kurtosis. In Figure 9. we can see a greater density of higher gain scores for the pilot group. Now, considering the comparison of our data, the distributions and p-value this statistical difference may not be a large or as the significance than what we saw in the first data set. Our Pilot group does show overall better performing student but not to a major degree. With this we can still conclude that there is a statistically significant difference, but it appears to not have as much of a larger impact as we have seen previously. Students taught in the Pilot style tend to understand and retain information slightly better and perform slightly better on tests versus those who were taught using traditional methods.

Putting together everything we have discovered we can make some larger conclusions. Our first conclusion is that student who score lower than the median on their pre-test are, on average, are more likely to make larger gains on their post-test. The second conclusion is that the Pilot group typical has an overall larger number of gains and better performance though, that difference is not something major and could be argued with further testing that there could be even less to no difference.

Overall teachers and professors catering to the needs of all students, those who perform poorly or greatly is the most important aspect of how one can learn more in a classroom environment. One important aspect of our data is that it may be up to personal preference when it comes to the teaching style that works for oneself. In our data, students who may have preferred the traditional style could’ve been place in the pilot group and visa-versa. This is something that could have significant statistical weight on our results but, with what data we have and what we know about the data our best statistical, and analytical skills were used to give a conclusive answer to our issue.

References

Nolan, D. A., & Speed, T. (2001). Stat labs: Mathematical statistics through applications. New York, NY: Springer.