Project Report: Design an A/B Test

Feng Li

Aug 03, 2016

(Project Instructions)

1. Experiment Design

1.1 Metric Choice

1.1.1 Invariant Metrics

1. **Number of cookies**: That is, number of unique cookies to view the course overview page.
2. **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).
3. **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.

Invariant metrics should not change across the experimental and control groups. Because the free trial screener pops out after clicking on the “Start free trial” button, the number of cookies to view the course overview page and the cookies to click the “Start free trial” button should remain unchanged during the experiment. And the ratio of these two variables, click-through-probability, should also remain unchanged.

1.1.2 Evaluation Metrics

1. **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.
2. **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.

Evaluation metrics are expected to show the corresponding changes across the experimental and control groups. Gross conversion and net conversion are chosen as evaluation metrics in that they reflect how much the tested screener influence the enrollment in the free trial and payment after free trial respectively.

For example, if some students indicate fewer than 5 hours available per week, they might decide not to enroll (start free trial) following the screener’s suggestion. So lower gross conversion and net conversion are expected in the experimental group.

1.1.3 Other Metrics

1. **Number of user-ids**: That is, number of users who enroll in the free trial.
This variable alone can't provide useful information about the experiment. We're interested in its proportion in total number of cookies that click on the "Start free trial" button, that is, probability of enrolling.

2. **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.

Retention equals probability of payment divided by probability of enrolling, thus no additional value to keep as an evaluation variable.

If possible, we can also include the free trial drop-out rate and paid drop-out rate, that is, the number of user who quit the free trial or paid learning due to lack of time divided by number of enrollment in each period. With these metrics, we can observe whether the number of frustrated students who quit is reduced as the screener is used.

1.2 Measuring Standard Deviation

In the experiment, the sample size is 5000 cookies visiting the course overview page. This results in 400 cookies that click the "Start free trial" button, and 82.5 enrollments in the free trial, according to the baseline values.

The number of clicks and enrollments follows a binomial distribution, with the standard deviation sqrt(p(1-p)/n). This calculation yields the standard deviation of:

Gross conversion: 0.0202
Net conversion: 0.0156

The analytic estimate would match the the empirical variability in the experiment, because the units of analysis and units of diversion are the same (cookie).

1.3 Sizing

1.3.1 Number of Samples vs. Power

Bonferroni correction is not used, with reasons explained later.

To calculate number of page views required for one group in the experiment, we first calculate the sample size for each evaluation metric, and then pick the largest one.

We allow a type I error alpha = 0.05, type II error beta = 0.2, and the power is 1-beta = 80%. The baseline conversion rates are given in the baseline values table, and the minimum detectable effects are prespecified.

Gross conversion
baseline conversion rate: 20.625%
dmin: 1%
number of clicks: 25835
number of pageviews: 645875

Using a sample size calculator, we can calculate the number of clicks needed: 25835. So the number of pageviews needed for control and experimental groups is 25835/(3200/40000)*2 = 645875.

With the same method, we can calculate the number of pageviews for net conversion.

Net conversion
baseline conversion rate: 10.93125%
dmin: 0.75%
number of clicks: 27413
number of pageviews: 685325

So the number of pageviews required for the experiment is 685325.

1.3.2 Duration vs. Exposure

With 685325 page pageviews required, while unique pageviews per day on Udacity is 40000 in the baseline, 50% of total traffic is needed to be diverted to the experiment, resulting in 35 days to run the experiment. Considering the free trial will last for 14 days, 35 days is long enough to observe the payment after free trial.

The experiment constitutes no greater than minimal risk, because the screener is a mild reminder about time commitment. None of the participants could suffer physical or mental harm as a result of this experiment, nor sensitive data will be gathered.

2 Experiment Analysis

2.1 Sanity Checks

2.1.1 Number of Cookies to View Page and Click Button

We expect the pageviews and clicks are divided evenly between the control and experimental groups. Using an expected rate of diversion of 0.5, we can construct a 95% confidence interval around it. We can check if these invariant metrics are reliable by examining whether the observed rate of diversion is within the confidence interval.The calculation is done using R.

p = 0.5
alpha = 0.05
Z score for 95% confidence interval: 1.96

Number of pageview
confidence interval: [0.4988204, 0.5011796]
observed proportion of experimental size: 0.4994258
pass: Yes

Number of clicks
confidence interval: [0.4958845, 0.5041155]
observed proportion of experimental size: 0.4996278
pass: Yes

2.1.2 Click-through-probability on Button

We expect more or less the same click-through-probability across groups. Using the observed CTP in the control group, we can construct a 95% confidence interval. We can compare the two CTPs by examining whether or not the observed rate in the experimental group lies in the confidence interval.

95% confidence interval of CTP in control group: [0.08121036, 0.08304127]
observed CTP in experimental group: 0.08218244
pass: Yes

2.2 Result Analysis

2.2.1 Effect Size Tests

A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)

Using the given data, we could calculate the difference of each evaluation metric and its variance, and then construct a 95% confidence interval.

Gross conversion
difference: -0.02055487
variance: 1.909799e-05
95% confidence interval: [-0.02912032, -0.01198943]
dmin: ±0.01
statistical significance: Yes
practical significance: Yes

Net conversion
difference: -0.004873723
variance: 1.179217e-05
95% confidence interval: [-0.01160431, 0.001856864]

dmin: ±0.0075
statistical significance: No
practical significance: No

2.2.2 Sign Tests

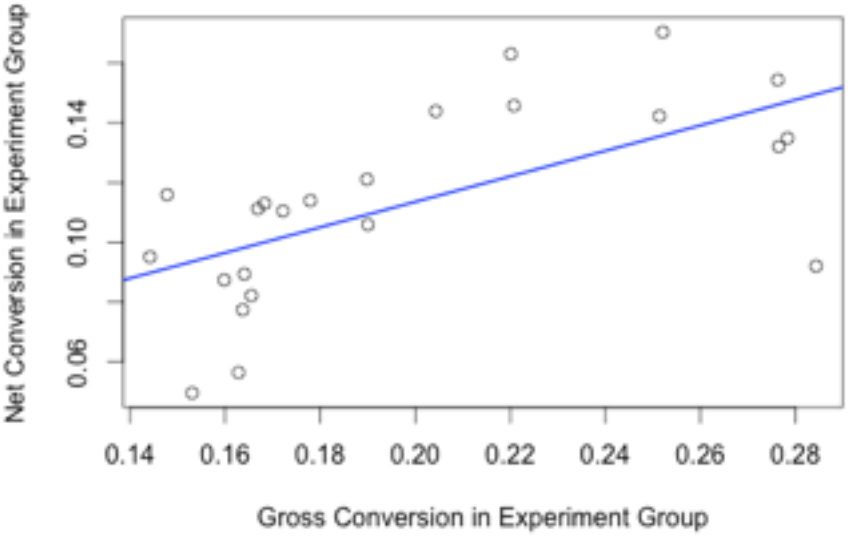
To perform a sign test, we can use the binom.test function in R. More specifically, first we create a column with value FALSE or TRUE indicating if there is a positive or negative difference day-by-day across groups. Then we count the occurrences of TRUE, and get the p-value in the binom.test result.

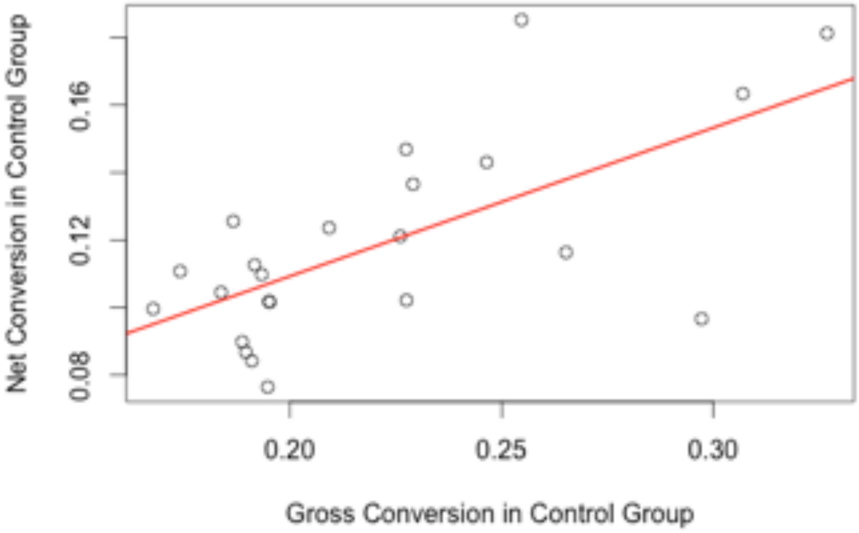
alpha = 0.05
Gross conversion
number of successes = 4, number of trials = 23
p-value = 0.002599
statistical significance: Yes

Net conversion
number of successes = 10, number of trials = 23
p-value = 0.6776
statistical significance: No

2.2.3 Summary

The following are plots of net conversion against gross conversion in control and experimental groups. The coefficient is 0.9869 in control, and 0.9277 in experimental group. So the two evaluation metrics are likely positively correlated.





As stated on wikipedia, Bonferroni correction can be conservative in this situation. The correction comes at the cost of increasing the probability of producing false negative and consequently reducing statistical power. So Bonferroni method is not used during the analysis phase.

The gross conversion rate is both statistically and practically significant, dropping in the experimental group by approximately 2%. So our hypothesis is supported, that the screener will reduce the number of students that enroll from initial click.

Net conversion rate dropped by almost 0.49%, indicating that the screener had a negative effect on the number of students who would pay after the free trial. However, this change is neither statistically or practically significant.

2.3 Recommendation

The screener proved to have a negative effect on the number of students who enroll from initial click on the “Start free trial” button, and also not successful to increase the number of students who pay after free trial. In fact, it appeared to increase the rate at which students quit after free trial.

If Udacity’s goal is only to increase the likelihood to enroll and pay for course, I recommend not to launch this screener.

However, if there are other considerations like to increase the probability of completion, more effectively allocate coach resources, or improve the overall students experience, there should be more follow up experiments and evaluation metrics to observe.

3 Follow-Up Experiment

A follow up experiment could be based on motivation to motivate students to commit more time to online learning. For example, we could send email to those who enrolled in a course but haven’t logged into Udacity for a certain amount of time, say one week, and examine how this would affect the time spent on Udacity.

The unit of diversion is user ID, because the objective of the experiment is the users who have already enrolled in courses, whether in free trial or paid learning.

We could use cohort as sample, that is, users who enroll and enter the experiment at the same time, to ensure the total number of user IDs stays the same.

Another invariant metric is the courses enrolled by each user ID, for different courses may require different amount of devotion, or have different levels of attraction to users.

Hypothesis: sending reminder email to those who enrolled but haven’t logged in for one week could increase the amount of time spent on Udacity.
Unit of diversion: user ID
Invariant metrics: number of user ID, courses enrolled
Evaluation metrics: average time spent by per user ID