# A/B Testing – Final Project

Cameron Roberts

## Experiment Design

### Metric Choice

The experiment is focusing on the users who click the "start free trial" button. We want to test whether a message asking about the time available for study will make users more carefully consider their circumstances and ability to commit. So we are interested in how many users go on to enrol in the free trial *after* seeing this message.

The metric which measures precisely this is *gross conversion.* We hope that the experiment group would have a smaller value for this metric indicating that some users did not enrol after reading the message.

At the same time, it was also indicated we do not want to significantly reduce the number of students to continue past the free trial and go on to complete the course.

Two metrics, *retention* and *net conversion* provide possible ways of evaluating this. Retention measures how many students which start the free trial go on to pay while net retention measures how many users who click on the "start free trial" button become paid subscribers. So while we want to see gross conversion reduce, we would like to see no significant reduction in these two evaluation metrics. If they were to increase, it would be a bonus because we would be increasing paying customers while frustrated users would be reduced.

In the list provided there are also some invariants we can measure to ensure there are no underlying differences between our control and experiment groups. This includes the *number of cookies* to visit the course overview page, the *number of clicks* on the "start free trial" button and *click-through-probability*. Each of these measure activity *before* the experiment message is shown and thus should not be significantly different between groups.

Finally, there was a final metric for consideration: number of user ids. This metric will not be used as either an invariant of evaluation metric. It is not suitable since the number of users who start the free trial will be dependent on the number of clicks on "start free trial". The ratio of these two values is gross conversion which is better choice.

### Measuring Standard Deviation

#### *Gross Conversion*

The analytical estimate of the standard deviation of the metric is given by $\sqrt{\hat{p}(1-\hat{p})/n}$ where $\hat{p}$ is the estimate of the probability of enrolling given a click on "start free trial" and $n$ is the number of cookies to click on "start free trial". $\hat{p}$ has been given as 0.2063 and for 5,000 page views, $n = 400$ which is 5,000 times the click through probability on "start free trial", 0.08. Applying these numbers to the formula, the standard error is 0.0202.

#### *Net Conversion*

Here we use the same formula, $\widehat{SE} = \sqrt{\hat{p}(1-\hat{p})/n}$, but $\hat{p}$ is now the estimate of the probability of payment given a click on "start free trial", 0.1093, and $n$ is the number of cookies to click on "start free trial", 400. Hence the standard error is 0.0156.

### *Retention*

Again, using the formula for the standard error with $\hat{p}$ as the estimate of the probability of payment given enrolment, 0.53 and $n$ is the number of cookies, 400 times the enrolment probability of 0.2063 giving 82.5. With these values we get the standard error of retention as 0.0549.

For the first two metrics, gross conversion and net conversion, the unit of analysis is the number of unique cookies to click on the "start free trial" button. The unit of diversion in our experiment is also a cookie so we can have reasonable faith in our analytical estimates which should align closely with the empirical ones. However, for retention, the unit of analysis is user-id which is different to the unit of diversion, so it may be worth calculating the empirical variability in this case, if time permits.

## Sizing

### *Number of Samples vs. Power*

I will not use the Bonferroni correction for this experiment. The Bonferroni correction is more conservative than using the original alpha value. While it will control the family wise error rate, in our particular experiment this is probably not desired. If our decision to launch was based on the success of *any* test it may be worth using the correction so that we control FWER and reduce type I errors (false positives). But in our experiment we base our decision on the results of *all* tests. We don't want to see a significant reduction in retention or net conversion. A side effect of the Bonferroni correction is that it also increases type II errors, ie. false negatives. This means we would be more likely to incorrectly diagnose that the introduced message had a significantly negative impact on students going on to complete the course.

So, with $\alpha = 0.05$ and $\beta = 0.2$ I use the online calculator here to determine the sample size required to meet the power for each metric:

### Gross conversion

$d_{min} = 0.01$, so 25,835 clicks on "start free trial" are required in each group giving a total of 51,670 across both. To get the number of page views we divide by the click through probability on "start free trial", 0.08 giving 645,875 page views.

### Net conversion

$d_{min} = 0.0075$, so 27,413 page views are required in each group giving a total of 54,826 across both. Similarly, this translates to 685,325 page views.

### Retention

$d_{min} = 0.01$, so 39,115 enrolments are required in each group giving a total of 78,230 across both. To translate this to the number of page views required, we need to divide by both the probability of enrolment 0.2063 and the click through probability on the "start free trial" button 0.08. This results in the number of page views required being 4,741,212 rounded to the nearest integer.

Retention has by far the largest requirement on sample size here, so this the number of page views we will seek.

### *Duration vs. Exposure*

I see this experiment being of low risk. We are not tracking sensitive or harmful data as a part of the experiment. Although, retention is measured using user-ids which are linked to credit card numbers so extra care should be taken in this area not to expose these.

We need a considerably large number of page views to get the sample size we require. If we direct 50% of traffic we will require 238 days to run the experiment, which is far too long. Even 100% percent would require 119 days.

Since the retention metric requires so many page views, I will recommend this metric is no longer used for evaluation. With this decision, we now require only 685,325 page views to meet power requirements for the two conversion metrics. For this number of views, a 50% exposure would require 35 days for our experiment duration. This is still a little long. Since the experiment is low risk, we can divert 100% of traffic allowing us to complete the experiment in 18 days.

## Experiment Analysis

### Sanity Checks

We've selected three invariants for our sanity checks: *number of cookies*, *number of clicks* and *click-through-probability*. We expect there to be no significant difference between these metrics between the experiment and control groups.

### *Number of cookies*

In the experiment data we have 345,543 unique cookies in the control and 344,660 in the experiment. With these numbers we construct a 95% binomial confidence interval: $0.5 \pm (1.96 \times 0.0006) = [0.4988, 0.5012]$. The ratio of page views in the control to the total is $345{,}543/690{,}203 = 0.5006$ which is within these bounds.

### *Number of clicks*

Here we have 28,378 clicks in the control and 28,325 in the experiment group. The 95% confidence interval is $0.5 \pm 1.96 \times 0.0021 = [0.4959, 0.5041]$. The ratio of clicks in the control to the total is $28{,}378/56{,}703 = 0.5005$ which is within the confidence interval.

### *Click-through-probability*

The click-through-probability of the control group is $28{,}378/345{,}543 = 0.0821$ and for the experiment group is $28{,}325/344{,}660 = 0.0822$. The 95% binomial confidence interval based on the pooled probability and pooled standard error is $\pm 1.96 \times 0.0007 = [-0.0013, 0.0013]$. The difference between the CLPs for the control and experiment is 0.0001 which is within bounds.

All of our invariant metrics fall within their respective confidence intervals and thus the sanity checks have passed. We can now continue to our analysis of the evaluation metrics.

### Result Analysis

#### *Effect Size Tests*

##### Gross conversion

The gross conversion is calculated as the number of enrolments over the number of clicks. Care must be taken to use only the days for which we have both these values. Doing so we have $3{,}785/17{,}293 = 0.2189$ for the control group. Likewise, the gross conversion for the experiment group is $3{,}423/17{,}260 = 0.1983$. This gives a difference $\hat{d} = -0.0206$. Now, the estimate of the standard error is

$$\widehat{SE} = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{1/n_{exp} + 1/n_{cont}}} = \sqrt{\frac{0.2086(1-0.2086)}{1/17{,}260 + 1/17{,}293}} = 0.0044.$$

So, in this case we have a confidence interval of $-0.0206 \pm 1.96 \times 0.0044 = [-0.0291, -0.0120]$. The confidence interval does not include zero which means that at this alpha level there is significant evidence to suggest we can reject the null hypothesis that there is no difference between the gross conversion for the two groups. The practical significance $d_{min} = -0.01$ also falls outside this interval.

#### Net conversion

The net conversion is defined as the number of paying students over the number of clicks. As we did for gross conversion, we only use the days for which we have data for both. The control group has a net conversion of $2{,}033/17{,}293 = 0.1176$. For the experiment group we have $1{,}945/17{,}260 = 0.1127$. So, our difference is $\hat{d} = -0.0049$. The estimate of the standard error is 0.0034 so we have a confidence interval of $-0.0049 \pm 1.96 \times 0.0034 = [-0.0116, 0.0019]$. This confidence interval includes zero so there is not enough evidence to reject the null hypothesis. The practical significance of $-0.0075$ also falls within this interval.

### *Sign Tests*

#### Gross conversion

There are 23 days for which we can calculate the gross conversion in the control and experiment. Of these, 19 days have a lower value for the experiment group. The *p*-value for the sign test using these values is 0.0026 which is significant at alpha level 0.05.

#### Net conversion

Again we have 23 days, but for net conversion only 13 values are smaller in the experiment than the control group. The *p*-value for this sign test is 0.6776 which is not significant at the 0.05 alpha level.

These two results agree with the effect size tests we performed in the previous section.

## Recommendation

The effect test on gross conversion was significant and the practical significance was met (ie. was in the confidence interval), so we succeeded in reducing gross conversion as a result of the experiment. However, we also wanted to ensure there was no significant reduction in net conversion. While we could not reject the null hypothesis since zero was in the confidence interval for this test, the confidence interval also included the practical significance value. As a result, we don't have enough evidence to suggest the difference is not large enough to worry about.

So, based on these results, it is my recommendation that Udacity does not launch the change tested in the experiment.

## Follow-Up Experiment: How to Reduce Early Cancellations

It may be worth testing if varying the length of the free trial period has an effect on reducing early cancellations. If the length of the free trial is shorter it may make visitors considering enrolling think more carefully upfront about whether they have enough time to commit. I propose testing making the free trial a shorter 7 days. My hypothesis is that this will reduce the number of early cancellations thus improving the student experience.

The unit of diversion will be the cookie, and visitors to the Udacity course page would be presented with either a 14-day (control) or 7-day (experiment) free trial length depending on which group they are allocated to. i.e. The experiment is conducted *before* clicking on "start free trial" button.

We will use click-through-probability on the "start free trial" button (ie. no. clicks divided by number of cookies) as an evaluation metric to measure the proportion of visitors to the page which begin a free trial. We want to see this metric reduce, but we hope not at the expense of final enrolments.

We would then also evaluate conversion to see how many users become full enrolments. ie. Conversion will be defined as the number of user-ids who make a payment divided by the number who began a free trial. We do not want to see a significant decrease here and an increase would be welcomed.

We would use the number of cookies to visit the Udacity page per day as an invariant metric an ensure there is no significant difference between the number assigned to the control and experiment groups.

As mentioned earlier, the unit of diversion in our experiment is the cookie. A risk, however, with this is that if a user revisits the page with a fresh cookie (perhaps after clearing or using a different device), they may be alarmed to see a different trial period length.

## References


1. https://discussions.udacity.com/t/when-to-use-bonferroni-correction/37713/6
2. https://en.wikipedia.org/wiki/Bonferroni_correction
3. http://stats.stackexchange.com/questions/65942/how-and-when-to-use-the-bonferroni-adjustment
4. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112991/
5. http://www.evanmiller.org/ab-testing/sample-size.html
6. https://discussions.udacity.com/t/final-project-hypothesis/37176/7
7. https://discussions.udacity.com/t/p7-empirical-variance-and-anlytical-variance/38868/2
8. https://en.wikipedia.org/wiki/Binomial_distribution
9. http://www.r-tutor.com/elementary-statistics/non-parametric-methods/sign-test
10. https://discussions.udacity.com/t/can-you-test-for-a-null-effect/165338