Free Trial Screener AB Test

August 22, 2016

1 Free Trial Screener AB Test

1.1 Overview

At the time of this experiment, Udacity courses currently have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead. This screenshot shows what the experiment looks like.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

1.2 Experiment Design

1.2.1 Metric Choice

1.2.2 Invariant Metrics

- **Number of cookies**: Number of unique cookies to view the course overview page.
- **Click-through-probability**: Number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.

1.2.3 Evaluation Metrics

- **Gross conversion**: 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.
- **Retention**: 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.
- **Net conversion**: 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.

1.2.4 Metrics Not Used

- **Number of user-ids**: That is, number of users who enroll in the free trial.
- **Number of clicks**: Number of unique cookies to click the "Start free trial" button.

1.2.5 Explanation

- **Number of cookies**: This was chosen because as an invariant metric because it is the unit of diversion, and the number should not vary between control and experimental groups since the experiment occurs after this metric is recorded.
- **Click Through Probability**: CTP is a better option as an invariant metric than number of clicks because it accurately tracks how many independent users move on to the next page. Also, since the intervention occurs after the "Start free trial" button is clicked, the CTP would not be affected by the experiment.
- **Gross conversion**: A statistically significant change in the experiment will tell us how the experiment affected a user in enrolling in the free trial. Since we are expecting a decrease in user ids in the experiment, we are hoping to see a decrease in gross conversion since the experiment will dissuade visitors who are not ready for the program.
- **Retention**: The experiment would filter out visitors that would not have time to progress during the 14 day trial. As a result, the resulting proportion of user-ids to continue past the trial period should have a a nonnegative stastically significant change here.
- **Net conversion**: Since we assume the intervention would not discourage the visitors that would continue past the 14 day trial period, we expect this metric to have nonnegative practical significiance. For this reason, it makes an excellent choice as an evaluation metric.
- **Number of user-ids**: This may be actually used an evaluation metric, but it would not be a good one. With this metric alone, we are unable to tell if the experiment had a positive or negative effect with a statistically significant change between experiment and control since we are not able to normalize between the two groups.
- **Number of clicks**: This metric can be used as an invariant metric because it occurs prior to the intervention, but it is generally used if there was an UI change. The CTP is a much better choice as an invariant metric than this one. There also might be a slim chance that this metric may be unreliable if there were usability issues during the time when the experiment was launched or in the control.

1.2.6 Expected Results

If the experiment is successful, I expect the users who don't have the time commitment will not complete checkout and enroll into the trial. As an effect, the number of user-ids should statistically decrease, and of those users who wish to enroll into the trial, a higher percentage of them should

convert into paying users. The expectation of the evaluation metrics based on my conjecture are as follows: * **Gross conversion**: I expect the number of user-ids to complete checkout and enroll to decrease. This metric, as an effect, should stastically decrease. * **Retention** and **Net Conversion** should show not show negative practical significance, since the purpose of the experiment is to reduce the number of students frustrated due to not having enough time, without significantly decreasing the number of students continuing past the the trial.

1.2.7 Measuring Standard Deviation

Baseline Values:

Metrics	Values
Unique cookies to view page per day	40000
Unique cookies to click "Start free trial" per day	3200
Enrollments per day	660
Click-through-probability on "Start free trial"	0.0800
Probability of enrolling, given click	0.2063
Probability of payment, given enroll	0.5300
Probability of payment, given click	0.1093

Standard Deviation for Evaluation Metric:

Metrics	Baseline Values	SE	SE/5000
Gross Conversion	0.2063	0.0072	0.0202
Retention	0.5300	0.0194	0.0549
Net Conversion	0.1093	0.0055	0.0156

For Gross and Net Conversion, the analytical estimates would be comparable to the empircal variablility due to the fact that the denominator used to calculate those values were using the the unit of diversion, which was the cookie.

The Retention metric's analytical estimates may differ from the empirical variability. Due to this chance, the retention metric will not be used as an evaluation metric.

1.2.8 Sizing

Number of Samples vs. Power Bonferroni correction was not used in this phase because Gross Conversion and Net Conversion are highly correlated with each other. Using Bonferroni correction would too conservative to use, requiring too many pageviews required for the experiment.

```
In [75]: # Samples Needed calculated using
         # http://www.evanmiller.org/ab-testing/sample-size.html
         # Required pageviews for Gross Conversion
         samples_needed = 25839.
         pageviews = 2 * samples_needed / 0.08
         print("Gross Conversion Pageviews Required: {}".format(pageviews))
```

```
Gross Conversion Pageviews Required: 645975.0
```

```
In [76]: # Required pageviews for Retention
         samples_needed = 39115.
         pageviews = 2 * samples_needed / 0.0165
         print("Retention Pageviews Required: {}".format(pageviews))
```

```
Retention Pageviews Required: 4741212.12121
```

```
In [77]: # Required pageviews for Net Conversion
         samples_needed = 27411.
         pageviews = 2 * samples_needed / 0.08
         pageviews
         print("Net Conversion Pageviews Required: {}".format(pageviews))
```

```
Net Conversion Pageviews Required: 685275.0
```

Duration vs. Exposure This experiment will not affect other parts of the services that Udacity provides. Paying customers who are already enrolled will not be affected by the experiment. The experiment being launched also does not represent any drastic change. With that said, I consider this experiment being of **low risk**. It will be launched on all traffic.

With a baseline of 40,000 pageviews a day, the experiment running on all traffic the number of pageviews required for 474,1212 views is 118 days, which is too long of a duration. The Retention evaluation metric **will not** be used.

The experiment will run for 18 days to cover the 685,275 pageviews required to fulfill the requirements for both Retention and Net Conversion.

1.3 Experiment Analysis

1.3.1 Sanity Checks

```
In [79]: # Number of Cookies
         import math
         pageviews_cont = 345543.
         pageviews_exp = 344660.
         pageviews_tot = pageviews_cont + pageviews_exp
         p = 0.5
         std_error = math.sqrt(p*(1-p)/(pageviews_tot))
         print("Standard Error: {}".format(std_error))
         margin = std_error * 1.96
         ci = [0.5-margin,0.5+margin]
         observed = pageviews_cont/pageviews_tot
         print("Confidence Interval: {},\nObserved: {}".format(ci, observed))
```

```
Standard Error: 0.000601840740294
Confidence Interval: [0.49882039214902313, 0.5011796078509769],
Observed: 0.500639666881
```

```
In [80]: # Click Through Probability
         clicks_cont = 28378.
         clicks_exp = 28325.
         ctp_cont = clicks_cont / pageviews_cont
         ctp_exp = clicks_exp / pageviews_exp
         ctp_pooled = (clicks_cont + clicks_exp)/ (pageviews_tot)
         p = 0.8
         std_error = math.sqrt(ctp_cont*(1-ctp_cont)/(pageviews_cont))
         print("Standard Error: {}".format(std_error))
         margin = std_error * 1.96
         ci = [ctp_cont-margin, ctp_cont+margin]
         print("Confidence Interval: {},\nObserved: {}".format(ci, ctp_exp))

Standard Error: 0.000467068276555
Confidence Interval: [0.08121035975252971, 0.08304126739662393],
Observed: 0.0821824406662
```

Both invariant metrics are within range of the confidence interval.

1.3.2 Results Analysis

Effect Size Tests

```
In [86]: # Gross Conversion
         eclicks_cont = 17293.
         eclicks_exp = 17260.
         enroll_cont = 3785.
         enroll_exp = 3423.

         gross_conv_pool = ((enroll_cont + enroll_exp) /
                            (eclicks_cont + eclicks_exp))
         se_pool = math.sqrt(gross_conv_pool *
                             (1 - gross_conv_pool) *
                             (1/eclicks_cont + 1/eclicks_exp))
         print("Gross Conversion Pooled Probability: {}".format(gross_conv_pool))
         print("Std. Error Pooled: {}".format(se_pool))

         d = (enroll_exp / eclicks_exp) - (enroll_cont / eclicks_cont)
         margin = se_pool * 1.96
         ci = [d-margin, d+margin]
         print("Confidence Interval: {},\nd: {}".format(ci, d))

Gross Conversion Pooled Probability: 0.208607067404
Std. Error Pooled: 0.00437167538523
Confidence Interval: [-0.0291233583354044, -0.01198639082531873],
d: -0.0205548745804
```

The gross conversion for the experiment shows practical significance based on dmin = 0.01. It is statistically significant since 0 is outside of the confidence interval.

```
In [85]: # Net Conversion
         eclicks_cont = 17293.
         eclicks_exp = 17260.
         pay_cont = 2033.
         pay_exp = 1945.

         net_conv_pool = (pay_cont + pay_exp) / (eclicks_cont + eclicks_exp)
         se_pool = math.sqrt(net_conv_pool *
                             (1 - net_conv_pool) *
                             (1/eclicks_cont + 1/eclicks_exp))
         print("Net Conversion Pooled Probability: {}".format(net_conv_pool))
         print("Std. Error Pooled: {}".format(se_pool))

         d = (pay_exp / eclicks_exp) - (pay_cont / eclicks_cont)
         margin = se_pool * 1.96
         ci = [d-margin, d+margin]
         print("Confidence Interval: {},\nd: {}".format(ci, d))
```

```
Net Conversion Pooled Probability: 0.115127485312
Std. Error Pooled: 0.00343413351293
Confidence Interval: [-0.011604624359891718, 0.001857179010803383],
d: -0.00487372267454
```

With the resulting confidence interval, Net conversion is not practically significant for dmin= 0.0075. It is also not statistically significant due to 0 being inside the confidence interval.

Sign Tests Using http://graphpad.com/quickcalcs/binomial1.cfm

For Gross Conversion there were 4 successes out of 23 trials. This yielded a two tailed p-value of 0.0026. This is statistically significant for alpha = 0.05/2.

For Net Conversion there were 10 successes out of 23 trials. This yielded a two tailed p-value of 0.6776. This is not statistically significant for alpha = 0.05/2.

Summary As stated before Bonferroni correction was not used in this phase because Gross Conversion and Net Conversion are highly correlated with each other. Using Bonferroni correction would too conservative to use, requiring too many pageviews required for the experiment.

We are also choosing not to use Bonferroni correction because we are requiring both Net Conversion and Gross Conversion to show statistical and practical significance. Bonferroni correction would be useful in a case where any metric showing statistical significance is enough to trigger the launch of an experiment.

1.3.3 Recommendation

Effect size and sign test results showed that then net conversion metric did not render any practical or statiscal significance. However, it does suggest a negative decrease in net conversion because of

the resulting negative practical significance value. We also did not meet our d_{min} criteria. Because of this, I am choosing not to launch this experiment.

Gross Conversion, however did show practical and statistical significance, which indicates that the experiment did influence behavior in deciding whether to enroll in the trial. The d value for gross conversion is negative, which means that there were fewer users who opted to enroll in the trial session.

Although launching this experiment can in fact increase overall student experience and improve coaches' capacity to support students who are likely to complete the course, since our results potentially suggest a decrease in net conversion, there might be a risk of udacity losing potential people who enroll into the program.

1.4 Follow-Up Experiment

On Udacity, for each project that needs to be completed, there is a recommended due date for each project specified for a given cohort. However, for each project there currently isn't something that users can use to see if they are on track to complete the project on a week to week basis. It is very easy for a student to fall off track and not meet the recommended due date for each project.

My hypothesis, is that if there was such a mechanism to help keep students on track on a week to week basis, our current evaluation net conversion metric will increase.

1.4.1 Experiment Details

- **Null Hypothesis**: A week to week schedule on what to complete, starting from the date a user starts the 14 day trial period, will not increase Retention.
- **Alt Hypothesis**: A week to week schedule on what to complete, starting from the date a user starts the 14 day trial period, will increase Retention.
- **Unit of Diversion**: User-id is the unit of diversion. We would like to compare the retention after the experiment with user-id's who are in the control group against the experimental, which has this "week by week calendar feature" available for them.
- **Invariant Metric**: The number of user-ids, that is the number of users who enroll in the free trial will be used as an invariant metric, since the unit of diversion we are using is the cookie.
- **Evaluation Metric**: Retention will be used as the evaluation metric, because we want to see if our experiment shows an increase of enrollment for users who have access to this week by week calendar resource.

1.5 References

- Click Through Rate vs. Click Through Probability
- Udacity Forum Discussion
- Udacity Lesson
- Z-Score Table
- http://graphpad.com/quickcalcs/binomial1.cfm
- http://www.evanmiller.org/ab-testing/sample-size.html
- http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full