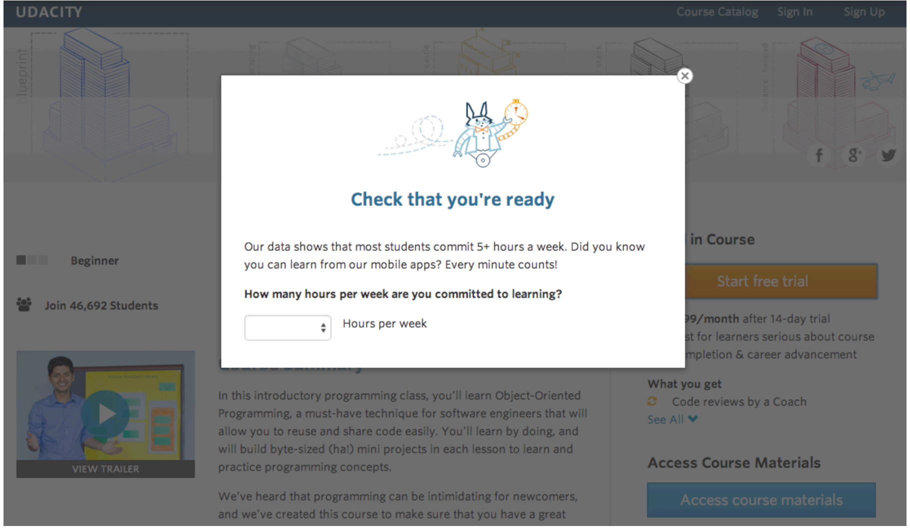
A/B Test Final Project

(For Udacity Nanodegree)
Jie Hu, jie.hu.ds@gmail.com

Summary

Udacity is a for-profit educational organization, well known for its professional online coaching and learning of technical skills. It offers course materials for free but charged for items including verified certification, coaching and project review. Students can choose to start a course for free and will be charged after 14 days. Since most of the courses require hard work with sufficient amount of time, it will be beneficial to let students know before enroll a course, so that less proportion of students will leave before free trial ends and hence save resources such as coaching to students who were more likely to pay and even complete the course. Therefore, a screener between click "Start free trial" and filling checkout information is designed.



Check that you're ready

Our data shows that most students commit 5+ hours a week. Did you know you can learn from our mobile apps? Every minute counts!

How many hours per week are you committed to learning?

Hours per week

99/month after 14-day trial
st for learners serious about course
mpletion & career advancement

In this introductory programming class, you'll learn Object-Oriented Programming, a must-have technique for software engineers that will allow you to reuse and share code easily. You'll learn by doing, and will build byte-sized (ha!) mini projects in each lesson to learn and practice programming concepts.

We've heard that programming can be intimidating for newcomers, and we've created this course to make sure that you have a great

After click "Start free trial" button, the screener will show up and ask how many hours a student will be committed to learning? If answer less than 5 hours, the student will be recommended to use free materials access (though checkout is also accessible), else student will go through normal checkout process and fill the information of payment.

In this project, I use A/B test to figure out whether it's really worth to add the screener. The steps are:

- Experiment Design: select metric based on business target, measure standard deviation and recommend the least size of the A/B test to reach enough power. Then determine the duration and percentage of traffic the test will be exposed to.
- Experimental Analysis: check sanity and analyze the result of the test, including if there's a significant difference, and do sign check to see in details.

Experiment Design

Metric Choice

- Evaluation Metrics: Gross conversion, Retention, Net conversion
- Invariants: Click-through-rate, Number of cookies (view the course overview), Number of clicks (on “start free trial”)

Recall that the launch criteria for metrics here are required by our targets:

- Decrease the number of frustrated students who leave free trial because they don’t have enough time
- Do not reduce the number of students to continue past the free trial and pay for the course

All the criteria for metrics should be based on these 2 target, so if under the hypothesis, the metric is relevant to these two targets, I will choose as evaluation metrics, if independent to the hypothesis, then I will choose as invariants.

Now consider each of the metrics we have here:

- **Number of cookies** (*Number of unique users to visit the course overview page*): it happens before screener and independent of our experiment, so I choose as an **invariant**
- **Number of User-ID**: If our hypothesis is true that screener will discourage some students to enroll the course, the number of user-ids of students who choose to enroll the course will certainly be changed, so it’s a good choice as evaluation metrics. However, it’s better to use normalized, or a probability here because we want to know how much the screener can contribute to the change we want, so here I won’t use it as evaluation metric. A better choice will be gross conversion rate as below.
- **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. It’s a pretty good invariant because it won’t be significantly changed with or without experiment, and it’s normalized and will be more accurately show the change. So I will choose as an **invariant**.
- **Number of clicks** (*Number of unique cookies to click the start free trial button*): it happens before screener and independent of our experiment, so I choose as an **invariant**
- **Gross conversion** (*Number of users who enrolled in the free trial/Number of users who clicked the Start Free Trial button*): It directly depends on the experiment because, under my hypothesis, after click, the screener should have negative effect and decrease the number of customers who are unlikely to unroll and even pay. So I choose as **evaluation metrics**. In my experiment, control group without screener is assumed to have higher gross conversion than experiment group with screener.
- **Retention** (*Number of user-ids to remain enrolled for 14-day trial period and make their first payment/Number of users who enrolled in the free trial*): Similar as mentioned in gross conversion metric, I assume users in experiment

group will be aware of time commitment to this course and thus retention rate will be higher due to less proportion of customers who enroll and left without payment. So I choose retention as an **evaluation metric**.

- **Net conversion** (*Number of user-ids remained enrolled for 14-day trial and at least make their first payment/Number of users clicked the Start Free Trial button*): under our hypothesis, we don't expect to see big changes in number of customers enrolled the course, and therefore, an increased or unchanged net conversion will be welcome. So I choose net conversion as one of the **evaluate metrics**

Measuring Standard Deviation

With 5000 cookies, I expect to see:

$$\frac{5000 * 3200}{40000} = 400$$

enrollments (start free trial).
Since the metrics are probability, each customer can be considered independent with others and there're 2 possible outcomes for each metrics, binomial distribution will be a good fit and I apply the formula:

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

And get standard deviation:

- Gross conversion: 0.0202
- Net conversion: 0.0156
- Retention: 0.0549

For gross conversion and net conversion, they use number of cookie (unit of diversion) as denominator, and such denominator is equal to unit of analysis, therefore the analytical estimate will be comparable to the empirically variability.

For retention, the denominator is "Number of users enrolled the course" which is different from the unit of diversion, therefore, the analytical estimate and empirical variability are different.

Sizing

Number of Samples vs. Power

I do not use Bonferroni correction for all the below steps because in this experiment, all the metrics are highly correlated.

Because there were too big size of retention, 4,741,212 pageviews are required and the duration for even 100% traffic will be over 100 days, so I finally removed this metrics and turned to gross and net conversion, which have much smaller size.

Here're my calculation by Evan's Awesome A/B Tools:

- Gross conversion: baseline conversion rate: 20.625%, minimum detectable effect is 1%, Power: 80%, alpha: 5%

  Sample size is 25835 * 2 = 51670 enrollments, and by:
  $\frac{PV}{51670} = \frac{40000}{3200}$, I get 645875 pageviews are needed.

- Net conversion: baseline conversion rate: 10.93%, minimum detectable effect is 0.75%, Power: 80%, alpha: 5%

  Sample size is 27413 * 2 = 54826 enrollments, and by:
  $\frac{PV}{54826} = \frac{40000}{3200}$, I get 685325 pageview are needed.

I choose the larger size, which is 685,325 page views are needed.

Duration vs. Exposure

I choose 60% of traffic to be tested, and it needs 685325 / 22000= 29 days.
I don't think there's big risk because there's no big cost to make such change and diversion. It's a kind of slight UI change.

Risk Analysis

As we are not dealing with data (Political attitudes, personal disease history, sexual preferences), we are not specifically extract privacy information, and because it's a recommendation to help users know more about the course, so no one will get hurt by such experiment and there's not big risk to run such test. Besides, the duration of such test is within 30 days, which will be a low risk acceptable experiment.

There is one tiny problem come to my mind that will be a little risky: Getting less customers start free trial for the company. However, this save a lot of investment on employing more coaches to help students who actually are not interested complete the course. So such problem can certainly be ignored.

Experiment Analysis

Sanity Checks

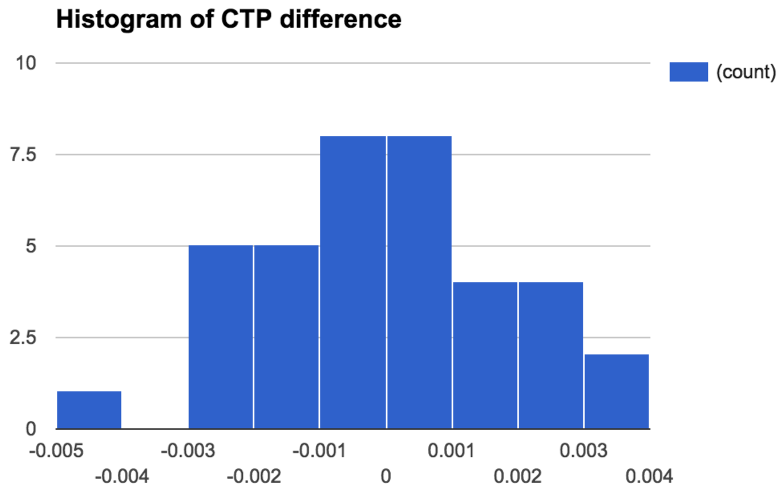
Sanity check is a necessary step to check whether invariants stay the same after we implement experiment. If not, we can not make any conclusion based on the experiment.

1. Number of cookies:
Total Control group pageview: 345543
Total Experiment group pageview: 344660
Total pageview: 690203
Probability of cookie in control or experiment group: 0.5
SE = sqrt(0.5*(1-0.5)*(1/345543+1/344660) = 0.0006018

Margin of error (me) = SE * 1.96 = 0.0011796
Confidence Interval = [0.5 - me, 0.5 + me] = [0.4988, 0.5012]
Observed value = 344660/690203 = 0.5006

2. Number of clicks:
Total Control group clicks: 28378
Total Experiment group cliks: 28325
Total clicks: 56703
Probability of cookie in control or experiment group: 0.5
SE = sqrt(0.5*(1-0.5)*(1/28378+1/28325) = 0.0021
Margin of error (me) = SE * 1.96 = 0.0041
Confidence Interval = [0.5 - me, 0.5 + me] = [0.4959, 0.5041]
Observed value = 28378/56703 = 0.50046

3. Click-through-probability



It seems a normal distribution around 0. Here I run a test:
Probability of cookie in control group: 28378/345543 = 0.0822
SE = sqrt(0.0822 * (1-0.0822)/ 345543) = 0.000467
Confidence interval = [0.0822 - 1.96*SE, 0.0822 + 1.96*SE] = [0.0812, 0.0830]
Observed value in experiment group: 28325/344660 = 0.0822, which is within confidence interval

Because all confidence intervals indicate we failed to reject null hypothesis, so all sanity checks passed.

Result Analysis

Effect Size Tests

- Gross conversion:

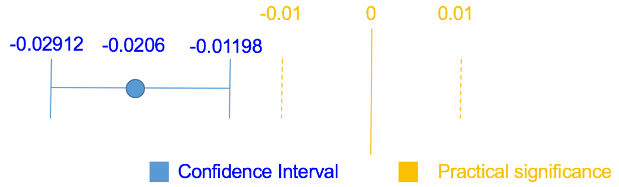
Group	Enroll	Click	Gross Conversion	Difference
Control	3785	17293	0.21887	-0.02055
Experiment	3423	17260	0.1983198	

$$\hat{p}_{pool} = \frac{3423 + 3785}{17293 + 17260} = 0.2086071$$

$$S_{pool} = \sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})(\frac{1}{17293} + \frac{1}{17260})} = 0.00437$$

95% confidence interval is:

(d – 1.96*0.00437, d + 1.96 * 0.00437) = (-0.02912, -0.01198)



From the plot we can see gross conversion is both statistically and practically significant

- Net conversion:

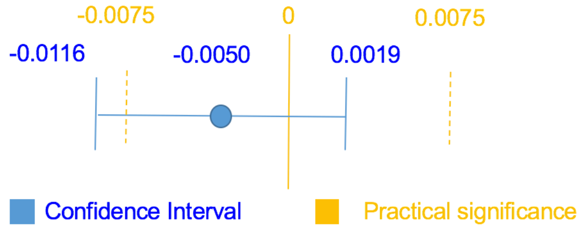
Group	Enroll	Click	Gross Conversion	Difference
Control	2033	17293	0.117562	-0.0048737
Experiment	1945	17260	0.1126883	

$$\hat{p}_{pool} = \frac{2033 + 1945}{17293 + 17260} = 0.1151275$$

$$S_{pool} = \sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})(\frac{1}{17293} + \frac{1}{17260})} = 0.0034$$

95% confidence interval is:

(d – 1.96*0.0034, d + 1.96 * 0.0034) = (-0.0116, 0.0019)



Net conversion is neither practically significant nor statistically significant. Further analysis or test is needed.

Sign Tests

- Gross conversion:
  4 out of 23 days are positive, which has p-value of 0.0026 < 0.05 if we consider it as flip a coin and do the random test. So gross conversion metric does show a significant difference between control and experiment groups
- Net conversion:
  10 out of 23 days are positive, which has p-value: 0.6776 > 0.05, if we consider it as flip a coin and do the random test. So net conversion metric doesn't show a significant difference between control and experiment groups

Summary

Though as the number of metrics increases, it's more likely to inflate the false positive rate, I didn't use Bonferroni correction because we hope gross conversion decrease while keep net conversion same or increasing. I would expect both of these two metrics meet the expectations, not any of them. So Bonferroni correction is not necessary here.

Recommendation

Final decision:
Based on our analysis, it's delighting to see there's significant decrease in the proportion of students who left free-trial, the screener will save a lot of money and time. And the sign-test results are encouraging, too. However, one tradeoff to be noticed is that net conversion in effect size test is neither practical significant nor statistical significant, and the confidence interval involve parts either in or outside practical interval, which means **there will be a high risk to lose customers if we launch the screener**.

From the experiment results, if we use normal distribution to estimate the probability to make such mistakes, we have:
$Z^* = (-0.0075+0.005)/0.0034 = -0.7353$
And the corresponding probability (one-sided) is 23.1%! We have big probability to make such mistake and lose at least 0.75% customers in average. If we have 40,000 pageviews each day, 23.1% the chance, we might lose **at least**: 40000*0.0822*0.75% = 24.66 customers. Besides, we never know if in other dimensional thinking, if it's worth to lose these customers. It quite possible some customers might not finish the course, but they might be likely to introduce the course to other students if they left only because they didn't have enough time. So it's better to think more in a bigger picture.

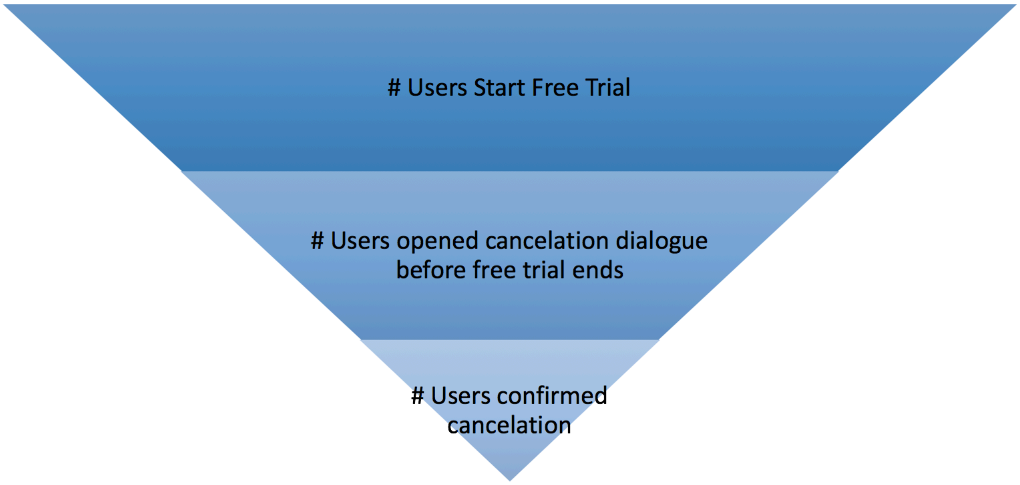
In the next section I will discuss a possible solution to do further experiment to reduce losing customers in free trial stage. However, I would like to recommend not to launch this screener in a hurry based on above analysis.

Follow-Up Experiment

In order to reduce early cancellations, here I design another experiment:

Business Objective: decrease proportion of frustrated students who leave the course before free trial ends.
Design: add a window to show the help methods frustrated students can immediately get help. For example, link to specific topics about how to learn this course well in forum, contact coach or / and email us.
Hypothesis: a new dialog window will decrease the percentage of the students leaving before free trial ends
Unit of diversion: user_id because it's more stable to track the user

The funnel here will be simple:


# Users Start Free Trial
# Users opened cancelation dialogue before free trial ends
# Users confirmed cancelation

And based on the funnel, we have our metrics:

Evaluation metric:
# Users confirmed cancelation before free trial ends

Invariants (Both invariants happen before cancelation, so independent to the experiment):

1) # Users start free trial
2) # Users opened cancelation dialog before free trial ends

After users start free trial, their user-ids will be tracked.
Experiment Group: after the users click and open the cancelation dialogue, they will be exposed "help dialogue" to encourage them to ask for help and continue learning, however, there's still choice to continue cancelation after which leads to cancelation process
Control Group: after the users click and open the cancelation dialogue, they will be lead to cancelation process without "help dialogue".

Reference

Online Test Size Calculator: http://www.evanmiller.org/ab-testing/sample-size.html
Udacity A/B Test Course in Nanodegree:
https://classroom.udacity.com/nanodegrees/ud120-nd/syllabus