A/B Testing Analysis and Design

By: Sohrab Rahimi

Introduction:

In this project, Udacity (a company which focuses on online education), likes to run an experiment and see whether or not adding a message after a clicking on the “start free trial” button in their website would have significant effects or not. The idea is to ask the student about the amount of free time he/she has to spend on the course. If they answer less than 5 hours per week, then a warning message shows up advising him/her that the course will require more time and they will not be successful with less than 5 hours. With this in mind, the student will still have an option to go ahead with the free trial or not. The hypothesis of this idea is that the student will be enlightened about what he/she might expect and hence, avoid frustration without significantly reducing the course enrollment. If this hypothesis is true, Udacity will manage to improve the experience of students throughout the class. In the following steps, we are going to test this hypothesis through A/B testing. The following table is what Udacity has collected for this experiment:

Unique cookies to view page per day:	40000
Unique cookies to click "Start free trial" per day:	3200
Enrollments per day:	660
Click-through-probability on "Start free trial":	0.08
Probability of enrolling, given click:	0.20625
Probability of payment, given enroll:	0.53
Probability of payment, given click	0.1093125

1. Experiment Design:

1.1. Metric Choice:

In the first step, we need to choose our metrics. In doing so, one should note that the invariant should be those metrics chosen before the treatment (the appearance of the message in this case) so that we rest assured that they will be the same in both experiment and control groups (i.e. invariant metrics should be completely independent from the treatment). Whereas, evaluation metrics are those we expect to be different in two groups and we want to measure and draw our conclusion on. Therefore, two sets of metrics are used for this experiment:

1. **Invariant Metrics:**

Invariant metrics refers to those metrics that will be the same in both control and experiment groups. On this account, these metrics help us to have control over the way we divide our sample into experiment and control groups. For this study, therefore, we use three invariant metrics:

1. The number of cookies to view the course overview page (it’s a good invariant matrix because we expect this to be approximately the same in both experiment and control groups)

2. The number of cookies to click “Start free trial” (since the message is supposed to show up after clicking this button, we expect this metric to be the same in both experiment and control groups).
3. Click-through-probability on “Start free trial” (for the same reason as number 2)

2. Evaluation Metrics:

Unlike invariant metrics, these metrics vary between experiment and control groups. Evaluation metrics, as the naming suggests, will be our criteria for deciding on whether the change has been effective or not. The data that Udacity has collected provides three useful evaluation metrics: Retention, Gross Conversion, and Net Conversion. Among the other metrics, the user-id seems to be useful at first but since it is not normalized, it is not reliable and therefore we will use Gross Conversion instead.

Retention: 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.

Gross Conversion: Number of user-ids to complete checkout and enroll in the free trail divided by number of unique cookies to click the "Start free trail" button.

Net Conversion: 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 trail" button.

What should be expecting from these metrics? We should expect that the Gross Conversion decrease since it represents the number of people who sign in for the free trial and we expect this number decrease because the message may discourage some students who are not willing to spend enough time on this course. On the other hand, we expect both Retention and Net Conversion to see no decrease, since the desired output is that putting up the warning message will not affect the number of students who proceed to the first payment.

1.2. Measuring Standard Deviation

In this section we calculate the standard deviation for each evaluation metric that we chose from the previous step. We assume a Bernoulli distribution for all probabilities and therefore the standard deviation can be calculated as follows:

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

Where P is the metric’s probability and N is the population. Recall that our values have been calculated for 3200 clicks. Therefore, we will need to first calculate the standard deviation for 3200 clicks and then estimate that for 50000 clicks. Below are the standard deviations for 3200 clicks:

Sd (Gross Conversion) = $\sqrt{\frac{\mathbf{0.20625(1-0.20625)}}{\mathbf{3200}}}$ **= 0.00715**

Sd (Retention) = $\sqrt{\frac{0.53(1-0.53)}{660}}$ = **0.01942**

Sd (Net Conversion) = $\sqrt{\frac{0.1093125(1-0.1093125)}{3200}}$ = **0.0055159**

In order to calculate the standard deviation for 50000 sample size, we should multiply the ones we calculated by $\sqrt{\frac{40000}{5000}} = 2.8282$. After doing this, the following standard deviations will result:

Sd (Gross Conversion) = 0.0202
Sd (Retention) = 0.0549
Sd (Net Conversion) = 0.0156

Analytic estimate will be comparable to the empirical variability if the unit of diversion and unit of analysis are equal. In case of Gross Conversion and Net Conversion, the unit of analysis and conversion are the same (i.e. number of cookies). In case of retention the two are not the same, as the denominator for the estimate is "Number of users enrolled the courseware" which is not the same as the unit of diversion. Accordingly, the analytical and empirical estimates are different.

1.3. Sizing

1. Number of Samples vs. Power

As the metrics chosen are probably highly correlated, using Bonferroni correction does not make sense as it is too conservative. To calculate the sample sizes I used the online calculator suggested in the course [1]:

Measure	Gross Conversion	retention	Net Conversion
Baseline Conversion	0.20625	0.53	0.1093125
Minimum Detectable Effect	0.01	0.01	0.0075
alpha	0.05	0.05	0.05
beta	0.2	0.2	0.2
Sample size per variation	25835	39115	27413
Total Sample Size	51670	78213	54826
Total Page Views	642475	4739879	679300

While retention rate would have been a great measure, it requires 4739879 total page views which is substantially higher than other two evaluation metrics. If we use retention, with 40,000 daily page views traffic, the experiment would take at least 119 days which is not reasonable. Therefore, using the other two measures make more sense. Based on the other two metrics, we will need at least 679300 page views to be able to accomplish this experiment with both metrics.

2. Duration vs. Exposure

According to our data, the daily traffic is 40k page views. We can use the entire traffic as there seems to be no ethical risk associated with this experiment for Udacity neither would it harm the users. We need about 679300 page views. By dividing the two we conclude that we need at least 17 days to collect enough samples for our experiment.

2. Experiment Analysis

2.1. Sanity Checks

Page views in Control : 345543
Page views in Experiment : 344660
Clicks in Control : 28378
Clicks in Experiment : 28325

Where the probability for a cookie being in the control or experiment groups is 0.5. Accordingly:

1. The number of cookies to view the course overview page:

SE = $\sqrt{\mathbf{0.5}(\mathbf{1}-\mathbf{0.5})((\frac{\mathbf{1}}{\mathbf{345543}}+\frac{\mathbf{1}}{\mathbf{344660}})}$ *= 0.0006018*
Margin of Error = SE (1.96) = 0.0011796

Confidence Interval = [0.4988,0.5012]
Observed value = 0.5006

2. The number of cookies to click “Start free trial”:

Using the same formula, we get the following:

Confidence Interval = [0.4959,0.5041]
Observed value = 0.5005

3. Click-through-probability on “Start free trial”:

Confidence Interval = [0.0812, 0.0830]
Observed value = 0.0822

We can see that all the observed values are within the confidence levels and therefore they all pass the sanity check.

2.2. Results Analysis

1. **Effect Size Tests**

In this step we will calculate a confidence level for the difference between the evaluation metrics between the experiment and control groups. We can first calculate this interval for the Gross Convention with the following data:

Measure	Control Group	Experiment Group
Clicks	17293	17260
Enrollment	3785	3423
Gross Conversion	0.218874	0.198319

SE = 0.00437167
Margin of Error = SE(1.96) = 0.008568
Pooled Probability = 0.2086
d^ = -0.02055
CI = [-0.0291, -0.0120]

We can see that Gross Conversion is statistically significant as it does not contain zero and it is also practically significant as it doesn't contain d hat. Now for the Net Conversion:

Measure	Control Group	Experiment Group
Clicks	17293	17260
Enrolment	2033	1945
Net Conversion	0.117562	0.112688

SE = 0.0034341
Margin of Error = SE(1.96) = 0.0067
Pooled Probability = 0.2086
d^ = -0.0049
Confidence Interval = [-0.0116, 0.0018]

Since this confidence level contains zero and d hat, it is neither statistically nor practically significant

2. Sign Test:

For alpha = 0.5, using the online calculator [2] we get the following results:

Metric	P-value	Reject the null hypothesis
Gross Conversion	0.0026	yes
Net Conversion	0.6776	No

3. Summary

In this experiment we wanted to test whether adding a message after clicking, for a specific group of users (depending on the amount of time they are willing to spend on the course) will improve the students' experience of the course without significantly affect the number of students who make the first payment. We chose three invariant metrics: Number of cookies, clicks and click through probability. We also chose a three evaluation metrics: Retention, Gross Conversion and Net Conversion, however since Retention required a lot of page reviews we didn't go ahead with it.

Our null hypothesis was that the experiment group and the control group show no significant difference in Net Conversion and Gross Conversion. We then ran the practical and statistical significance test without using the Bonferroni Correction, as Bonferroni is used to adjust for type I error (false positive) at the expense of power, or increased type II error [3] which is not helpful for our case. In our case, type I error is not as important as the type II is: we want the results to be significant so we don't want to be conservative about rejecting the null hypothesis, at the expense of type II error. What we want to avoid from in our case, is falsely accepting the null hypothesis since both the evaluation metrics should be satisfied for us to launch. After running the sanity checks, we looked into the effect size of each metric and we found out that only the Gross Conversion is influenced by this change and not the Net Conversion.

4. Recommendation

The results of this test is pretty close to what we expected at first. We see a significant decrease in the Gross Conversion metric, meaning that overall, the number of students who enroll in the free trial significantly decrease with that message showing up. While the Net conversion didn't change significantly (neither practically nor statistically), which means that the number of people who make the first payment doesn't change whatsoever. To be on the safe side though, other tests may be helpful. Recall that the confidence interval calculated for the Net conversion was [-0.0116, 0.0018]. Although this interval includes zero and therefore not statistically significant, most of it lies below zero (86.5 % of it). Accordingly, it would be safer if we look into another data set to see if the Net Conversion will still be insignificant as we want to make sure that the number of people who proceed to the first payment will not decrease.

5. Follow up experiment:

Other than the amount of time one spends on each course, another factor might be the person's knowledge domain and background as a person with relevant background might actually need less time to accomplish the course. One experiment then, to create an introductory course with which touches on some of the topics in the program so that the students prepare for the type of knowledge they may need to accomplish these courses. This introductory course could be pretty similar to P0 project which Udacity already offers. So, it would be interesting to see to what extent this course helps students to gain an understanding of what to expect from this course. My hypothesis is that those students who take this introductory course are more likely to care about whether or not they will be able to finish this program or not and therefore, will increase the chance of finishing the program.

To design this experiment, we will define two groups: the control group, are those students who started the free trial but were not exposed to the introductory course at first, and the experiment group will be those who also started the free trial but were first directed to the introductory course as part of their trial. The unit of diversion will be the user-id because we don't want to confuse the users and we clearly want to separate those users exposed to the introductory course from those who weren't. The invariant metrics, will be the number of user-ids since users are exposed to the course after starting the free trial. The evaluation metric will be the graduation rate (i.e. the number of students who finished the program over the population of those who started the fulltime trial) and also, the overall duration it took those students who started the free trial to accomplish the entire program. The null hypothesis, in this case, will be that the introductory course will have no impact on the graduation rate and the duration of the program. If we see practically and statistically significant increase in the graduation rate and decrease in program duration for those students who took the introductory course, we will conclude that this course has had an effect on the program's success.

References:

[1] http://www.evanmiller.org/ab-testing/sample-size.html

[2] http://graphpad.com/quickcalcs/binomial1.cfm

[3] http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full