Udacity AB Testing / Data Analysis ND

Jean Paul Barddal, M.Sc.

December 5, 2016

1 Introduction

In this project I will perform an AB Test for an experiment conducted by Udacity to verify if a change in their enrolling process presented in the website would lead to less frustrated students, that leave the course because they don't have enough time to dedicate.

Udacity has two options on the home page: "start free trial" and "access course materials". If students click on "start free trial", they will be asked to enter credit card information, and will be automatically enrolled in a free trial of the paid version of the course for 14 days. After this timespan, they will be charged unless they cancel their enrollment first. Conversely, if students click on "access course materials", they will be able to (i) watch the lectures, and (ii) take the quizzes for free, but they will neither receive coaching support, a certificate, nor submit their final project for feedback.

In this experiment, Udacity tested a change where if the student clicked "start free trial", they would be asked how much time they had available to devote to the course. If they indicated that they would be able to dedicate 5 or more hours per week, then they would be taken to the checkout process as usual. On the other hand, i.e. less than 5 hours per week would be dedicated, then a message would pop up indicating the Udacity courses usually require a greater time of commitment for successful completion, and the students would be suggested to access the course materials for free. In this case, the student could either enroll as usual, or access the course materials for free instead.

The hypothesis to be tested is that if we set clearer expectations for students upfront, they would be less frustrated since less students would leave the free trial for not having enough time – also, it would be important to verify if we were not significantly reducing the number of students that

continue past the trial and eventually complete the course. If this hypothesis holds, Udacity could then improve the overall student experiment and improve coaches' capacity to support students who are likely to complete the course.

This report is divided as follows. Section 2 introduces the required statistics necessary for the proposed AB test. Section 3 analyzes the results of the statistics and reports the outcome of the AB test. Finally, Section 4 proposes a follow-up experiment to reduce early cancellations.

2 Basic Statistics

Udacity has provided the required data in here[1] and here[2].

Given all of the possible metrics, we need to pick both variant and invariant metrics. Variant metrics will determine if our hypothesis hold, while invariant ones are not expected to significantly change across the experiment and control groups and shall be used to check the integrity of our test.

This choice is somewhat naive since the number of cookies (correlated to the number of page views), clicks and click-through-probability should remain unchanged regardless of the experiment set we're dealing with. All of the remainder can change, but I will discuss the ones I believe are relevant for our test in the following topics.

- **Gross conversion rate**: this metric could determine whether the screener has effect on enrollments,
- **Net conversion rate**: this metric could be used to measure if the screener proved to change the completion rate in the 14-day span,
- **Retention rate**: again, could be used to check if the screener had any effects on the 14-day dropout rate.

If the hypothesis depicted in the introduction is correct, we would expect to see changes in all these three metrics.

- We would expect a **DECREASE** in the *grose conversion rate* since dropping students would be 'filtered' by the screener

[1] `https://docs.google.com/spreadsheets/d/1MYNUtC47Pg8hdoCjOXaHqF-thheGpUshrFA21BAJnNc/edit#gid=0`

[2] `https://docs.google.com/spreadsheets/d/1Mu5u9GrybDdska-ljPXyBjTpdZIUev_6i7t4LRDfXM8/edit#gid=154400404`

- We need the *net conversion rate* to remain **UNCHANGED** as the number of students to continue past the free trial and complete the course should **NOT** be affected
- Finally, the *retention rate* should be **HIGHER** as the students likely to drop would not have enrolled, and those who enrolled would be unlikely to drop

2.1 Variability

Before conducting the experiment, Udacity gathered data to get daily cookies, enrollments, click through probability, gross conversion, net conversion and retention. I will refer to these data as the 'baseline'.

First, I will need to make a few assumptions. The first is that we will need a fair amount of cookies per day from each group. Since we do have a lot of cookies available, let's say we need 10,000 cookies to work on, 5,000 of each group. The second is that the distribution of the data is gaussian, which will allow me to compute the standard deviation using a simple approximation. First, we need to scale the fraction of pageviews in the sample over the pageviews in the baseline, which is:

$$\frac{5000}{40000} = 0.125$$

So, given the 3,200 clicks and 660 enrollments in the baseline data, we can predict 400 clicks and 82.5 enrollments a day in the sample.

The rates of the evaluation metrics, i.e. gross rate (gc), retention (r) and net conversion (nc), which are as follows:

$p_{gc} = 0.20625 \quad p_{nc} = 0.10931 \quad p_r = 0.53$

Now, we need to remember that all of these metrics follow a binomial distribution! And thus, se now can compute the standard deviation for all the metrics as follows:

$$\sigma = \sqrt{\frac{p(1-p)}{n}} \tag{1}$$

$\sigma_{gc} = 0.0202 \quad \sigma_{nc} = 0.0156 \quad \sigma_r = 0.0549$

In the post experiment data, the number of cookies per day was higher than our estimation. There were nearly 10,000 a day rather then 5,000. Doubling

the sample size to 800 and enrolls to 165, we can recompute the analytic estimate for the standard deviation, which will be:

$\sigma_{gc} = 0.0143 \quad \sigma_{nc} = 0.0110 \quad \sigma_r = 0.0388$

Of the three metrics, the analytical standard deviation computation of retention is unlikely to match the one seen in the experiment. This is due the fact that the unit analysis for retention is 'user-id', while the unit of diversion for the experiment is cookies (remember, user-id is only associated to users that enrolled!). Conversely, the gross and net conversion rates are expected to match since the empirical standard deviation are the same as in the experiment.

2.2 Sizing

Following the project description, we need to know the number of page views required to conduct our analysis. The project requires we adhere to a type I error rate of $\alpha = 0.05$ and type II error rate of $\beta = 0.20$. For the selected metrics, the minimum detectable effect (d_{min}) has been pre-specific as a business decision:

$d_{min}(gc) = 0.01 \quad d_{min}(nc) = 0.0075 \quad d_{min}(r) = 0.01$

We now can use a sample size formula to determine the required number of samples for each metric n_{min}. Since the idea here is to evaluate all metrics, the final sample size should be the one the maximizes.

Each metric has its own unit (clicks or enrolls), so values need to be rescaled given the ratio seen in the baseline.

- Ratio of page views to clicks = 0.08
- Ratio of page views to enrolls = 0.0165

$n_{min}(gc) = 2 \times \frac{25835}{0.08} = 645,875, \quad n_{min}(nc) = 2 \times \frac{27413}{0.08} = 685,325$, and

$n_{min}(r) = 2 \times \frac{39115}{0.0165} = 4,741,213$

The biggest sample is our limiting factor (also referred as retention rate), so we would need **4,741,213 page views** to conduct the analysis.

2.3 Duration vs. Exposure

Given the required page views, the exposure can be specified by determining the risk of the experiment, and then we can determine the experiment duration. The exposure depends on the risk involved and because the screener is a warning sign about the time required to complete the course, it constitute

near-zero risk. Since no students would suffer any physical risk during the experiment, and since no sensitive data is collected, a complete exposure (100%) is a safe choice.

Dividing the total number of page views by the page views per day in the baseline (40,000), we obtain a duration of **119 days** were Udacity would have to divert it's entire traffic. That's almost **4 months(!)**, so I believe the duration of the experiment should be reduced. A possibility would be excluding the retention as a metric and consider the next limiting metric, i.e. net conversion. This way, we would need 685,275 page views, and thus, we would need to run the experiment for **18 days**.

3 Analysis

3.1 Sanity Checks

Now, we need to go back to the invariant metrics to see if our assumptions are met. We expect that cookies and clicks are evenly split between the control and experiment subsets. So, using a expected rate diversion of 50%, we can compute the standard deviation (again, assuming a gaussian distribution), and construct a 95% confidence interval (CI) around our expectation (E). By comparing the observed rate, we can check if these two invariant metrics are reliable.

Working with $p = 50\% = 0.5$, $\alpha = 0.05$ and $Z = 1.96$ (from the Gaussian statistics table), we have:

- Cookies
  $\sigma_{cookies} = \sqrt{\frac{0.5(1-0.5)}{345543+344660}} = 0.006$
  $E_{cookies} = Z \times \sigma_{cookies} = 0.01176$
  $CI_{cookies} = p \pm E_{cookies} = [0.48824, 0.51176]$

- Clicks
  $\sigma_{clicks} = \sqrt{\frac{0.5(1-0.5)}{28378+28325}} = 0.002$
  $E_{clicks} = Z \times \sigma_{clicks} = 0.00392$
  $CI_{clicks} = p \pm E_{clicks} = [0.49608, 0.50392]$

Fortunately, both cookies and clicks pass the sanity check. For the click-thru-rate (CTR in the formulas below), we should observe more or less the same value across the two groups. We can follow the same rationale here, but instead, we should compare the observed rates in the control and experiment groups. This test will determine if the two rates come from the

same distribution (which is the hypothesis earlier stated, i.e. it is really invariant).

$\frac{28378}{345543} = 0.082$

$\sigma_{CTR} = \sqrt{\frac{0.0821(1-0.0821)}{345543}} = 0.000467$

$E_{CTR} = Z \times \sigma_{CTR} = 0.00091532$

$CI_{CTR} = p \pm E_{CTR} = [0.0811, 0.0830]$

Since 0.082 is inside CI_{CTR}, it also passes the sanity check!

3.2 Significance

For each metric, I will now test for statistical and practical significance. The minimum detectable effect is the smallest difference that we need to obtain between the experimental and control groups for us to assume that they are practically significant. For each metric, I will compute the rate in each group and then compute their differences. These new difference variables will be used to construct confidence intervals as I did in the sanity check. No Bonferroni correction is needed here since the final outcome will require that each of the evaluation metrics present significant results. In practice, if a single metric fails, the initial hypothesis will be denied.

General Assumption: $\alpha = 0.05$ **and** $Z = 1.96$ (Reminder: we can't work with retention since we do not have enough data!)

- **Gross Conversion**

  $r_A = 0.2188$

  $r_B = 0.1983$

  $\hat{d} = -0.0205$

  $\sigma_A^2 = \frac{0.2188(1-0.2188)}{17293} = 9.88 \times 10^{-6}$

  $\sigma_B^2 = \frac{0.1983(1-0.1983)}{17260} = 9.21 \times 10^{-6}$

  $\sigma_{\hat{d}}^2 = 9.88 \times 10^{-6} + 9.21 \times 10^{-6} = 1.9098 \times 10^{-5}$

  $\sigma_{\hat{d}} = 0.004$

  $E_{\hat{d}} = 8.5652 \times 10^{-3}$

  $CI = [-0.0291, -0.0120]$

  $d_{min} = -0.01$

- **Net Conversion**

$r_A = 0.1176$
$r_B = 0.1127$
$\hat{d} = -0.0049$

$\sigma_A^2 = \frac{0.1176(1-0.1176)}{17293} = 5.9990 \times 10^{-6}$
$\sigma_B^2 = \frac{0.1127(1-0.1127)}{17260} = 5.7931 \times 10^{-6}$
$\sigma_{\hat{d}}^2 = 5.9990 \times 10^{-6} + 5.7931 \times 10^{-6} = 1.1792 \times 10^{-5}$

$\sigma_{\hat{d}} = 3.4340 \times 10^{-3}$
$E_{\hat{d}} = 6.7228 \times 10^{-3}$
$CI = [-0.0116, 0.0018]$
$d_{min} = -0.0075$

Gross conversion is then statistically and practically significant. Net conversion is not statistically significant, but the negative lower bound of the detectable effect is within the range of the confidence interval. Our goal in practical significance is to check that we do not have changes in net conversion rates. We the the lower bound of the confidence interval to be smaller than our minimum detectable effect which should be excluded from the confidence interval, and thus, net conversion is not practically significant.

3.3 Sign Tests

Another possibility would be testing each of the metrics individually via a binomial sign test. In this case, each day of the experiment is used to compare to see if there is a positive (+1) or negative (-1) difference across the groups. We could denote a positive difference as a success and a negativa as a failure. Then, a comparison between the p-values would lead us to determine significance. For instance, the gross conversion rate has only 4 out of 23 successes for a two-tailed p-value of 0.0026. This is way smaller the our individual type I error of 0.025, and thus, it indicates statistical significance of this metric. On the other hand, net conversion has 10 out of 23 successes and a two-tailed p-value of 0.6776, showing that this metric is not statistically significant.

3.4 Overview of the Results

The effect size tests showed that gross conversion is both statistically and practically significant, while net conversion is neither. The gross conversion

rate dropped approx. 2%, and thus, the screener was effective at reducing the number of students that enrolled from the initial click. Yet, the net conversion dropped approx. 0.5%, indicating that the screener had a negative effect on the number of students that completed the trial. This is a drawback of the screener: it makes students to evade the free trial since they don't think they have enough time to devote 5 hours/week to the course. Although the decrease of the gross conversion supports our hypothesis, we can't follow this screener technique since the net conversion dropped and invalidates the initial hypothesis.

3.5 Recommendation

Following the previous discussion, the screener was effective at reducing the number of people from clicking to enroll, but it failed w.r.t in maintaining the number of students who continued the course after the trial. In practice, we had a decrease in the ratio of students who kept on working on the nano degrees. Therefore, we should **NOT** proceed with the launch to the Udacitys website.

4 Follow-up Experiment

My concern with the current proposal is that students are somewhat 'scared' and lead to think they don't have enough time to dedicate compared to what they actually have. In opposition to presenting the screener upfront, maybe Udacity could measure the number of hours dedicated by the student during the free trial, and after a few days (maybe after the first 7 days), the screener could pop up if the student hasn't dedicated enough time to either motivate the student to dedicate more time or drop without being charged. On the other hand, if the student is showing a good progress and is dedicating enough time, he would be just continue the trial, and consequently, the course. Evidently, Udacity doesn't want any students to drop, so tips, advices and job offers for people with the skills retained with this nano degree could be shown, all aiming at inspiring the student to keep on working on the Nano Degree. In this case, the AB test would be similar: we could compare the **same metrics** for the subset of students that use the current version of the website against the ones that get the screener after a few days using the free trial.