# Introduction

This document includes all required content for the final project for the Udacity course [AB Testing](https://www.udacity.com/course/ab-testing--ud257). The final project asked to analyze a new feature in its courseware to determine if it met the business objective. Specifically, this experiment tested a change where students who selected to start a free trail of the premium version of the course were prompted with a message describing how much time it usually takes for a student to be successful in a course, the prompt can be seen [here](https://drive.google.com/file/d/0ByAfiG8HpNUMakVrS0s4cGN2TjQ/view). The hypothesis was that if Udacity could reduce the number of students who started the free trial without reducing the number of students who eventually paid, those students in the free trial would get a better experience. The data and analysis below shows the process of determining a recommendation for whether or not the change should be implemented.

# Experiment Design

## Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

**Invariant**

* **Number of cookies:** That is, number of unique cookies to view the course overview page. (dmin=3000). This is an invariant so we can insure to compare similar amounts of unique visitors across the different groups.
* **Number of clicks:**That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (dmin=240). Similar to the number of cookies, I want to ensure that there are consistent numbers being considered in the experiment.
* **Click-through-probability:** That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. (dmin=0.01). This is almost exactly the same as above, the experiment is testing how the warning affects follow through after signing up for the free trail.

**Evaluation Metrics**

* **Gross conversion:**That is, number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button. (dmin= 0.01). While we ultimately want to have more people make a payment, we also want to monitor how many people are starting the trial, and in fact, may want to see less people choosing to enroll as a result of the “heads up” provided by the prompt.
* **Net conversion:**That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button. (dmin= 0.0075). Similar to above, this metric is comparing the number of user-ids over total number of cookies to click on the start free trail. This gives us a good idea of how many people clicked on the start free trail and then stayed with the program through 14 days. We would expect this metric to stay the same and possible increase if the experiment worked as hypothesized.
* **Retention:**That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. (dmin=0.01). Retention is going to be used because it’s going to help indicate if the warning is influencing students to remain past the 14 weeks. I would expect this to increase if the feature is working. This is one step further in the funnel, than net-conversion, seeing how many folks who have signed up stay past the 14 days.

**Not Either**

* **Number of user-ids:** That is, number of users who enroll in the free trial. (dmin=50).

## Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

**Standard Deviation**

Gross Conversion: 0.02023

Net Conversation: 0.0156

Retention: 0.05495

Since all of these metrics are probability metrics, one could start with an assumption that there is a binomial distribution, which is normal, and the analytical estimate of standard deviation would be valid. In addition to this, however, we must determine how the unit of diversion compares to the unit of analysis.

The other consideration with measuring the variability relates to the comparison of the unit of analysis vs. the unit of diversion. It has been found that when the unit of diversion and the unit of analysis are the same, the empirical and analytical variation are closer than if the two differ, in which case the analytical variation is an underestimate of the empirical. The table below shows the unit of analysis, the unit of diversion, and whether or not there is a potential for a large difference between the two. The table shows that both Gross and Net Conversion use the same unit of analysis as diversion, which means the analytical variation presented above is valid. However, the Retention metric has a different unit for both, which likely means that the empirical variation is greater than the analytical variation. This provides even more reason to exclude retention from the evaluation metrics because even the analytic estimate of variation requires a significantly higher number of page views that we would want for this experiment.

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Unit of Analysis | Unit of Diversion | Analytical or Empirical Variation? |
| Gross Conversion | Cookie | Cookie | Analytical |
| Net Conversion | Cookie | Cookie | Analytical |
| Retention | User-ID | Cookie | Empirical |

## Sizing

### Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

Number of Pageviews requires for each metric:

**Net Conversion:**

Baseline Conversion: .1093

Dmin: .0075

Beta: 20%

Alpha: 5%

Sample Size: 27411

Number of Page views per sample: 342,637.5

Number of Samples: 2

Total Page views needed: 685,275

**Retention**

Baseline Conversion: .53

Dmin: 0.01

Beta: 20%

Alpha: 5%

Sample Size: 39,087

Page Views per Sample: 2368909

Number of samples (experiment and control): 2

Total Page Views: 4737818

**Gross Conversion**

Baseline Conversion: 0.20625

Dmin: 0.01

Beta: 20%

Alpha: 5%

Sample Size: 25,812

Page Views per sample: 322650

Number of samples (experiment and control): 2

Total Page views: 645300

Because the retention metric requires such a significant number of page views and would require such a long duration (>100), we will not use it. Given this, the net conversion requires slightly larger number of page views with 685,275. Also, because Retention requires such a large number of pageviews (>4 million) which would require a fairly long duration experiment, this metrics will not be used as an evaluation metric.

### Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

|  |  |
| --- | --- |
| Number of Page views (unique cookies) | 40000 |
| Diversion % | 0.5 |
| Total in Experiment per day | 20000 |
| Total Page Views Desired | 685225 |
| Total Days Needed | 34.26125 |
|  |  |
| Actual Days (round up to get the amount needed) | 35 |

Given the traffic of 40,000 cookies per day, 35 days will be needed to achieve the proper sample size for both the experiment and control. I chose to divert 50% of all traffic because it intends to reduce the number of people who start the free trial, which could have negative consequences towards paying customers and overall revenue. Also, the 50% diversion requires 35 days, which might be a bit longer than desired, but is short enough to reduce the risk of paying customers. This experiment is slightly risky for Udacity, so a smaller diversion fraction has been chosen to reduce that risk.

# Experiment Analysis

## Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.)

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.**

I want to run sanity checks on three different variables:

1. Number of Cookies
2. Number of Clicks
3. Click-through-probability

The first 2 metrics we will be checking to see if the split between the groups is what we would expect when considering that they are randomly assigned to each group with a .5 probability. The third metric, click-through-probability, will check that the rate for both the experiment and control is within the acceptable amount, which will be a slightly different process than the first two.

**Comparing counts between experiment and control**

1. Number of Cookies and Clicks
   1. For these two metrics, I want want to compute the range of acceptable range of a proportion of cookies that fall into the “success” considering an expected .5 probability with a 95% confidence.
   2. To calculate this, I will assume a binomial distribution and consider the number within the “control” group to be the success. This will then produce an upper and lower bound in the form of a decimal on either side of the expected proportion of .5.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Lower Bound | Upper Bound | Observed Proportion in Control | Pass? |
| Page Views | 0.4988 | 0.5012 | 0.5006 | Yes |
| Clicks | 0.4959 | 0.5041 | 0.5005 | Yes |

* 1. For both of these metrics, the proportion of observations within the experiment and control match what would be expected, considering these metrics, the experiment is performing up to expectation.

1. Click-through-probability
   1. I also want to ensure that the experiment and control group do not vary significantly in regards to the number of people who click on “Start Free Trail” so using click through probability will help ensure this.
   2. In a slight variation of the calculation above, I will calculate the observed click through probability of the control group and then determine if the experiment probability falls within the 95% confidence interval that suggest the two groups are not different.
   3. The table below shows the results of these calculations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Control Probability | Lower Bound | Upper Bound | Experimental probability | Pass? |
| Click-through-probability | 0.0821 | 0.0812 | 0.0830 | 0.0822 | Yes |

* 1. Given the control and experimental probability, this metric also passes the sanity check, suggesting that the control and experimental group are not statistically different.

## Result Analysis

### Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)

For each of the evaluation metrics, Gross Conversion and Net Conversion, a comparison of the means is done by calculating the pooled standard error to calculate the confidence interval. The pooled standard error calculation is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metric |  | Lower Bound | Upper Bound | Statistically Significant? | Practically Significant |
| Gross Conversion | -0.02055 | -0.2912 | -0.0120 | Yes | Yes (=.01) |
| Net Conversion | -0.0049 | -0.0116 | 0.0019 | No | No (=.0075) |

Given these numbers, the experiment showed that a statistically and practically significant number of people chose not to enroll in the class, but that the ratio of users who actually paid for the class did not significantly change in either direction. Given this, it suggests that the experiment only deterred some users from participating in the trail but it did not impact the number of people who ultimately paid for the service.

### Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Gross Conversion (success = lower gross conversion) | Net Conversion (success = higher net conversion) |
| How many days did the experiment see a "success" |  | 19 | 10 |
| Total Days |  | 23 | 23 |
|  |  |  |  |
| Two Tailed P-Value |  | 0.0026 | 0.6776 |

The sign test showed that the chance of seeing the results displayed in the Gross Conversion metric were unlikely to appear due to chance alone (p value = 0.0026) but the Net Conversion results were far from statistically significant with a p-value of 0.6776.

### Summary

Because Gross Conversion and Net Conversion are likely correlated with one another, the Bonferroni correction is likely to be too conservative.

## Recommendation

Because this experiment did not make a statistically significant difference in the number of people who enrolled past the 14 day trail, the experiment is not resulting in increased revenue via paid subscribers. The experiment, however, does result in less users participating in the free trial. This type of result may be valuable to the business if coaching resources are low and there is a need to limit their workload, then it might be worth launching this experiment on a limited basis to see what impact it has on coach’s ability to provide assistance to employees.

# Follow-Up Experiment

The results of the experiment suggest that the relationship between those who choose to start the trial and those who choose to stay past 14 weeks is not affected by setting expectations. I would want to better understand the activity level of those who pay vs. those who don’t, especially right before the 14-week time frame. I would want to know if users who are active close to the deadline but do not pay would be more likely to enroll if given an extra week of a free trial, a “bonus week.” If during that week they got closer to completing the class and more engaged, then they may be more likely to pay after 3 weeks than just 2.

Because some of this analysis might be a bit difficult to conduct, I would recommend an experiment where some users who click start free trial are “selected” to receive a 3 week trial to see if they are more likely to pay for the service than those who only receive a 2 week trial.

I would use the cookie as the diversion metric to separate those who clicked on “Start Free Trial” and then track by user the number ratio who are converted.