Michael Christensen
A/B Testing Final Project
7/13/2016

Experiment Design
Metric Choice

Number of cookies: That is, number of unique cookies to view the course overview page. (d_{min}=3000)

Number of user-ids: That is, number of users who enroll in the free trial. (d_{min}=50)

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). (d_{min}=240)

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.(d_{min}=0.01)

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. (d_{min}= 0.01)

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.(d_{min}=0.01)

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. (d_{min}= 0.0075)

Invariant Metrics: number of cookies, number of clicks, click-through-probability

Evaluation Metrics: gross conversion, retention, net conversion

Invariant Metrics:

Number of cookies: The number of unique cookies that visit the course overview page is not expected to vary as Udacity updates the "start free trial" page as the users have not seen that page before they decide to visit it. This makes it a good invariant metric. The number of unique cookies would make a poor choice for an evaluation metric since the screener pops-up after clicking on the “Start free trial” button.

Number of clicks: The number unique cookies to click the free trial button (before the screener) is appropriate as an invariant metric since the screener pops-up after clicking on the “Start free trial” button. Number of clicks does not work as an evaluation metric by the same logic. That is that the experiment is being carried out after the unique cookie clicks on the “Start free trial” button.

Click-through-probability: The number of unique cookies to click the "Start free trial" button divided by the number of unique cookies to view the course overview page is well-suited as an invariant

metric since it is computed before the screener. The fact that the metric is computed before the screener makes it a poor choice as an evaluation metric.

Evaluation Metrics:

Gross conversion: Gross conversion is a good evaluation metric since the number of users who complete checkout and enroll in the free trial are expected to depend on the screener. This same logic makes gross conversion unsuitable as an invariant metric. If the experiment hypothesis is correct, I would expect gross conversion to be lower as those students likely to dropout during the 14-day trial would be filtered by the screener.

Retention: Retention is a good evaluation metric since the number of user-ids to remain enrolled past the 14-day boundary is expected to depend on the screener. The use of the screener disqualifies retention as an invariant metric. If the experiment hypothesis is correct, I would expect retention rates to be higher since those likely to drop would have not enrolled initially.

Net conversion: Net conversion is suitable as an evaluation metric since the ratio of number of user-ids to remain enrolled past the 14-day boundary to unique cookies to click on the “start free trial” button may depend on the screener. This same logic disqualifies net conversion as an invariant metric. If the experiment hypothesis is correct, I would expect net conversion rate to remain about the same (not statistically different). This could be the combined result of decreased gross conversion and increased retention.

Neither Invariant nor Evaluation Metric:

Number of user-ids: It doesn’t make sense to use user-id as an invariant metric given that user-id is not tracked for users that do not enroll (even if they are signed in when they visit the course overview page). In other words, the number of user-ids will be affected by the screener. As an evaluation metric, number of user-ids is somewhat redundant in that it is already captured by the gross-conversion evaluation metric (number of user-ids to complete checkout and enroll in the free trial divided by the number of unique cookies to click the “Start free trial” button”).

I would consider launching the experiment if either:
1. Retention increased
2. Gross conversion rate decreases and net conversion rate does not decrease. The results must be statistically and practically significant.

In either case it is likely that there are improvements to the student experience as well as the coach’s capacity to support students who are likely to complete the course

Measuring Standard Deviation

Given a sample size of 5000 cookies (in each group) visiting the course overview page it is necessary to scale the unique cookies to click “Start free trial” per day baseline data provided by Udacity.

5000 cookies/40000 cookies =0.125

Unique cookies to click the “Start free trial” per day: 3200

For sample: 3200*.125= 400 unique cookies to click the “Start free trial”

By the central limit theorem, the distribution of the rates (gross conversion, retention, and net conversion) is Gaussian. Standard deviation of these metrics can be calculated analytically using this formula:

std = sqrt(p*(1-p)/N)

The rates for the evaluation metrics are as follows:

Gross conversion: P_gc = 0.20625
Retention: P_r = 0.53
Net conversion: P_nc = 0.1093125

Standard Deviation (with N=400)
Gross conversion: std_gc = 0.0202
Retention: std_r = 0.0549
Net conversion: sdt_nc = 0.0156

The analytically calculated standard deviation for gross conversion and net conversion are likely to match closely with the empirical standard deviation since the unit of diversion (cookie) and unit of analysis (cookie) are the same. For retention, the unit of diversion (cookie) is not the same as the unit of analysis (user-id) and the analytic and empirical calculations of standard deviation will probably not match. If time permits, the empirical value for standard deviation should be computed for the retention metric.

Sizing

Number of Samples vs. Power

I decided not to use the Bonferroni correction during my analysis phase because the evaluation metrics are highly correlated and the resulting solution will be too conservative.

To calculate sample size for each evaluation metric, I used this website: http://www.evanmiller.org/ab-testing/sample-size.html

alpha=0.05 (significant level)
1-beta=0.2 (statistical power)
d_min = minimum detectable effect (pre-defined for each metric)
click through probability on “Start free trial” = 0.08

It should be noted that I multiply by 2 in the solutions below to consider both the control and experiment groups.

Gross conversion: (baseline conversion rate is 0.20625 and d_min is 0.01)
The required number of samples calculated from the online calculator is 25835. This is the number of clicks on "Start free trial" button. In order to get that number, we need
25835 / 0.08 * 2 = 645875 page views.

Retention: (baseline retention rate is 0.53 and d_min is 0.01)
The required number of samples calculated from the online calculator is 39115. This is the number of users who finished the 14 day free trial. In order to get that number, we need 25835 / 0.08 * 2 = 4741212 page views.

Net Conversion: (baseline conversion rate is 0.1093125 and d_min is 0.0075)
The required number of samples calculated from the online calculator is 27413. This is the number of clicks on the "Start free trial" button. In order to get that number, we need 27413 / 0.08 * 2 = 685325 page views.

Page views required to conduct the experiment: 4,741,212

Duration vs. Exposure

Given the relatively large number of page views required for the retention metric, I have decided to drop it at this point. It would require about 119 days to accumulate the required page views assuming 40,000 views per day. An experiment of this length could be costly to upkeep and may prevent other experiments from being run.

After dropping retention as an evaluation metric, I am left with gross conversion and net conversion. I choose net conversion to calculate duration since it has the larger required number of page views and is thus the limiting metric. In identifying what proportion of traffic to test the free trial screener on, it is necessary to consider various forms of risk. The experiment will not cause any physical, psychological, emotional, social, or economic harm to the participant. The information collected isn't sensitive and therefore, privacy and confidentiality aren't issues. Given this minimal risk, it is appropriate to run 100% of my traffic though the experiment (50% control group and 50% experiment group). I assume that since the experiment setup is fairly simple, there is no appreciable threat of a user interface bug popping up. If 100% of total traffic is diverted (40,000 unique cookies) then experiment duration will be ~18 days. This span of time allows the experiment to be carried out over multiple weeks and on both weekdays and weekends. This is desirable since it smooths out the effects of temporal variability in user behavior.

Experiment Analysis

Sanity Checks

For the invariant metrics I expect a roughly equal split of events between the experiment and control groups (probability = 0.5). Below I perform sanity checks of observed values using a 95% confidence interval around the expected value.

Number of cookies (Pass):
control group total: 345543
experiment group total = 344660
standard deviation = sqrt((0.5*0.5) / (345543 +344660)) = 0.0006018
margin of error = 1.96 * 0.0006018 = 0.0011796
lower bound = 0.5 – 0.0011796 = 0.4988
upper bound = 0.5 + 0.0011796 = 0.5012
Observed value: 0.5006

Number of clicks (Pass):
control group total = 28378
experiment group total = 28325
standard deviation = sqrt((0.5 * 0.5) / (28378 + 28325)) = 0.0021
margin of error = 1.96 * 0.0021 = 0.0041
lower bound = 0.5 - 0.0041 = 0.4959
upper bound = 0.5 + 0.0041 = 0.5041
observed = 28378 / (28378 + 28325) = 0.5005

Click through probability (Pass):
control value = 28378 / 345543 = 0.0821258
standard deviation = sqrt((0.0821258 * (1-0.0821258)) / 344660) = 0.000468
margin of error = 1.96 * 0.000468 = 0.00092
lower bound = 0.0821258 - 0.00092 = 0.0812
upper bound = 0.0821258 + 0.00092 = 0.0830
experiment value = 0.0821824

Result Analysis

Effect Size Tests

Gross conversion (Statistically Significant and Practically Significant)

I append the following to denote the group (control or experiment):

_cnt (control group)
_exp (experiment group)

N_cnt = clicks_controlled = 17293
X_cnt = enroll_controlled = 3785
N_exp = clicks_experiment = 17260
X_exp = enroll_experiment = 3423

p_pooled = (X_cnt + X_exp) / (N_cnt + N_exp) = 0.2086
se_pooled = sqrt(p_pooled * (1-p_pooled) * (1/N_cnt + 1/N_exp)) = 0.00437

Probability difference:
d = (X_exp / N_exp) – (X_cnt / N_cnt) = -0.02055

Upper and lower bounds:
lower = d - se_pooled = -0.0291
upper = d + se_pooled = -0.0120

The interval does not contain 0 and is therefore statistically significant. The interval also does not include ± .01 and is thus practically significant (d_min =0.01).

Net conversion (Not Statistically Significant or Practically Significant)

N_cnt = clicks_controlled = 17293
X_cnt = pay_controlled = 2033
N_exp = clicks_experiment = 17260
X_exp = pay_experiment = 1945

p_pooled = (X_cnt + X_exp) / (N_cnt + N_exp) = 0.1151
se_pooled = sqrt(p_pooled * (1-p_pooled) * (1/N_cnt + 1/N_exp)) = 0.00343

d = (X_exp / N_exp) – (X_cnt / N_cnt) = -0.0048

lower = d - se_pooled = -0.0116
upper = d + se_pooled = 0.0019

The interval is not statistically or practically significant as it contains 0.

Sign Tests

In order to perform a sign test I calculate the value of the metric for each day and then compare the control and experiment groups on this basis. I use the calculator at the below link to determine the p-value of my result (assuming 0.5 as the probability):

http://graphpad.com/quickcalcs/binomial1.cfm

Gross conversion (Statistically Significant):
Gross conversion is larger for the experiment group 4 out of the 23 days of the experiment. The two-tail P value is .0026. This is significant since the result is smaller than p=.05 (alpha).

Net conversion (Not Statistically Significant):
Net conversion is larger for the experiment group 10 out of the 23 days of the experiment. The two-tail P value is .6776. This is not significant since the result is larger than p=0.05 (alpha).

Summary

The Bonferroni correction was not used in the analysis because I am basing my launch decision on the statistical and practical significance of two individual metrics (gross conversion and net conversion). This correction may have been appropriate if were basing my launch decision on only the movement of one metric out of several. In the context of this experiment, in order for the screener to be launched my requirement was for gross conversion to decrease while net conversion increased (statistical and practical significance required). The effect size tests showed that gross conversion was both statistically and practically significant. This result supported the hypothesis that the screener was effective at reducing the number of enrolling students who otherwise may have dropped out during the free trial period since they could not make the time commitment. I found that the difference in net conversion rates between the control and experiment groups was not statistically significant and had a 95% confidence interval that extended into negative territory. This finding illustrates that the screener may reduce the number of students that would go on to finish the 14-day free trial and make a payment. Reducing revenue is not a desired outcome of the screener.

The result of the sign test was the same as the effect size test. The difference in gross conversion rates between the control and experiment groups is statistically significant. The difference in net conversion rates between the groups is insignificant.

Recommendation

The initial hypothesis was:

The screener 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.

Given this hypothesis, I believe that significant decrease in gross conversion in the experiment group qualifies as a positive and expected outcome. The insignificant difference in number of user-ids to remain enrolled past the 14-day boundary and make a payment also falls in line with the hypothesis. Since the 95% confidence interval for net conversion does include the practical significance boundary, however, it is possible that the metric decreased by an amount that matters to Udacity. This possibility is not an acceptable risk and more testing is required before launching the screener.

Follow-Up Experiment

Following enrollment in the free trial, I would like to play a video providing an overview of the course and the material it will cover. This video will be informative and provide tangible, real-world examples illustrating the applicability of the skills that the student will learn. The idea is that setting expectations upfront and linking class material to exciting real-world problems could invigorate student curiosity and motivate them to finish coursework. The end result is increased student retention as well as lowered costs and increased revenue for Udacity.

Hypothesis: Including an introductory video will increase student retention by a practically significant amount.

Setup: After enrolling, students will be randomly assigned into either the control or the experiment group. The students in the experiment group will be presented with an introductory video while the students in the control group will not.

Unit of diversion: This will be user-id as this is tracked after a student enrolls.

Invariant metric: Number of user-ids that enroll is a good invariant metric since this number isn't affected by the experiment.

Evaluation Metric: Retention will be the evaluation metric since this is what I am trying to improve.

If the experiments results in increased retention (that is statistically and practically significant (d_min=.01)) then I will launch the change. Increased retention will reduce costs and increase revenue.