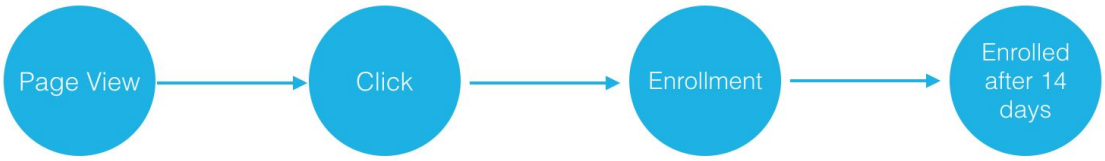
A/B Test Project

Introduction

This project is one of the projects for Udacity Data Analysis Nanodegree program. In this project, an A/B test experiment was conducted to study whether a change on the Udacity website improves the website efficiency. Udacity provide following description of their experiment:

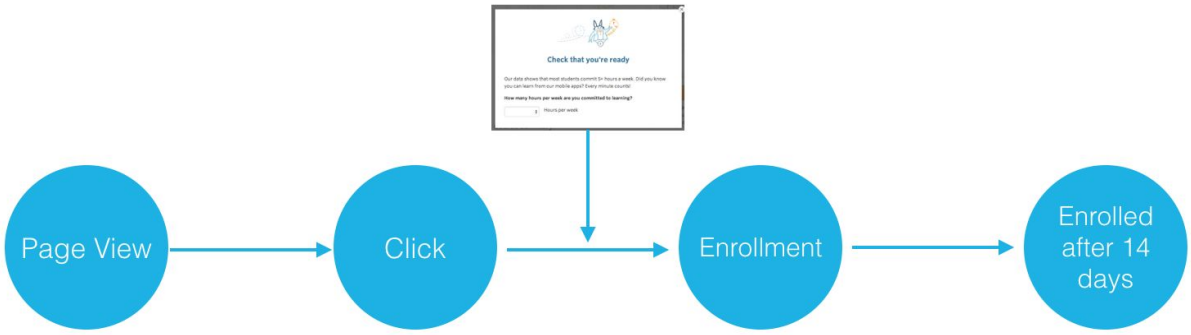
“At the time of this experiment, Udacity courses currently have two options on the home page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the 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.”

The original setting of Udacity website can be illustrated as the chart below:



“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 change Udacity wanted to test is to add a ‘free trial screener’ after the click step. The new setting is illustrated as below:



The intention of this changing is to help student set clearer expectations to reduce the number of students who started the free trial but didn't get enrolled. Udacity also cared about the effect of this change on the number of student who eventually completed the course. So the hypothesis of the this experiment is that this change will "reduce 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.

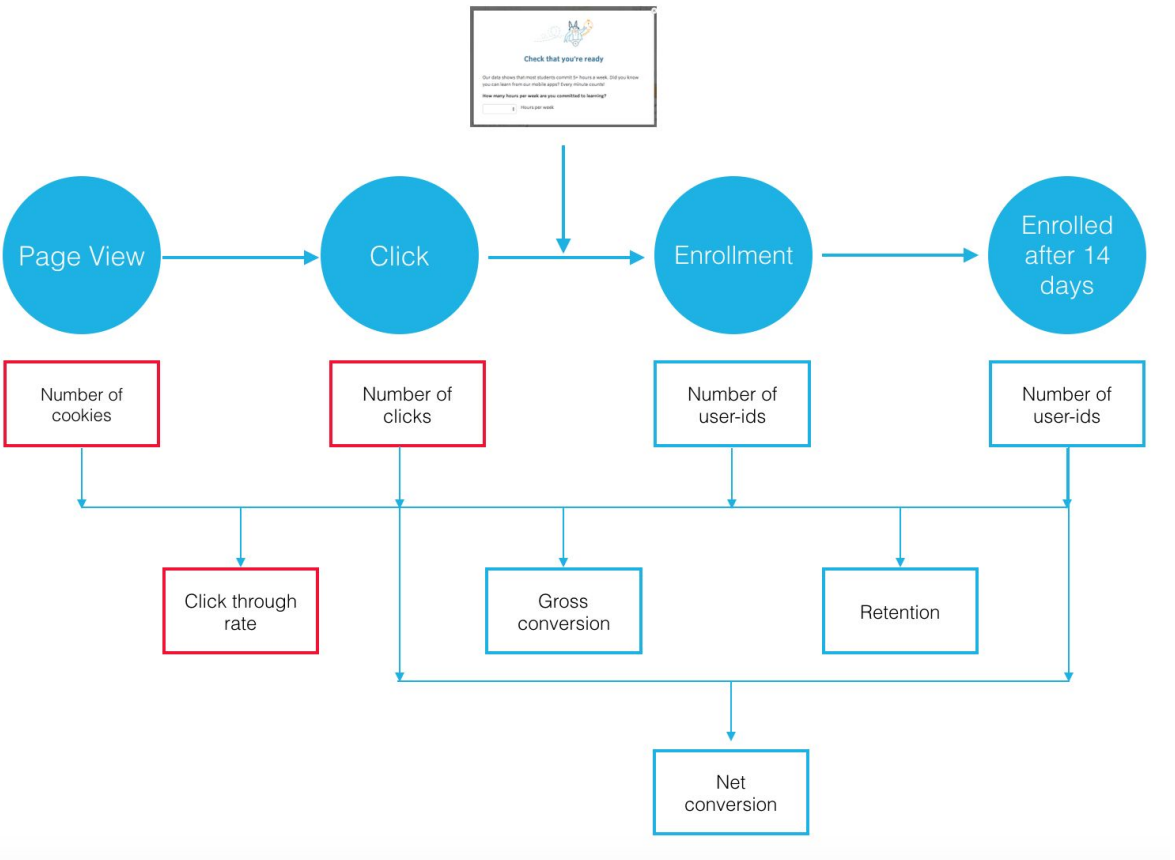
Experiment Design

Metric Choice

Data were collect along with the user's activities. Some metrics for this project:

- 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)



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

According to above chart, three metrics in red boxes are measured before adding the screener, therefore they are invariants.
Recall the goals of adding the screener are 1) decrease the number of enrollment 2) without significantly drop the number of students who stayed after the free trial. So our evaluation metrics are gross conversion (number of enrollment/ number of click), and Net conversion (number of stay/ number of click). To prove the effect of the screener, we expected to see a

decrease on gross conversion and an increase (or not significant decrease) on the net conversion. Therefore, our hypotheses for these tests are:

For Gross conversion: $H_0: D \geq -0.01, H_a: D < -0.01$;
For Net conversion: $H_0: D \leq 0.0075, H_a: D > 0.0075$;

Where D= rate after applying screener (experiment) - rate before applying screener (control)

Measuring Standard Deviation

From the baseline values, we found the rates for gross conversion and net conversion are 0.20625 and 0.1093125 respectively. The corresponding number of cookies clicked free trial of 5000 cookies viewed page is 400. Therefore the standard deviations of the two metrics can be calculated as below:

Metrics	Probability	N	Standard Deviation*
Gross Conversion	0.20625	400	0.0202
Net Conversion	0.1093125	400	0.0156

* $sd = \sqrt{\frac{p(1-p)}{N}}$

Sizing

Number of Samples vs. Power

Metrics	Probability	α	Power	Delta	N
Gross Conversion	0.20625	0.05	0.8	0.01	645875
Net Conversion	0.1093125	0.05	0.8	0.0075	685325

For Net Conversion, let p_0 = net conversion rate of control group, p_1 = net conversion rate of experiment group. The test hypothesis is:
$H_0: d_0 = p_1 - p_0 = 0 \quad H_a: d_1 = p_1 - p_0 = 0.0075$
Under H_0, d_0 ~ N(0, $\sqrt{2 * p(1-p)}$)

Under H_1, d_1 ~ N (0.0075, $\sqrt{p(1-p) + (p+0.0075)(1-p-0.0075)}$)

$Z_{0.975}\sqrt{\frac{2p(1-p)}{n}} = 0.0075 - Z_{0.8}\sqrt{\frac{p(1-p)+(p+0.0075)(1-p-0.0075)}{n}}$, solve for n we got:
n=25835, which is the number of clicks we need for each group.
To calculate the pageview, we need to refer to the baseline information, from which we know:
N=25835*2/(3200/40000)= 685325

Bonferroni correction is not necessary in this experiment, since both metrics criterions need to be satisfied to launch the click feature.

Duration vs. Exposure

The daily traffic of Udacity UI is 40000 pageviews. I choose to use 50% of the daily traffic and run the experiment for 32 days to get enough pageviews.

Experiment Analysis

Sanity Checks

Sanity check is a necessary step in AB test to make sure the data captured is the same across experiments. For this experiment, sanity checks were conducted on three invariant metrics: page views, clicks, and CTR.

	Pageviews	**Clicks**	**CTR**
Control group	345543	28378	0.0821
Experiment group	344660	28325	0.0822
Total	690203	56703	
Conditions	P=0.5	P=0.5	diff=0
SE	0.0006	0.0021	0.000661
Margin error	0.0012	0.0041	0.0013
Confidence interval	(0.4988, 0.5012)	(0.4959,0.5041)	(-0.0013, 0.0013)
Observed value	0.5006	0.5005	0.0001
Sanity Checks	**Pass**	**Pass**	**Pass**

The calculations can be found in R code.

Result Analysis

Effect Size Tests

- Gross conversion:

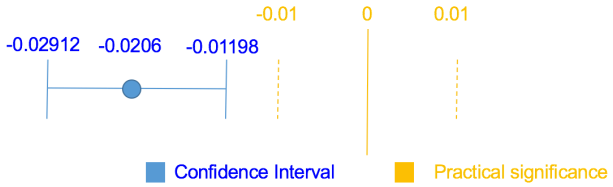
Group	Enroll	Click	Gross Conversion	Difference
Control	3785	17293	0.21887	-0.02055
Experiment	3423	17260	0.1983198	

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

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

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:

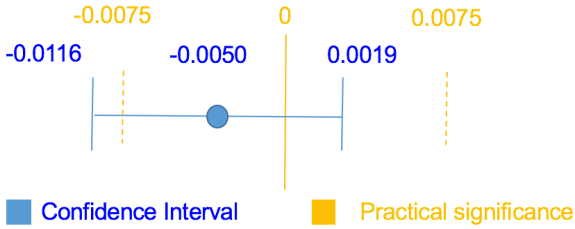
Group	Enroll	Click	Gross Conversion	Difference
Control	2033	17293	0.117562	-0.0048737
Experiment	1945	17260	0.1126883	

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

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

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.