A/B Testing Final Project

Ren Zhang
zhang_ren@bentley.edu

Experiment Design

Metric Choice

1. Number of cookies
   Used as invariant metric: this metric is used as the unit of diversion, it should be split evenly between the control and experiment groups.
   Not used as evaluation metric: we are not expecting changes happen to this metric between experiment and control.
2. Number of userids
   Not used as invariant metric: this number could change between two groups, since the student answered no to the question in the experiment group will not see the enrollment option.
   Not used as evaluation metric: Even though we would expect to see difference of the metric between two groups, we should also take considerations about the total number of cookies that clicked the button, so a ratio of this metric and number of clicks is a better evaluation metric.
3. Number of clicks
   Used as invariant metric: the event happens before the experiment and this metric should not change.
   Not used as evaluation metric: we are not expecting changes happen to this metric between experiment and control.
4. Click-through-probability
   Used as invariant metric: the event happens before the experiment, therefore should not be affected by it.
   Not used as evaluation metric: we are not expecting changes happen to this metric between experiment and control.
5. Gross conversion
   Not used as invariant metric: this metric could be different across two groups, since users in the experiment group may decide not to enroll in free trial based on their time availability.
   Used as evaluation metric: we would want to see whether the number of student enroll in program could be effected with the self-assessment question.
6. Net conversion
   Not used as invariant metric: this metric could be different since the enrolled users in the experiment group are aware of the time commitment upfront.
   Used as evaluation metric: we would want to see whether the number of student continued in program after the free trail could be effected with the self-assessment question.
7. Retention
   Not used as invariant metric: this metric could be different since the enrolled users in

the experiment group are aware of the time commitment upfront.
Used as evaluation metric: since this metric has the number of userids on the denominator, it is not a good evaluation metric for this test, though it may be an interest metric to look at in other settings.

For gross conversion metric:
We are expecting to see that the gross conversion for the experiment group is lower than the gross conversion for the control group, since the students enrolled in free trial are aware of the time commitment upfront. So we are expecting to see this metric to pass both the statistical and practical significance test.

For net conversion metric:
We are expecting to see that the net conversion in the experiment group is not significantly less than that of the control group. Even though we are expecting to see less students enroll in free trial in the experiment group, but if the hypothesis is true, we are expecting to see less free trial students to drop out after the trial as well. So we are hoping to see this metric not pass both the statistical and practical significance test.

In summary, if the net conversion metric did not pass the practical significance test, we can justify the change.

Measuring Standard Deviation

Given a sample size of 5000 cookies visiting the course overview page, based on the ratio shown in the baseline values, we would expect the number of unique cookies that clicked the "Start free trail" button to be:

$$\mathrm{N} = 5000 \times \frac{3200}{40000} = 400$$

Since the evaluation metrics are measures of probability, we assume that the distribution of both Gross conversion and Net Conversion follow binomial distribution, we can obtain estimate the standard deviations using the following formula:

$$\mathrm{SE} = \sqrt{\frac{p(1-p)}{N}}$$

For Gross conversion, the estimated standard deviation is calculated using $\mathrm{N} = 400$ and $\mathrm{p} = 0.20625$, which gives 0.0202306. Similarly, using $\mathrm{N} = 400$ and p = 0.1093126, we calculated the estimated standard deviation of Net Conversion to be 0.01560155.
Since for both metrics, the quantity on the denominator is the number of unique cookies to clicks the button, which happen to be our unit of diversion, thus we could expect that the analytical variability would match with the empirical variance.

Sizing

Number of Samples vs. Power

Since that we need the information from both tests to help us make a decision, we would not use Bonferroni correction in the analysis.

For Net conversion, using the online calculator, we set the baseline conversion rate to 10.93125%, set the minimum detectable effect to be 0.75%, use alpha = 5% and beta = 20%, the calculator give us the number of clicks needed in each sample to be 27411. Similarly we can obtain the number of clicks needed in each sample for Gross conversion is 25839.

Based on an 8% click-through probability, we would need at least $\frac{27411\times2}{8\%} = 685275$ pageviews.

Duration vs. Exposure

Since the required total number of pageviews is quite large, we would use 100% of traffic to speed up the data collection process so that we can minimize the effects of other unknown changes. Using the given 40000 pageviews per day, we would expect the experiment to run for $\left\lceil\frac{685275}{40000}\right\rceil = 18$ days.

We don't think there is much risk involved in this experiment for Udacity. The technical requirement is low, and the additional step for the user is just a quick self-assessment question so the effect on the user experience on Udacity website is very minimal.

Experiment Analysis

Sanity Checks

First check the number of cookies. The total number of cookies in the experiment group is 344660 and the total number of cookies in the control group is 345543. Assume a binomial distribution with probability 0.5, we can calculate the standard deviation to be:

$$\text{SD} = \sqrt{\frac{0.5\times0.5}{344660+345543}} = 0.0006018407$$

Since that as the sample size gets larger the binomial distribution approaches normal distribution. We can construct a 95 percent confidence interval for the probability for a cookie to be assigned to the control group as:

$$[0.5 - 1.96\times0.0006018407, 0.5 + 1.96\times0.0006018407]$$

Which is: $[0.4988204\ , 0.5011796]$, since the observed value is: $\frac{344660}{344660+345543} = 0.4993603$ which fall within the confidence interval, we passed the sanity check for this one.

For the number of clicks. The total number in the experiment group is 28325 and the total number in the control group is 28378. Similarly to above procedure, we calculated the 95%

confidence interval for number of clicks as $[0.4958845\ ,0.5041155]$, which happen to contain the observed value 0.5004673, and thus we passed the sanity check for this one.

For the click-through probability, the observed click-through probability for the control group is 0.082125814 and for the experiment group it is 0.082182441, and the pooled probability is:

$$\frac{28325+28378}{344660+345543}=0.08215409$$

The estimated pooled standard deviation is:

$$\sqrt{0.08215409\times(1-0.08215409)\times(\frac{1}{344660}+\frac{1}{345543})}=0.0006610608$$

The 95 percent confidence for the difference assume there is no difference between the probabilities should be:

$$[-1.96\times0.0006610608\ \ ,1.96\times0.0006610608\ \ ]$$

Which is equal to:

$$[-0.001295679\ ,0.001295679\ ]$$

And since the above confidence interval contain the observed difference 0.000056627, we passed the sanity check for this invariant metric as well.

Result Analysis

Effect Size Tests

For the Gross conversion metric:

The observed Gross conversion rate in the experiment group is $\frac{3423}{17260}=0.198319815$, the observed Gross conversion rate in the control group is $\frac{3785}{17293}=0.218874689$.

The difference between them is: $0.198319815-0.218874689=-0.020554875$

The pooled rate is estimated to be: $\frac{3423+3785}{17260+17293}=0.208607$.

The estimated pooled standard deviation is:

$$\sqrt{0.208607\times(1-0.208607)\times(\frac{1}{17260}+\frac{1}{17293})}=0.004371675$$

The resulted 95 percent confidence interval would be:

$$[-0.020554875-1.96\times0.004371675\ \ ,-0.020554875+1.96\times0.004371675\ \ ]$$
$$=[-0.02912,-0.01199]$$

The confidence interval does not contain 0, so we passed the statistical significance test, and since the upper bound is lower than -0.01, we passed the practical significance test.

For the Net conversion metric:

Using similar procedure as shown above, we can obtain the 95 percent confidence interval: $[-0.0116,0.001857]$.

Since the confidence interval contains 0, so we did not pass the statistical significance test for this one. Also, since -0.0075 is within the confidence interval, we can failed to pass the practical significance test for this one as well.

Sign Tests

For the Gross conversion metric:
There are 4 out of 23 days on which the gross conversion is higher in the experiment group, using the online calculator, the two-tailed p value for this sign test is 0.0026.

For the Net conversion metric:
There are 10 out of 23 days on which the net conversion is higher in the experiment group, using the online calculator, the two-tailed p value for this sign test is 0.6776.

Since as mentioned before we choose not to use Bonferroni correction, so based on a 0.05 cut off, only the sign tests for gross conversion metric passed the test.

Summary

There are no discrepancies between the effect size hypothesis tests and the sign tests.

We did not use Bonferroni correction, since we need the information from both tests to help us make the decision and this makes Bonferroni correction not suitable in our case.

Recommendation

Based on the test for the gross conversion metric, we can see that using a quick self-assessment question before letting users enroll in the program will cause less student enroll in free trail. Based on the test for the net conversion metric, even though it does not pass the significance tests, but the 95 percent confidence interval contains the -0.0075 practical significance boundary on the negative side. It would be risky if we launch the change. So my recommendation is that Udacity should not launch the change.

Follow-Up Experiment

We did see a significant drop in the gross conversion rate when adding a quick self-assessment question, but there is no significant change in net conversion rate. Our doubt is that a quick self-assessment question may not be sufficient in letting the students to set up a clear expectation of their time commitment after enroll in the program. Instead of the self-assessment question, we would want to test if showing a 1 minute video, in which the Udacity coach gives a brief overview on the time requirement/ work load to the user and only after that ask the user whether they want to enroll in free trail or just access the materials for free.

The hypothesis is similar to the hypothesis in this test, that when users get a clearer expectations the number of users who left the free trial would be reduced while the same time have no significant reduction in the number of enrollment.

Since this follow-up experiment is just a modified version of this test, we can still use the metrics as well as the unit of diversion we have chosen here.

Reference:
Sign and binomial test online calculator: http://graphpad.com/quickcalcs/binomial2/
A/B test sample size online calculator: http://www.evanmiller.org/ab-testing/sample-size.html