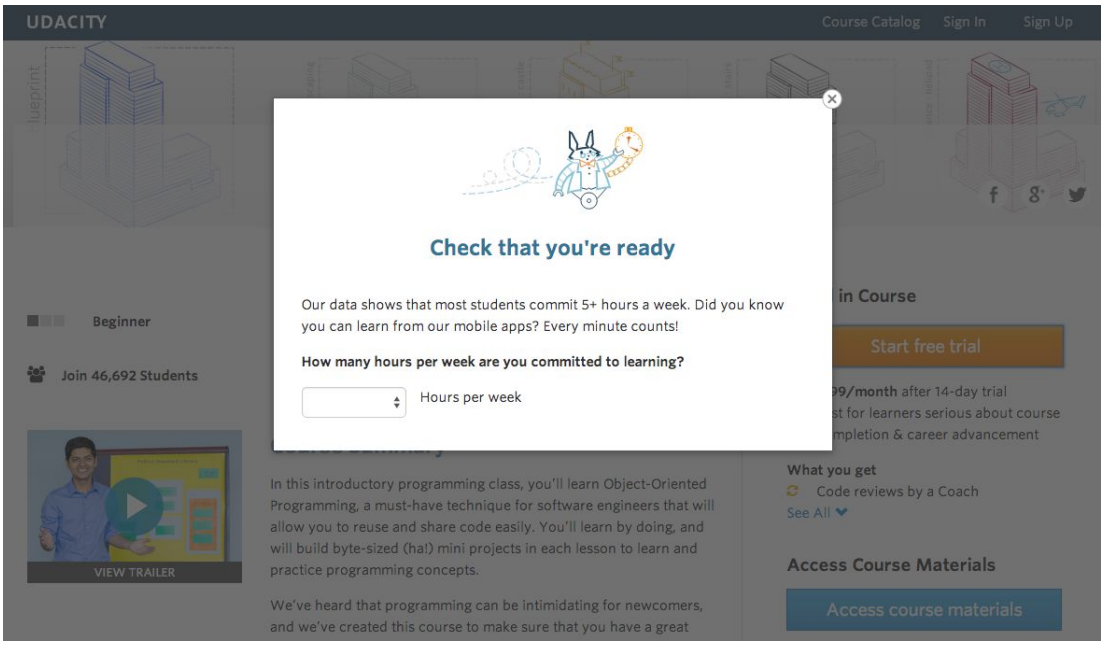
Kyle Shannon
7/23/16
Udacity P7: Design an A/B Test - Version 2

# Project Overview

Udacity has decided to run an A/B test to determine if a lightbox that asks a question and provides feedback will reduce student's frustration, improve coach's interactions with the correct students and provide good upfront feedback when students are signing up for a free trial. The lightbox will be implemented for the experimental group after these users have clicked the "start free trial" button (pictured below). The control group will not see this particular light box and will continue to signup for a free trial as normal.



"start free trial" button

From a business standpoint Udacity hopes this change will reinforce to potential paid customers that these courses involve hard word and on average students spend 5 hours per week completing courses. If this message has been well received we will might see a potential dip in the experimental group's **Gross Conversion**, although not a significant dip, coupled with an increase in the group's **Net Conversion**. Additionally we should hope to see the experimental group has a higher **Retention** rate.

# Experiment Design

## Metric Choice

### Invariant Metrics

Invariant metrics are not suppose to change between a control and experimental group. This makes them particularly useful when sanity checking, for example the distributions of these metrics should remain the same.

1. Number of Cookies:  In this case a cookie is a proxy for a unique user visiting the course page. We want to make sure that we have approx the same number of cookies in both control and experimental groups. In this case a cookie count is unique by day. In other words the same cookie visiting a site twice in a single day would be counted as one.
2. Number of Clicks: This is the number of times the “start free trial” button is clicked per cookie. This count should be the same between both groups as well.
3. Click-through-probability: This metric is the percentage of cookies that actually visit the course page AND click on the “start free trial” button. This metric should remain approx the same between both groups.

**Evaluation Metrics**

Evaluation metrics are the primary mover and shakers of the study. We want to see the difference between these metrics in the control and experimentation group. These metrics will be what we use to determine if there has been a valid change between both groups. In other words if there is a statistically significant finding and the metric has observed a minimum difference (dmin), then there may be cause to launch the change.

1. Gross Conversion($d_{min}$=0.01): User IDs (users must create a unique log in when they sign up) that start the free trial as a ratio to the number of cookies that clicked on the “start free trial” button.
2. Retention($d_{min}$=0.01): Looks at the ratio of user IDs that become paid customers (at least one payment) over the number of user IDs that at least completed checkout.
3. Net Conversion ($d_{min}$=0.0075): These are the number of user IDs that became paying customers (at least one payment) by the number of unique cookies that click the “start free trial” button.

**Unused Metrics**

1. Number of user-ids: This metric is not suitable for either category above, because user IDs might have already completed a trial or clicked on the button and that is how they became users, thus their distribution would be different

**Results Needed to Launch**

To launch this experiment, we want to decrease the number of unprepared students while not decreasing Udacity’s revenue. Decreasing the number of unprepared students will be measured by gross conversion, whereas net conversion and or retention will be the measurement determining change in Udacity’s revenue. If these criteria are met we may have reason to launch the experiment:

- If we see a practically significant decrease in gross conversion
- A statistically significant increase in retention
- and one of the following:
    - A statistically significant increase in net conversion or
    - No statistically significant change in net conversion, we are chiefly concerned with revenue not being harmed.

## Measuring Standard Deviation

In this experiment we are expecting to have 5000 cookies in each group. Below are the calculations used to derive the standard deviations.

| | | caculation | adjusted | |
|---|---|---|---|---|
| Unique cookies to view page per day: | 40000 | 5000/40000 | 0.125 | |
| Unique cookies to click "Start free trial" per day: | 3200 | 3200*0.125 | 400 | |
| Enrollments per day: | 660 | 660*0.125 | 82.5 | |
| Click-through-probability on "Start free trial": | 0.08 | | | |
| **Gross Conversion** - Probability of enrolling, given click: | 0.20625 | sqrt((0.20625*(1-0.20625))/400) | 0.02023060414 | standard deviation |
| **Retention** - Probability of payment, given enroll: | 0.53 | sqrt((0.53*(1-0.53))/82.5) | 0.05494901218 | standard deviation |
| **Net Conversion** - Probability of payment, given click | 0.1093125 | sqrt((0.1093125*(1-0.1093125))/400) | 0.01560154458 | standard deviation |
| | | | | |
| standard deviation = sqrt( (p(1-p)) / n) | | | | |
| 0.0202 | standard deviation | Gross Conversion | | |
| 0.0549 | standard deviation | Retention | | |
| 0.0156 | standard deviation | Net Conversion | | |

Because the user IDs are used as the unit of diversion in Retention, it is possible that the analytic estimate will differ from the empirical variability. I would not expect them to be different for Net Conversion and Gross Conversion, because they use unique cookies as the unit of diversion.

## Sizing

### Number of Samples vs. Power

I will not be using the Bonferroni correction, because the correlation is fairly good and this method would probably be too conservative. Additionally, we simply are not conducting many hypothesis tests at once. Therefore we should not have a runaway type one error rate. E.g. if we had 20 simultaneous tests at an alpha level of 0.05, we would have about a ~64% chance of observing one false positive. In this case using the Bonferroni might be a good choice, though it can be a bit too conservative. Other options might be the False Discovery Rate (FDR) or the positive False Discovery Rate (pFDR)

Calculated pageviews in the same google spreadsheet as above, shown below:

| | | | sample size | number of groups | total sample size | clicks or enrollements per page view | pageviews | days needed |
|---|---|---|---|---|---|---|---|---|
| Gross Conversion - Probability of enrolling, given click: | 0.20625 | dmin: .01 | 25835 | 2 | 51670 | 0.08 | 645875 | 16.146875 |
| Retention - Probability of payment, given enroll: | 0.53 | dmin: .01 | 39115 | 2 | 78230 | 0.0165 | 4741212 | 118.5303 |
| Net Conversion - Probability of payment, given click | 0.1093125 | dmin: .0075 | 27411 | 2 | 54822 | 0.08 | 685275 | 17.131875 |
| | | | | | | | | |
| Type 1 alpha rate: | 0.05 | | | | | | | |
| Type 2 beta rate: | 0.2 | | | | | | | |

I can immediately tell you that we should drop retention as with 100% of traffic diverted to it, it would take about 119 days, this seems like a bad business decision. So I will not be using it going forward.

Because retention is being dropped, we will **require 685,275 pageviews** to power this experiment.

### Duration vs Exposure

I think now that we are not using Retention, we could reasonably **divert 60% of the traffic which would take about 21 days to run this experiment.** That sounds like a logical business decision. Sure the experiment could be run faster in 17 days with 100%, but if there was a bug, or some problem that was not expected or anticipated then we could really mess business ops up in the short term. 60% is fairly more conservative, and taking a almost a month to run an experiment is by no means outlandish.

I would not consider this experiment to be fairly risky. It does not affect how paying students and non-paying students perform coursework, and because it is a simple technical change there is low risk for bugs to diminish the user experience. There may be some small risk to potential new students looking to enroll in the free trial, but I think the gains outweigh the risks overall.

## Experiment Analysis

### Sanity Checks

Before we can continue with our analysis we should perform some sanity checking on our invariant metrics' underlying assumptions. From the pageview/click data we can assume an even .50 split between both control and experimental groups and use these numbers to construct a 95% confidence interval to ensure what we are seeing is within expected ranges. Below are my results for creating the confidence intervals for each invariant metric:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Date | Pageviews | Clicks | Enrollments | Payments | | |
| 37 | Sat, Nov 15 | 8630 | 743 | | | | |
| 38 | Sun, Nov 16 | 8970 | 722 | | | | |
| 39 | | | | | | | |
| 40 | CONTROL | | | | PAGEVIEW | | |
| 41 | total | 345543 | 28378 | | standard error | 0.0006 | |
| 42 | combined | 373921 | | | MoE | 0.0012 | |
| 43 | prob of pageview/cookie in either group: | 0.5 | | | CI | 0.4988 | 0.5012 |
| 44 | | | | | p^ (hat) | 0.5006 | |
| 45 | EXPERIMENT | | | | | | |
| 46 | total | 344660 | 28325 | | CLICK | | |
| 47 | combined | 372985 | | | standard error | 0.0021 | |
| 48 | prob of pageview/cookie in either group: | 0.5 | | | MoE | 0.0041 | |
| 49 | | | | | CI | 0.4959 | 0.5041 |
| 50 | | | | | p^ (hat) | 0.5005 | |
| 51 | | | | | | | |
| 52 | | | | | CLICK THROUGH PROBABILITY | | |
| 53 | | | | | 28378/345543 = | 0.0821 | we use instead of 0.5 |
| 54 | | | | | SE | 0.0005 | |
| 55 | | | | | MoE | 0.001 | |
| 56 | | | | | CI | 0.0811 | 0.0831 |
| 57 | | | | | P^ (hat) | 0.0822 | |
| 58 | | | | | | | |
| 59 | | | | | | | |
| 60 | | | | | | | |

Sanity checks all passed for the three invariant metrics. Click Through probability was a bit tricky, because we want to know if the observed rate both come from the same population. To do this we needed to calculate the CI for one group and compare it to the observed rate for the other group.

## Result Analysis

### Effect Size Test

| | | |
|---|---|---|
| | alpha = 0.05 | |
| | Z score = 1.96 | |

| GROSS CONVERSION dmin=.01 | Control | Experiment |
|---|---|---|
| Clicks | 17293 | 17260 |
| Enrollments | 3785 | 3423 |
| gross conversion ratio | 0.2189 | 0.1983 |
| pool^ | 0.2086070674 | |
| SE pool | 0.004371675385 | |
| MoE | 0.0086 | |
| | | |
| d^ | -0.0206 | |
| CI | -0.0292 | -0.012 |
| | | |
| RESULTS: | | |
| Statistcal Sig? | yes | |
| Practical Sig? | yes | |

| NET CONVERSION dmin=.0075 | Control | Experiment |
|---|---|---|
| Clicks | 17293 | 17260 |
| Payments | 2033 | 1945 |
| net conversion ratio | 0.1176 | 0.1127 |
| pool^ | 0.1151274853 | |
| SE pool | 0.003434133513 | |
| MoE | 0.0067 | |
| | | |
| d^ | -0.0049 | |
| CI | -0.0116 | 0.0018 |
| | | |
| RESULTS: | | |
| Statistcal Sig? | no | |
| Practical Sig? | no | |

| | | |
|---|---|---|
| Gross Conversion Confidence Interval | [-0.0292, -0.012] | statistically and practically significant. |
| Net Conversion Confidence Interval | [-0.0116, 0.0018] | statistically and practically NOT significant. |

### Sign Test

| | | control | exp | | control | exp | |
|---|---|---|---|---|---|---|---|
| Date | | enroll/click | enroll/click | success? | payment/click | payment/click | success? |
| Sat, Oct 11 | | 0.1951 | 0.1531 | | 0.1019 | 0.0496 | |
| Sun, Oct 12 | | 0.1887 | 0.1478 | | 0.0899 | 0.1159 | 1 |
| Mon, Oct 13 | | 0.1837 | 0.164 | | 0.1045 | 0.0894 | |
| Tue, Oct 14 | | 0.1866 | 0.1669 | | 0.1256 | 0.1112 | |
| Wed, Oct 15 | | 0.1947 | 0.1683 | | 0.0765 | 0.113 | 1 |
| Thu, Oct 16 | | 0.1677 | 0.1637 | | 0.0996 | 0.0774 | |
| Fri, Oct 17 | | 0.1952 | 0.1628 | | 0.1016 | 0.0564 | |
| Sat, Oct 18 | | 0.1741 | 0.1442 | | 0.1108 | 0.0951 | |
| Sun, Oct 19 | | 0.1896 | 0.1722 | | 0.0868 | 0.1105 | 1 |
| Mon, Oct 20 | | 0.1916 | 0.1779 | | 0.1127 | 0.114 | 1 |
| Tue, Oct 21 | | 0.2261 | 0.1655 | | 0.1211 | 0.0822 | |
| Wed, Oct 22 | | 0.1933 | 0.1598 | | 0.1098 | 0.0874 | |
| Thu, Oct 23 | | 0.191 | 0.19 | | 0.0842 | 0.1059 | 1 |
| Fri, Oct 24 | | 0.3269 | 0.2783 | | 0.1813 | 0.1349 | |
| Sat, Oct 25 | | 0.2547 | 0.1898 | | 0.1852 | 0.1211 | |
| Sun, Oct 26 | | 0.2274 | 0.2208 | | 0.1469 | 0.1457 | |
| Mon, Oct 27 | | 0.307 | 0.2763 | | 0.1634 | 0.1543 | |
| Tue, Oct 28 | | 0.2092 | 0.2201 | 1 | 0.1236 | 0.163 | 1 |
| Wed, Oct 29 | | 0.2652 | 0.2765 | 1 | 0.1164 | 0.132 | 1 |
| Thu, Oct 30 | | 0.2275 | 0.2843 | 1 | 0.1022 | 0.092 | |
| Fri, Oct 31 | | 0.2465 | 0.2521 | 1 | 0.1431 | 0.1704 | 1 |
| Sat, Nov 1 | | 0.2291 | 0.2043 | | 0.1366 | 0.1439 | 1 |
| Sun, Nov 2 | | 0.2973 | 0.2514 | | 0.0967 | 0.1423 | 1 |
| | count | 23 | 23 | | 23 | 23 | |
| | successes | | | 4 out of 23 | | | 10 out of 23 |

## Sign and binomial test

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 one-tail P value is 0.0013
  This is the chance of observing 4 or fewer successes in 23 trials.
- The two-tail P value is 0.0026
  This is the chance of observing either 4 or fewer successes, or 19 or more successes, in 23 trials.

## Sign and binomial test

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 one-tail P value is 0.3388
  This is the chance of observing 10 or fewer successes in 23 trials.
- The two-tail P value is 0.6776
  This is the chance of observing either 10 or fewer successes, or 13 or more successes, in 23 trials.

| | | |
|---|---|---|
| Gross Conversion, two tailed P value | 0.0026 | Statistically Significant |
| Net Conversion two tailed P value | 0.6776 | Not Statistically Significant |

**Summary**

I will not be using the Bonferroni correction, because the correlation between the metrics (e.g. cookies and user IDs) is fairly strong and with two tests the family-wise error rate will not be terribly high. If we were concerned with increased impact of type 1 errors (false-positive results) and we were testing several hypotheses on the same data set all at once, then we would probably want to adjust the alpha by lowering it with the Bonferroni. However in this case I feel it is too conservative of a method to use. Also, there was not a discrepancy between the sign tests and the effect size test.

### Recommendation

This is a tough call, but I might suggest to not recommend the change for two reasons. (i) the gross conversion was slightly negative and within the dmin=0.01 range, therefore we might see a slight decrease in number of users the enroll but ultimately do not pay. This could free up coaching resources for students that do plan to commit fully. However I might argue that students that do sign up for the trial and do not become paying customers probably do not interact with the coaches much anyways, so I am dubious to the claim that they take up more resources. I think they are so busy or focused on other things that they don't put the time in and wind up not becoming paying customers, or they lapse and one payment goes through, then they unenroll. (ii) net conversion was not statistically or practically significant. In fact it was slightly negative (part of the confidence interval) so we might actually wind up slightly decreasing student enrollment and ultimately revenue. Unless we knew for sure that students who signed up for the trial then quit did waste coaching time, this potential decrease might actually wind up hurting Udacity's revenue. Therefore I recommend to not add the extra screen to the lightbox.

## Follow-up Experiment

I think a good follow up experiment would be to look at students who sign up for the trial, become a paying customer then quit after one payment. I am interested in looking if these one time payments are students that either did not like the material or coaching, or just wanted to quit but forgot and then when they were changed they say the bill on their credit card and then they logged back in to unenroll from the coaching.

I would expect that students which forgot to cancel have a certain behavior pattern. Perhaps they signed up then were active, but at some point they stopped activity within the 14 day period. Then they were still inactive once the 14 day period ended, but had just one more case of activity, or maybe a few cases of activity during the 30 day first paid month in which they logged in and canceled. I would expect these users to contrast with other users that canceled after one month due to frustration. In these cases I would expect there to be much more activity during the 1 month they were enrolled.

I think if we knew better what type of user canceled right away, after just one month, we might be able to add that "do you have enough time" screen to the enrollment page. Because if most users that cancel after one month never really logged on, then we can say better that the extra screen, from the experiment I just did above, is going to be possibly detrimental to revenue. However if most students that cancel after one month have a ton of activity, then perhaps that screen would be welcomed, because many of those students that do not have the time or effort, or who are not ready will not wind up starting the free trial and might just go for the free course option instead.

So we would split students into two groups as before. Students that receive the message at enrollment about time commitment and those who do not see it. But this time we will be looking for how students behave after the free trial period and they have become paying customers.

Hypothesis: Users that accidentally enrolled for paid services for 1 month will have some level of measurable activity vs.  users that canceled paid services after 1 month due to frustration.

Unit of Diversion: User-ID (because we care about which users actually canceled after 1 month, so in order to have a paid subscription they need to have a unique user-ID and we want to know which users

Invariant Metrics: number of user-IDs - because we want to know how many users are being put into each group and account for this.

Evaluation Metrics: Will be cancelation rate, within first payment period, before paying again a 2nd time. Also we can use activity rate to determine how active users are during the first payment period. We would want to have a cutoff. This cutoff will explain users that are hardly ever active, i.e. they probably forgot to cancel, vs. very active users.

kyle shannon - P7: A/B Testing - Udacity
List of websites and resources for AB testing:

1. https://discussions.udacity.com/t/struggling-with-final-project-effect-size-tests/34035/7
2. https://discussions.udacity.com/t/p7-how-to-calculate-p-value-for-sign-test/38614
3. http://graphpad.com/quickcalcs/binomial1/
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/pdf/i1949-8357-4-3-279.pdf
5. https://en.wikipedia.org/wiki/Bonferroni_correction
6. https://en.wikipedia.org/wiki/A/B_testing#A.2FB_testing_tools_comparison
7. https://rajivgrover1984.blogspot.com/2015/11/ab-testing-overview.html
8. http://mathworld.wolfram.com/BonferroniCorrection.html