Design an A/B Test
Udacity Nanodegree
Final Project

A/B Testing is the test that we want to test for particular product. Usually A/B testing works for testing changes in elements in the web page. A/B testing framework is following sequence:

- Design a research question.
- Choose test statistics method or metrics to evaluate experiment.
- Designing the control group and experiment group.
- Analyzing results, and draw valid conclusions.

A/B testing is used to validate whether the changes that we have applied in our product is significantly affected our users instead of relying solely on the expert opinion.

This project is about Udacity's Free-Trial-Screening regarding students taking online courses and the screening is based on student's awareness of minimum "5 or more hours per week" time requirement to enroll in the course.

Research question

"The hypothesis was that this 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".

Experiment Design

Choice of Metrics:

- **Number of cookies**: is good invariant metric. The course overview page is in the Home page, it is the page that has the "start-free-trial" button and "access course materials". So the users don't see the experiment and control, which means number of cookies will not be affected by the Free Trial Screener at all. Moreover, it is chosen as the unit of diversion in this case and it should not vary.

- **Number of clicks**: is good invariant metric. A click happen before user is aware of the experiment, that is before they decide to click the "start-free-trial" button hence it is independent.
- **Click-through-probability**: is good invariant metric. Similarly, users don't see the "start-free-trial" button before they first browse the home page and make a click. In addition, as Click-through-probability is computed as a fraction of number of clicks and cookies, which are good invariant metrics.
- **Number of user-ids:** is neither invariant metrics nor good evaluation metrics. The user would be assigned to a group when they first logged in. This means the enrollment depends on the rendering of "start free trial" page and this result to different values in the control and experiment group. Since it varies, it is not an invariant metric. Also, the number of users who enrolled can change a lot with respect to number of clicks on the same day. To avoid redundancy and to marginalize the values, it is better to use Gross conversion (number of user-ids divided by number of "start free trial" clicks) than user-id.
- **Gross Conversion**: is a good evaluation metric. The number of users who decide to start the free trial are affected by the screener. For instance, the message box that suggested a commitment of "5 or more hours per week" filters out, and probably decreases, the users who don't have that much amount of time. From the business point of view, this might show whether cost of enrolment is successfully reduced. In other words, it is a good metric to check if the experiment makes a significant difference in the enrollment. Although number of clicks is invariant, enrollment is affected and this makes Gross Conversion not a good invariant metric.
- **Retention**: is a good evaluation metric but not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment. It is a good evaluation metric to check if the experiment makes a significant difference either in the enrollment or payment. In other words, the screener might reduce enrollment (the number of users who feels they don't have the required time commitment may decide not to enroll in the free trial) or show a positive payment guarantee (only few users decide to enroll in the free trial knowing the time commitment requirement, are more likely to remain enrolled after the 14-day period and make a payment).
- **Net conversion:** is a good evaluation metric . It is the product of Gross Conversion and Retention and is directly dependent on the effect of the experiment.

For this experiment, I choose number of cookies, number of clicks and click-through-Probability as Invariant metrics and gross conversion, retention and net conversion as Evaluation metrics.

Our goal is 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. We will launch only if gross conversion significantly decrease but net conversion either remain constant or doesn't decrease.

Measuring Standard Deviation

Standard deviation = sqrt(p(1-p)/N)
Sample size = 5000

Gross conversion:

Unique cookies to click "Start free trial" per day= 3200
Probability of enrolling, given click= 0.20625
Click-through-Probability = 0.08
Number of Clicks = 0.08*5000 = 400

Standard deviation = Sqrt(0.20625(1-0.20625)/400)= 0.0202

Retention:

Probability of payment, given enroll = 0.53
Probability of enrolling, given click= 0.20625
Number of enrollment (N) = 400 * 0.20625 = 82.5

Standard deviation = Sqrt (0.53 (1-0.53)/ 82.5) = 0.0549

Net Conversion:

Unique cookies to click "Start free trial" per day =3200
Probability of payment, given click = 0.1093125

Standard deviation = Sqrt (0.1093125 (1-0.1093125)/ 400)= 0.0156

When unit of analysis is same as unit of diversion, variability tends to be lower and closer to analytical estimate. In this case, our unit of diversion is number of cookies, and both Gross Conversion and Net Conversion are using number of cookies as denominator. This indicates that analytical estimate would be comparable to the empirical variability.

For Retention, the denominator is " number of user-ids to complete checkout. " and not same as the unit of diversion (number of cookies). The unit of analysis and the unit of diversion are not the same; hence the analytical and the empirical estimates are different.

Sizing

Number of Samples vs. Power

Bonferroni can't be used when metrics has high correlation as it's characterized as too conservative and strict to make the metrics pass. There is high correlation between metrics in our case, so I decided not to use Bonferroni correction.

Using Evan Miller Calculator (http://www.evanmiller.org/ab-testing/sample-size.html), the number of pageviews to power the experiment gives the following result for each evaluation metrics.

Gross Conversion :

Probability of enrolling, given click (base line conversion rate) = 20.625%
Minimum detectable effect (dmin)= 1%
Samples needed: 25,835
Minimum page view for each group: 25,835* (40000/3200) = 322,937.5

Net Conversion:

Probability of payment, given click (base line conversion rate) = 10.93125%
Minimum detectable effect (dmin)= 0.75%
Samples needed: 27,413
Minimum page view for each group: 27,413* (40000/3200) = **342,662.5**

Retention:

Probability of payment, given enroll (base line conversion rate) = 53%
Minimum detectable effect (dmin)= 1%
Samples needed: 39,115
Minimum page view for each group: 39,115* (40000/660) = **2,370,606**

The number of required pageviews for Retention is too large, which makes it unrealistic to be taken as evaluation metric due to the longer time it requires. Therefore, we use gross conversion and net conversion as evaluation metrics and we take the largest sample (27,413) in order to figure out the maximum duration the experiment might take. The samples result is for one group, so we have to double it for both groups. That is;

Total page view for both experiment & control group = 2*342662.5 = **685,325**

Duration vs. Exposure

Making a decision regarding how long our experiment need to take, and how many users we want them to see the experimental features is an important aspect because it affects our experiment.

With daily traffic of 40000, I would direct all of them (100%) to the experiment. This means it would take us approximately **18 days** (685325/40000 = 17.133) for the experiment.

The whole experiment is not considered risky in the sense that neither the users nor Udacity is negatively affected by this. Students are enrolled with the awareness that time dedication is required and if they are willing to proceed and enroll depends on their ultimate choice and they have the option to access course materials if otherwise. In addition, students are not required to give personal information and there is no issue of privacy. In this case, the experiment won't bring any harm from the students' side. Udacity website is also on the safe side, no additional expenses and no changes in database is required to be made due to the experiment.

Experiment Analysis

Sanity Checks

Invariant metrics for this experiment are already identified. But at this stage we can't directly interpret the results. We have to do sanity checks for the invariance metric in order to see if the results are comparable for both experiment and control groups, to make sure that the filtering is the same for both, and the data captured is the same across experiments.

Probability of cookies and clicks in control or experiment groups = 0.5
S.E = sqrt(p(1-p)/n) = sqrt(0.5 * 0.5 / (nctrl + nexp))
Margin of error (me) = 1.96 * S.E
Confidence Interval = (0.5 – me, 0.5 + me)
Actual value observed = number of assignments to control group / number of total assignments

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+344660))) = 0.0006018
Margin of error (me) = 1.96 *SE = 0.0011796
Confidence Interval = (0.5- 0.0011796, 0.5+ 0.0011796)] = **(0.4988, 0.5012)**
Observed fraction (p(hut)) = 344660/690203 = 0.5006

Since the observed fraction of successes or cookies in the controlled group fall within the confidence interval, number of cookies has **passed sanity check**.

Number of clicks
Total Control group clicks = 28378
Total Experiment group clicks= 28325
Total clicks= 56703
Probability of cookie in control or experiment group= 0.5
SE = sqrt(0.5*(1-0.5)*(1/(28378+28325)) = 0.0021
Margin of error (me) = 1.96 *SE = 0.0041
Confidence Interval = (0.5- 0.0041, 0.5+ 0.0041)] = **(0.4959, 0.5041)**
Observed fraction (p(hut)) = 28378/56703 = **0.5004**

Since the observed fraction of successes or cookies fall within the confidence interval, number of clicks has **passed sanity check**.

Click- through- probability

Click-through-probability is number of clicks on "Start free trial" divided by the number of cookies. Since both invariant metrics, number of cookies and number of clicks, passed the sanity check, click-through probability will surely pass the Sanity check. For confirmation purposes, the results are given below.

Total Experiment group pageview= 344660
Control value probability= 28378/345543 = 0.0821258
SE = sqrt(0.0821258*(1-0.0821258)/ 344660) = 0.000468
Margin of error (me) = 1.96 *SE = 0.00092
Confidence Interval = (0.0821258- 0.00092, 0.0821258+ 0.00092)] = **(0.0812, 0.0830)**
Experiment value observed = 28325/344660 = **0.0821**

Since experiment value observed fall within the confidence interval, Click- through- probability has **passed sanity check**.

Result Analysis

Effect Size Tests

Gross Conversion

Number of users who enrolled in the free trial divided by Number of users who clicked the Start Free Trial button

Here, we consider only the number of observations after the 14 days of trial period.

Control group:
Number of clicks (n_{ctr}) =17293
Enrollment (E_{ctr}) =3785

Experiment group:
Number of clicks (n_{exp}) = 17260
Enrollment (E_{exp}) =3423

P (hut) = Eexp + Ectr / (nctr + nexp) = (3423 + 3785)/ (17293 + 17260) = 0.2086

SE = sqrt(P (hut) * (1- P (hut)) * (1/(nctr + 1/nexp)))
= sqrt(0.2086*(1-0.2086)*(1/17293+1/17260)= 0.004372

Margin of error (me) = 1.96* SE = 1.96 * 0.004372 = 0.00857

d (hut)= Eexp / nexp - Ectr / nctr = 3423/ 17260 – 3785/ 17293 = −0.02055

Confidence interval = (d (hut)− me, d (hut) + me)= (−0.02055 -0.00857, −0.02055 +0.00857)
= (−0.0291, −0.0120)

Gross conversion is **statistically significant** because the confidence interval does not include 0, and it is **practically significant** because the confidence interval does not include the practical significance boundary (d_{min} = 0.01).

Net Conversion

Number of user-ids remained enrolled for 14 days trial and at least make their first payment divided by Number of users clicked the Start Free Trial button

Control group:
Number of clicks (n_{ctr}) =17293
Payments (P_{ctr}) =2033

Experiment group:
Number of clicks (n_{exp}) = 17260
Payments (P_{exp}) =1945

P (hut) =Pexp + Pctr / (nctr + nexp) = (1945 + 2033)/ (17293 + 17260) = 0.1151

SE = sqrt(P (hut) * (1- P (hut)) * (1/(nctr + 1/nexp)))
= sqrt(0.1151*(1-0.1151)*(1/17293+1/17260)= 0.003434

Margin of error (me) = 1.96* SE = 1.96 * 0.003434= 0.00673

d (hut)= Pexp / nexp - Pctr / nctr = 1945/ 17260 – 2033/ 17293 = −0.00487

Confidence interval = (d (hut)− me, d (hut) + me)= (−0.00487-0.00673, −0.00487+0.00673)
= (−0.0116, 0.0019)

Net conversion is **neither statistically nor practically significant**, because the confidence interval include 0 and it touches lower boundary (d_{min}= 0.0075).

Sign Tests

Sign Test Calculator (ttp://graphpad.com/quickcalcs/binomial2/)

Gross Conversion

Number of "successes": 4
Number of trials (or subjects) per experiment: 23
Sign test. If the probability of "success" in each trial or subject is 0.500, then:
The two-tail **P value is 0.0026**

Since p-value (0.0026) is less than alpha level (0.025), the change is statistical significant.

Net Conversion

Number of "successes": 10
Number of trials (or subjects) per experiment: 23
Sign test. If the probability of "success" in each trial or subject is 0.500, then:
The two-tail **P value is 0.6776**

Since p-value (0.6776) is greater than alpha level (0.025), the change is not statisticaly significant.

Summary

In our case, the null hypothesis is that there is no difference in the evaluation metrics between control and experiment groups. If we are going to launch the experiment, null hypothesis must be rejected for both evaluation metrics (gross conversion AND net conversion) and criteria for practical significance must be met.

The Bonferonni correction is a method for controlling type I errors (false positives) when using multiple metrics in which relevance of ANY of the metrics matches the hypothesis. In this case the risk of type I errors increases as the number of metrics increases.

In our case in which ALL metrics must be relevant in order to launch, the risk of type II errors (false negatives) increases as the number of metrics increases. In other words, our acceptance criteria require statistically significant differences for ALL evaluation metrics (gross AND net conversion). Therefore, the use of the Bonferonni correction is not appropriate.

Recommendation

From the analysis of Gross conversion it is clear that the free trial screener managed to reduce the number of frustrated students who enrolled in the free trial. Unfortunately, results of Net conversion not only show a decrease in the number of students making a payment, but also affected revenue negatively since lower d_{min} is included in the confidence interval. Therefore, I wouldn’t suggest launching the free trial screener.

Follow-Up Experiment: How to Reduce Early Cancellations

We know that the ultimate goal of Udacity is to decrease early cancellation (number of frustrated students) but to increase conversion rate (the number of students who make at least one payment). I would suggest if Udacity gave the chance of coaching or one to one meeting to all the users who clicked the start free trial button, net conversion would show an increase. The set up being the same, assuming that users who clicked the free trial button are aware of the minimum time requirement, couching at this stage might help to guide and convince users to make them proceed with their enrollment and at least make first payment.

Null Hypothesis: coach session arrangement will not increase net conversion

Unit of diversion: The unit of diversion will be user-id as the change takes place after a student creates an account and enrolls in a course.

Invariant metrics: user-id. It will be considered before the experiment and hence equally distributed among control and experiment groups
Evaluation metrics: The evaluation metric will be Retention.

A statistically and practically significant increase in Retention would indicate that the change is successful, therefore experiment can be launched.

References:

- http://graphpad.com/quickcalcs/binomial2/
- http://www.evanmiller.org/ab-testing/sample-size.html
- http://napitupulu-jon.appspot.com/categories/ab-testing.html