Basic Inferential Analysis of Significant Test-Score Gains in College Mathematics

James P. Donohue

March 4, 2022

Abstract

This study is aimed to use basic inference to ascertain if there is a significant difference in gain scores between two groups of students: those who scored less than the pre-test median, and those who scored at or above the pre-test median score. A pre/post-test by design covers all of the topics which a student will be studying during a semester. The results of this study can aid the educators who provided us this dataset to come to a sound conclusion on the preparedness of their students, and the robustness of their education methods. For our first data set we are provided two columns, pre-test and post-test percentage grades (scaled to lie between 0 and 1) for a cohort of 155 pre-service teachers taking a college mathematics course. From these values a third column was derived for academic gains. For our second dataset, mathematics pre-test and post-test scores for a cohort of students in a mid-west university. The cohort of students is split into two groups: “traditonal” and “pilot”. Unsurprisingly, simply “eyeballing” the histograms for Groups A and B can tell us that the difference in gains is statistically significant. However, it’s hard to make any strong conclusions based on the distributions of the histograms of the “Pilot” and “Traditional” groups. “Eyeballing” these distributions is definitely not enough.

Background and Significance

A pre/post-test by design covers all of the topics which a student will be studying during a semester. While taking the pre-test at the beginning of a semester, students are not expected to know the answers to all of the questions; however, they should be expected to utilize previous knowledge to predict rational answers. When taking the same test called a post-test at the end of a semester, students should be expected to answer more questions correctly based on an increase in knowledge and understanding. Pre and post-tests are assessment tools for measuring the preparedness and performance of students. The goal of this study is to ascertain if there is a significant difference in gain scores between two groups of students: those who scored less than the pre-test median, and those who scored at or above the pre-test median score. The results of which can aid the educators who provided us this dataset come to a sound conclusion on the preparedness of their students, and the robustness of their education methods.

Methods

For our first data set we are provided two columns, pre-test and post-test percentage grades (scaled to lie between 0 and 1) for a cohort of 155 pre-service teachers taking a college mathematics course. From these values a third column was derived for academic gains, calculated by the provided formula: gain := ([post-test score] - [pre-test score])/(1- [pre-test score]). Our next step was to split the cohort of 155 into two groups, Group A: those who scored below the median score, and Group B: those who scored greater than or equal to the median test score for the entire cohort of 155 students. We were able to generate this split using the third column we calculated for gains previously. To analyze the difference in average gain of both groups, we then generated the descriptive statistics for both Group A and Group B, and created histograms to visualize these findings. By merely "eye-balling" the charts and statistics, one can easily determine whether or not the difference in gain scores for both groups is statistically significant. To observe the significance of both groups' means, a more thorough bootstrapping approach was utilized. For 1,000,000 iterations, a t-test algorithm was constructed.

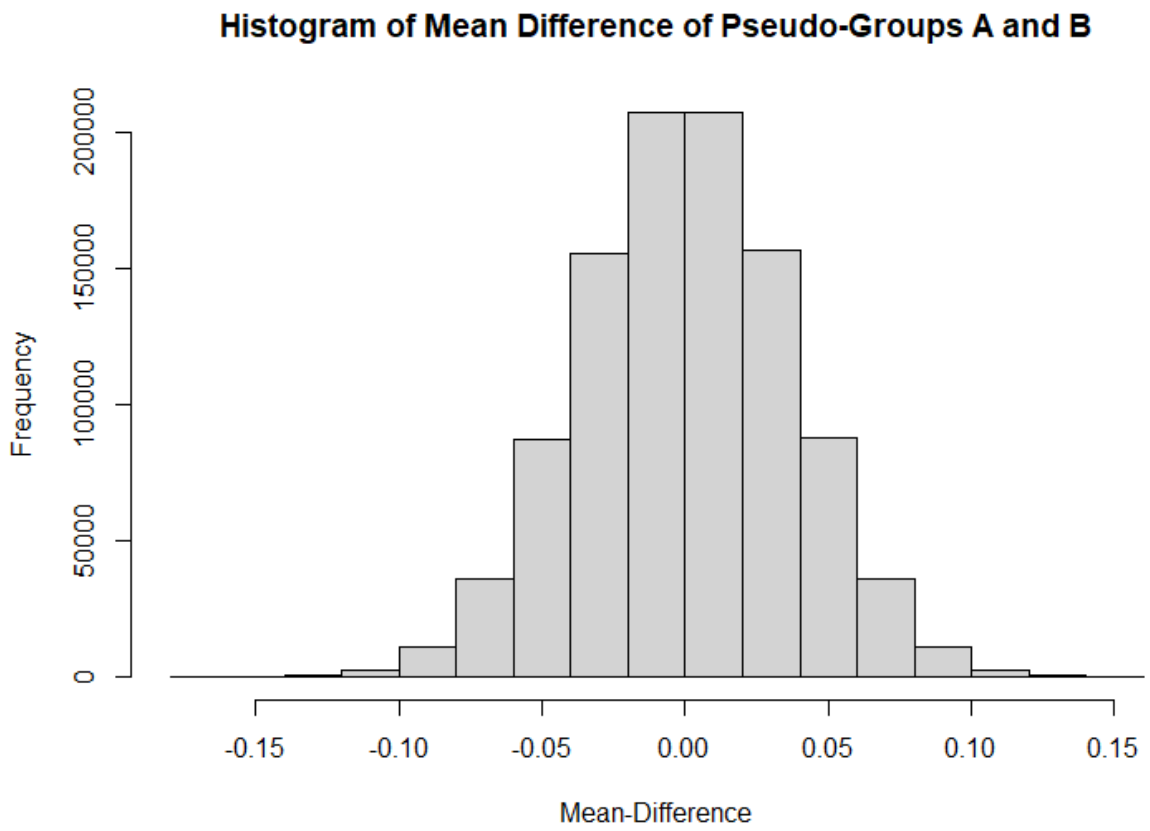
```
x = 0;
for (i in 1:1000000) {

  PgainA = mean(sample(gain, 73))
  PgainB = (mean(gain)-(PgainA*(73/155)))/(82/155)
  diflist = (PgainA - PgainB)
  meanDifList <- c(meanDifList, diflist)
  if (PgainA - PgainB > meanDif) {x = x + 1}
```

```
}
```

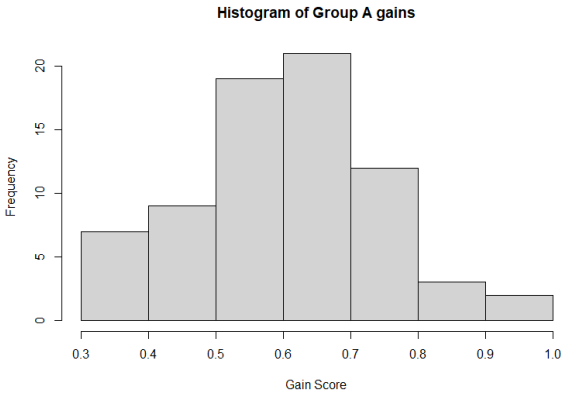
In the above algorithm, a random sampling of the entire cohort occurs each iteration, constructing two "pseudo-groups" for A and B, whose sizes mirror that of the original groups. At the end of the 1,000,000th iteration, our computed x value represents a value a million times larger than our P-value, once we generated x, simply dividing by a million, for each iteration of the algorithm, provided us with the proper p value for the dataset. Another product of this algorithm is the generation of a list of all differences of means between both pseudo-groups over the million iterations, we use this data to plot a histogram of differences, which we compare to our p-value to find exactly how uncommon the provided distribution of test-scores is.

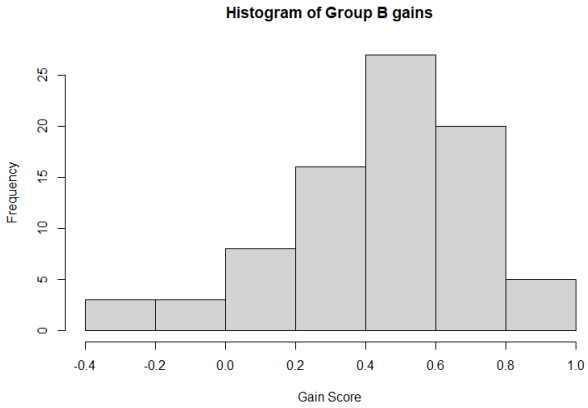
For our second dataset, mathematics pre-test and post-test scores for a cohort of students in a mid-west university. The cohort of students is split into two groups: "traditonal" and "pilot". The "traditional" group were taught by a lecture style method using existing texts, while the "pilot" group were taught with a modified text that emphasized greater student engagement. In order to observe a significant and meaningful difference in the mean gain scores between the two groups the same methods listed above for the initial group were used.

Results



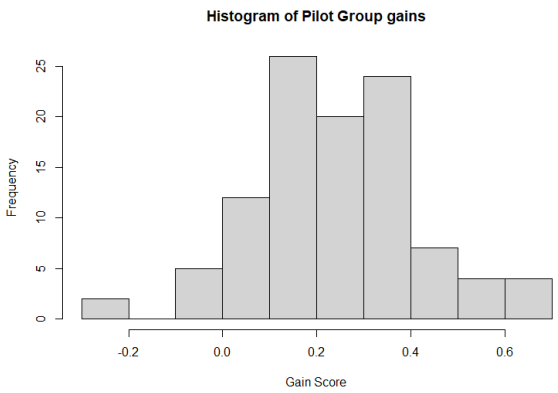
The results of our bootstrapping revealed that the mean-difference for 1,000,000 samples of pseudo-groups A and B had an almost perfect gaussian distribution. This makes sense for such a high distribution. What this tells us about the actual Groups A and B is that the actual group's distribution is very uncommon, with a mean difference of 0.160335 percentage gains.

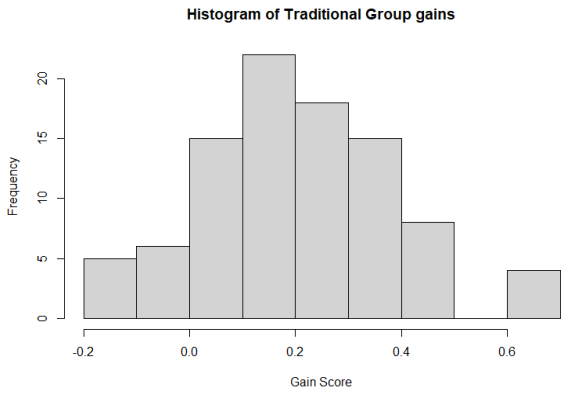




GAIN (%)	Mean	Median	Std. Dev.	Skewness	Kurtosis	5-Point Summary
Group A	0.6037353	0.6119403	0.1366652	0.06689977	2.729604	[0.3250 0.5068 0.6119 0.6037 0.6875 0.9286]
Group B	0.4434003	0.4621127	0.262181	-0.5386288	3.123546	[-0.2500 0.2632 0.4621 0.4434 0.6250 0.9730]

Unsurprisingly, simply "eyeballing" these histograms can tell us that the difference in gains is statistically significant. While the peak gains for Group A and Group B are adjacent, between 60-70% and 40-60% respectively, it appears both groups are oppositely skewed, although very slightly. Those who scored below the median skew very, very slightly up in score, while those at or above the median are skewed moderately downward. We understand that with a P-value of 2/1,000,000 (avg. after a handful of tests using the above algorithm) the observed difference is unlikely to be due to chance.





GAIN (%)	Mean	Median	Std. Dev.	Skewness	Kurtosis	5-Point Summary
"Pilot"	0.2042765	0.1944444	0.177572	0.2355234	3.224763	[-0.19444 0.08861 0.19444 0.20428 0.30864 0.66234]
"Traditional"	0.2465833	0.2437146	0.1641452	0.1215444	3.439531	[-0.2069 0.1429 0.2437 0.2466 0.3704 0.6486]

It's hard to make any strong conclusions based on the distributions of the above histograms. "Eyeballing" these distributions is definitely not enough. On top of that, bootstrapping the gains for both groups Pilot and Traditional yields an x value of 50163 on average. That's a p-value of 0.05. It's realistic, but hard to draw any conclusions from, as 0.05 is usually the cut-off point for rejecting the null. In this case, it seems that the case that the distributions occurred the way they did by chance and not by some external factor, is too likely to be rejected. (Though this is subjective). A p-value of 0.05 suggests the null being true is less likely than flipping heads 4.32 times in a row. With a high kurtosis of 3-3.5, we can also see that there are a handful of outliers for both groups.

Conclusion

With the given data it is safe to conclude there is a significant difference in gain scores between Groups A and B. This difference, as illustrated with the calculated P-value of near-zero, we reject the null and conclude that our results are not due to chance but the product of reason and causality. In theory, it could be that those who scored better on the pre-test were simply more equipped initially, and their moderate loss can be explained by simply trending toward the average. However, this is purely theoretical, and the providers of this dataset should confirm this for themselves or conduct additional observations to identify the probable cause of this difference. No strong conclusions can be drawn from the difference in mean

gains of the “Pilot” and “Traditional” groups. The ambiguity of their histograms, and the outcome of their t-test straddling the fine line of rejecting nullity, it would be irresponsible to assert any significance to our findings.