P7: An A/B Test on Free Trial Screener for Udacity

Experiment Design

Metric Choice

- **# of cookies:** the number of unique visitors to the page is not expected to change as a result of this experiment as the screener pops up after a user clicks the start button. So it's an appropriate invariant metric and not a good evaluation metric.
- **# of user-ids:** although the number of user-ids can provide some information about enrollments, it's not a reliable metric. For example, if the number of user-ids in experiment is different from the number of user-ids in the control group, we would not be able to know how much of the difference is due to difference in the number of cookies. Gross conversion is a better choice because it normalizes the number of user-ids based on the number of cookies for the corresponding group
- **# of clicks:** the number of clicks on the start button is not expected to change as a result of this experiment as the screener pops up after a user clicks the start button. So it's an appropriate invariant metric and not a good evaluation metric.
- **Click-through-probability:** click-through-probability does not change because we're experimenting a change that happens after the user clicks on enroll. So it's an appropriate invariant metric and not a good evaluation metric.
- **Gross conversion:** this metric adequately captures the effect of the screener as it's defined as enroll/click. It's expected to change due to the experiment.
- **Net conversion:** this metric adequately captures the effect of the screener as it's defined as pay/click. It's expected to change due to the experiment.
- **Retention** potentially a good evaluation metric but it was later excluded due to impractical sample size needed

I chose number of cookies and the number of clicks; They shouldn't change as they are related to steps *before* the free trial screener pops up. Also, click-through-probability does not change because we're experimenting a change that happens after the user clicks on enroll.

In the result of the experiment, I'll be looking at evaluation metrics to ensure they behave as expected. Specifically, I expect to see a significant decrease in gross conversion since we would like to reduce the number of frustrated students who leave the free trial. Additionally, I expect to see no significant decrease in net conversion since the hypothesis also mentions no significant reduction in the number of students who continue past the free trial. Both must be met in order for us to launch the experiment.

Measuring Standard Deviation

CM: In choosing whether to use an analytical or empirical estimate for the evaluation metrics we should check the denominators of our metrics (or the unit of analysis) and the unit of diversion. If they are the same we can use the analytical estimate.

In order to calculate the standard deviation, we first need to get the N (sample size) for each metric. There are 5000 cookies so we need to translate that into the unit of diversion associated with each metric:

- **Gross Conversion**: The denominator for this metric is click. We first need to find the number of clicks out of 5000 cookies:

Ng = (3200/40000)*5000 = 400 ⇒ SEg = sqrt(.20625(1-.20625)/400) ⇒ **SEg = .0202**

- **Retention** (later removed from evaluation metrics list, due to high pageviews needed)

The denominator for this metric is enroll, we first need to find the number of enrollments out of 5000 cookies:

Nr =(660/40000)*5000 = 82.5 ⇒ SEr = sqrt(.53(1-.53)/82.5) ⇒ **SEr = .0549**

- **Net Conversion**: same process as used for gross conversion:

Nn = (3200/40000) * 5000 = 400 ⇒ SEn = sqrt(.1093125(1-.1093125)/400) ⇒ **SEn = .0156**

In order to see if the analytic estimate is comparable to empirical variability, we need to compare the unit of diversion of the unit of analysis for each metric. The unit of diversion is cookies. The unit of analysis for gross conversion and net conversion is cookies. So, analytic variability tends to match empirical variability. But for retention, the unit of analysis is user-ids. Therefore, analytic and empirical variability don't match.

~~assumption on each metric having a binomial distribution is valid. The criteria for that is:~~
~~N*p > 5~~
~~Where N is the sample size and p is the estimated probability.~~
~~Ng*Pg = 82.5~~
~~Nr*Pr = 43.725~~
~~Nn*Pn = 43.725~~

Sizing

Number of Samples vs. Power

we would launch only if all evaluation metrics show significance (ie significant decrease in gross conversion AND no decrease in net conversion). Therefore, there would be no need to use Bonferroni correction.

I used the online calculator (http://www.evanmiller.org/ab-testing/sample-size.html) with the following information:
Alpha = 5 %
Beta = 20 %

- Gross Conversion:

Sample size: 25, 835
Translate into pageviews: 25835*40000/3200 = 322937.5

- Retention:

Sample size:39,087
Translate into pageviews: 39087*40000/660 = 2368909.091
Note: Too many!! Exclude from evaluation metrics!

- Net Conversion:

Sample size: 27, 413
Translate into pageviews: 27413 * 40000/3200 = 342662.5
Note: This is greater than Gross conversion sample size, thus it'll give the minimum number of pageview we'll need in order to consider both metrics
⇒ Total number of pageviews (both control and experiment groups) = 342662.5 * 2 = **685,325**

Duration vs. Exposure

There has to be a balance between duration and exposure. We don't want to divert all the traffic and risk it in favor of a shorter duration. On the other hand, we do not want the experiment to take forever.
There is not a high risk associated with this experiment (ie it would not impose any physical, psychological and emotional, social, and economic harm to our participants). I don't think we are exceeding the minimal risk as defined through the course (" probability and magnitude of harm that a participant would encounter in normal daily life."). Additionally, running the experiment for too long would prevent us from being able to run other experiments as well as it would be more costly. So I'd propose a 100% exposure.
Duration = 685325/(40000* 0.5) **~ 18 days (meets the "few weeks" requirement)**
Exposure = **100%**

Experiment Analysis

Sanity Checks

Number of cookies:
Each cookie is assigned to either control or experiment groups, so the estimated probability is .5 (binomial). Since totals are high, we can assume a normal distribution.
Create a 95% CI.

SD = sqrt(0.5*0.5/(345543+344660)) = 6.0184e-4
m = SD*Z* = 6.0184e-4 * 1.96 = 1.1796e-3
CI: (0.4988,0.5012)
Observed probability: 345543/(345543+34660) = **0.5006** falls within CI
⇒ **Passed!**

Number of clicks on "Free Trial" button
Assumptions above.

SD = sqrt(0.5*0.5/(28378+28325)) = 2.0997e-3
m = SD*Z* = 2.0997e-3 * 1.96 = 4.1155e-3
CI: (0.4959,0.5041)
Observed probability: 28378/(28378+28325) = **0.5005** falls within CI
⇒ Passed!

Click Through Probability on "Start Free Trial"
Ncont = 345543 Xcont = 28378
Nexp = 344660 Xexp = 28325
Ppooled = (28378+28325)/(345543+344660) = .0821
SEpooled = sqrt(.0821*(1-.0821)*((1/345543)+(1/344660))) = 0.00066
m = +/- 1.96*0.0007 = **+/- 0.0013**
d = (28325/344660)-(28378/345543) = **0.00006**

Result Analysis

Effect Size Tests

As mentioned above, I do not intend to use Bonferroni correction.

- Gross Conversion:

Pcont = 3785/17293 = 0.2189;
Pexp = 3423/17260 = 0.1983 ⇒ dhat = -0.0206
Ppooled = 0.2086
SEpooled = 0.0044 ⇒ m = SE*z*= +/-1.96*0.0044 = +/- 0.0086
⇒ **CI = (-0.0291, -0.012);**
dmin = 0.01 ⇒ **both practically and statistically significant**

- Net Conversion:

Pcont = 1945/17260 = 0.1127;
Pexp = 2033/17293 = 0.1176 ⇒ dhat = -0.0049
Ppooled = 0.1151
SEpooled = 0.0034⇒ m = SE*z*= +/-1.96*0.0034= +/- 0.0067
⇒ **CI = (-0.0116, 0.0018)**;
dmin = 0.0075 ⇒ **neither practically nor statistically significant**

Sign Tests

Gross Conversion: 19 out of 23 negatives (success)
Net conversion: 13 out of 23 positives (success)
Using QuickCalc's binomial calculator (p = 0.5):

- Gross conversion: pvalue = **0.0026** ⇒ **statistically significant (desired ✓)**
- Net conversion: pvalue = **0.6776** ⇒ **not statistically significant (desired ✓)**

Summary

we would launch only if we see a significant decrease in Gross Conversion AND no significant decrease in Net Conversion. We are conservative enough by saying both conditions must hold true. Therefore, there would be no need to use Bonferroni correction. (1)

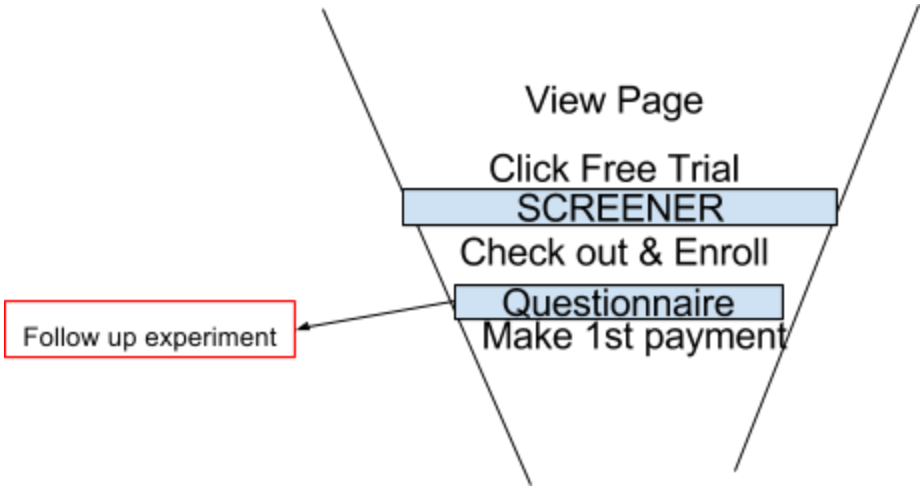
Recommendation

Based on the result of the analysis, I would recommend investigating more before launching the experiment. As shown above, Gross Conversion has shown both practical and statistical significance (thus, behaving as expected in order to launch). In addition, there is neither statistically nor practically significant change in Net Conversion. However, net conversion's CI has an overlap with the negative side of of the practical significance boundary. Based on the CI there's a chance that net conversion decreases to value lower than specified by the business requirement (-dmin). Therefore, I'd recommend further testing and investigation before launching this experiment.

Follow-Up Experiment

I propose to add a refund incentive where students get 5% of the money back if they complete the course. The hypothesis is that this incentive reduces the number of students who cancel early. An appropriate evaluation metric could be retention, which will show the probability of making the first payment given enroll. I expect this metric to show an increase as a result of this experiment. I recommend using user-id as the unit of diversion, as it's more convenient and more stable than a cookie. It also captures all the information we need as our experiment happens after enroll (ie, when users are assigned ids). My choice of invariant metric is number of user-ids as it is expected to not change with the experiment.

~~I would suggest we add a questionnaire questionnaire that pops up if a student intends to leave the free trial. The student would be asked a multiple choice question about why s/he is leaving. One choice would be "this course was too time consuming." We expect the rate at which users select this choice (let's call it frustration rate!) go down as a result of the screener.~~



~~The unit of diversion would be user-id as the questionnaire shows up while they are on free trial (and therefore they have user-ids.) An appropriate evaluation metric would be the probability of becoming frustrated (!), given click. My hypothesis is that if the free trial screener proposed in this experiment work well, we should be able to see a significant decrease in this metric.~~

(1) A really useful comment from Udacity review on Bonferroni, for future references:

“The answer is correct, we need both metrics to meet our expectations in order to launch, ideally you would explain why that means that we don’t need the Bonferroni correction, you could find an interesting discussion here: https://discussions.udacity.com/t/bonferroni-correction/201344/14

Some hints: When we are considering multiple metrics at the same time, and we need all of them to inform our decision (in our case, we are looking at both gross and net conversion to decide whether to launch or not), the risk is of a type II error. That is not what the Bonferroni correction is designed for. The Bonferroni correction is designed to reduce the risk that one metric is deemed significant by mistake. If we were in the situation where we need just one metric to meet expectations in order to launch an experiment, then we would need Bonferroni. In our case we would need multiple metrics to match our expectations to launch the experiment therefore Bonferroni is neither necessary nor helpful.”