A/B Testing Experiment

Greg Hein – 10/22/2016

Final Project – Udacity Data Analyst Nanodegree: A/B Testing

Experiment Design

Experiment Details

At the time of this experiment, Udacity courses had two options on the home page: "start free trial", and "access course materials". If the student clicked "start free trial", they would be asked to enter their credit card information, and then they would be enrolled in a free trial for the paid version of the course. After 14 days, they would automatically be charged unless they canceled first. If the student clicked "access course materials", they would be able to view the videos and take the quizzes for free, but they would not receive coaching support or a verified certificate, and they would not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead.

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. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course. The experiment will test this hypothesis.

The unit of diversion is a cookie, although if the student enrolled in the free trial, they were tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that did not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

Metric Choice

There were 7 different metrics tracked and recorded in the study:

- **Number of cookies:** That is, number of unique cookies to view the course overview page $(\mathrm{d}_{min} = 3000)$.
- **Number of user-ids:** That is, number of users who enroll in the free trial $(\mathrm{d}_{min} = 50)$.
- **Number of clicks:** That is, number of unique cookies to click the "Start free trial" button, which happens before the free trial screener is triggered $(\mathrm{d}_{min} = 240)$.
- **Click-through-probability:** That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page $(\mathrm{d}_{min} = 0.01)$.
- **Gross conversion:** That is, number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button $(\mathrm{d}_{min} = 0.01)$.
- **Retention:** That is, 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 $(\mathrm{d}_{min} = 0.01)$.
- **Net conversion:** That is, 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 trial" button $(\mathrm{d}_{min} = 0.0075)$.

From these seven metrics, two were chosen as invariant metrics. Invariant metrics are metrics that we expect to be similar in both our control and experiment groups and will provide a sanity check that our experiment is set up properly, and functioning as expected. There are three metrics that are recorded prior to the questioning of the students time commitment. Because they occur before the change we are testing, they should be approximately the same in both control and experimental groups and would make good invariant metrics. The invariant metrics chosen for the experiment are: **Number of Cookies**, and **Number of Clicks**. The third possible invariant metric, **Click-Through-Probability**, will be omitted from the experiment due to the fact that it is defined as **Number of Clicks** divided by **Number of Cookies**, (encompassing the two invariant metrics chosen) and would not be expected to provide any additional information.

There are three evaluation metrics chosen for this experiment. Evaluation metrics are metrics that we expect may show a difference between our control and experiment groups and provide information as to whether or not the change may have had any effect. The three evaluation metrics chosen are: **Gross conversion**, **Retention**, and **Net conversion**. **Gross conversion** was chosen due to the fact that it may provide some information as to whether the additional questioning of students had an effect on enrollment into the free trial. **Retention** and **Net conversion** were chosen due the the fact that they may provide information as to whether the additional student questioning resulted in a change in either the drop out, or completion of the 14 day free trial.

One metric, **Number of User-ids**, was not chosen as either an invariant or evaluation metric. It would not be appropriate as an invariant metric due to the fact that it is not tracked until after the questioning of students time commitment, and therefore would not necessarily be expected to be the same in the control and experimental groups. It is also not a good evaluation metric. The number of enrollments would be expected to be different in control and experimental groups, and may provide some information. But, the **Number of User-ids** will be influenced by the split of cookies into the two groups. This could potentially skew the results and make evaluation difficult. The **Number of User-ids** are not normalized. The chosen evaluation metrics do not have this problem, and therefore are better choices.

As stated earlier, the hypothesis for this experiment 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” For this experiment, the launch criteria will be: to require a practically significant decrease in Gross Conversion (showing a marked influence due to the change), and no decrease in Net Conversion (to verify no resulting loss in revenue). The reason that the evaluation metric of Retention is not included in the launch criteria will be shown later.

Measuring Standard Deviation

The Standard Deviations of our evaluation metrics with a sample size of 5000 cookies visiting the course overview page are calculated below. The data was provided by Udacity in a file of Baseline Values. The number of clicks and enrollments follow a binomial distribution, so the evaluation metrics should have a normal distribution.

$$\text{Number of unique clicks: } 0.08 \times 5000 = 400$$

$$\text{Number of unique user-ids: } 0.08 \times 5000 \times 0.20625 = 82.5$$

$$\text{Standard Deviation Formula: } \mathrm{SD} = \sqrt{\frac{p(1-p)}{N}}$$

$$\text{Gross Conversion: } \mathrm{SD} = \sqrt{\frac{0.20625(1-0.20625)}{400}} = 0.0202$$

$$\text{Retention: } \mathrm{SD} = \sqrt{\frac{0.53(1-0.53)}{82.5}} = 0.0549$$

$$\text{Net Conversion: } \mathrm{SD} = \sqrt{\frac{0.1093(1-0.1093)}{400}} = 0.0156$$

We would expect that with both Gross conversion and Net conversion, the analytic estimate should be comparable to the empirical variability due to the fact that their measure of analysis is unique clicks, the unit of diversion of the experiment. Retention will likely be different due to the fact that its measure of analysis is unique user-ids, which is not the unit of diversion in the experiment

Sizing

Number of Samples vs. Power

The Bonferroni Correction was not used in the analysis phase of the experiment. The choice not to use the Bonferroni Correction was made due to the chosen evaluation metrics and launch criteria.

The number of samples for each evaluation metric are calculated below ($\alpha = 0.05$ and $\beta = 0.2$). The sample size for each group were found with a sample size calculator:

Gross conversion: $25{,}835$ clicks per group
Retention: $39{,}115$ enrollments per group
Net conversion: $27{,}413$ clicks per group

These calculations were then used to determine the number of page views necessary for each evaluation metric:

Gross conversion: $25{,}835 \times (40,000/3,200) \times 2 = 645,875$
Retention: $39{,}115 \times (40,000/660) \times 2 = 4,741,212$
Net conversion: $27{,}413 \times (40,000/3,200) \times 2 = 685,325$

As currently configured, this experiment will require $4{,}741{,}212$ page views for proper evaluation.

Duration vs. Exposure

This experiment does not seem to pose much risk for Udacity. With only one additional pop-up question it is not at all likely to affect Udacity, its website, or the potential students. For that reason 100% of Udacity traffic will be diverted to this experiment.

The Baseline values given by Udacity indicate that there are $40{,}000$ unique cookies to view a page per day. With 100% exposure the amount of time needed for the experiment will be:

$$4{,}741{,}212/40{,}000 = 119 \text{ days}$$

This will be too long of an experiment to run for Udacity, so Retention will no longer be considered as one of the evaluation metrics. This changes the amount of time needed for the experiment to:

$$685{,}325/40{,}000 = 18 \text{ days}$$

Experiment Analysis

The results of the study were provided by Udacity and used in this analysis.

Sanity Checks

As discussed earlier, we expect our invariant metrics to be approximately equal in both the control and experimental groups. To verify that, we will calculate whether or not the observed values fall within the 95% confidence interval around the 50% mark.

Number of Cookies:

$$\alpha = 0.05,\ z^* = 1.96,\ \hat{p} = 0.5,\ N = 690203,\ SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

$$\mathrm{m} = \mathrm{z}^* \times SE = 0.001180$$

$$\mathrm{CI} = [0.4988,\ 0.5012],\ \mathrm{Actual} = 345543/690203 = 0.5006$$

Number of Clicks:

$$\alpha = 0.05,\ z^* = 1.96,\ \hat{p} = 0.5,\ N = 56703,\ SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

$$\mathrm{m} = \mathrm{z}^* \times SE = 0.004115$$

$$\mathrm{CI} = [0.4959,\ 0.5041],\ \mathrm{Actual} = 28378/56703 = 0.5005$$

For both invariant metrics, the actual value falls within the confidence interval. Both sanity checks pass.

Result Analysis

Effect Size Tests

To analyze the results of our experiment, the first test run are Effect Size Tests on our evaluation metrics. The 95% confidence interval is determined for each metric, and from that interval the metric is evaluated for statistical significance, and practical significance.

Gross Conversion $(\mathrm{d}_{min} = 0.01)$**:**

Experimental Rate: $\hat{\mathrm{r}}_{exp} = 3423/17260 = 0.1983$

Control Rate: $\hat{\mathrm{r}}_{control} = 3785/17293 = 0.2189$

$$\hat{\mathrm{d}} = \hat{\mathrm{r}}_{exp} - \hat{r}_{control} = -0.0206$$

$$\hat{\mathrm{r}}_{pool} = (3423 + 3785)/(17260 + 17293) = 0.2086$$

$$\mathrm{SE} = \sqrt{\hat{r}_{pool}(1 - \hat{r}_{pool})(1/N_{exp} + 1/N_{control})}$$

$$= \sqrt{(0.2086)(1 - 0.2086)(1/17260 + 1/17293)} = 0.004372$$

$$\mathrm{m} = \mathrm{z}^* \times SE = 1.96 \times 0.004372 = 0.008568$$

$$\mathrm{CI} = [\hat{\mathrm{d}}\text{ - m, }\hat{\mathrm{d}} + \mathrm{m}] = [\text{-0.02912, -0.01199}]$$

This result is statistically significant, because the confidence interval does not contain zero. It is also practically significant, because the confidence interval does not contain our $\mathrm{d}_{min} = -0.01.$

Net Conversion $(\mathrm{d}_{min} = 0.0075)$**:**

Experimental Rate: $\hat{\mathrm{r}}_{exp} = 1945/17260 = 0.1127$

Control Rate: $\hat{\mathrm{r}}_{control} = 2033/17293 = 0.1176$

$$\hat{\mathrm{d}} = \hat{\mathrm{r}}_{exp} - \hat{r}_{control} = -0.004873$$

$$\hat{\mathrm{r}}_{pool} = (2033 + 1945)/(17293 + 17260) = 0.1151$$

$$\mathrm{SE} = \sqrt{\hat{r}_{pool}(1 - \hat{r}_{pool})(1/N_{exp} + 1/N_{control})}$$

$$= \sqrt{(0.1151)(1 - 0.1151)(1/17260 + 1/17293)} = 0.003434$$

$$\mathrm{m} = \mathrm{z}^* \times SE = 1.96 \times 0.003434 = 0.006731$$

$$\mathrm{CI} = [\hat{\mathrm{d}}\text{ - m, }\hat{\mathrm{d}} + \mathrm{m}] = [\text{-0.01160, 0.001857}]$$

This result is not statistically significant, because the confidence interval contains zero. It is also not practically significant, because the confidence interval contains our $\mathrm{d}_{min} = -0.0075.$

Sign Tests

In addition to the Effect Size test, a second analysis is done for each of our evaluation metrics with a Sign Test. A Sign and Binomial Test calculator was used in this analysis.

$$\text{Number of Trials/Days of Experiment: } 23,\ \ \alpha = 0.05,\ Probability = 0.5$$

Gross Conversion:

$$\text{Number of days where Experiment result} > \text{Control result} = 4$$

$$\text{Two tailed p-value: } 0.0026$$

This result is statistically significant because our $\text{p-value} = 0.0026$ is less than $\alpha = 0.05$.

Net Conversion:

$$\text{Number of days where Experiment result} > \text{Control result} = 10$$

$$\text{Two tailed p-value: } 0.6776$$

This result is not statistically significant because $\text{p-value} = 0.6776$ is greater than $\alpha = 0.05$.

Summary

As stated earlier, the Bonferroni Correction was not used in the analysis phase of the experiment. The Bonferroni Correction is recommended if **any** of the metrics being satisfied would result in launch, a case where it is important that a type I error (false positive) be avoided. In the launch criteria for this experiment, **all** (both) metrics must be satisfied to result in the launch of the experiment. When **all** metrics must be satisfied, type II errors (false negatives) would be more problematic and need to be avoided. The Bonferroni Correction helps to minimize the problems caused by type I errors, at the expense of increased type II errors.

In this experiment, both the Effect Size tests and Sign tests have yielded similar results. In both tests the reduction in Gross Conversion was statistically significant. In the Effect Size test the reduction in Gross Conversion was also practically significant. The change in Net Conversion was not statistically significant in either test. In the Effect Size test it was also not practically significant.

Recommendation

As stated earlier, the launch criteria are: to require a practically significant decrease in Gross Conversion (showing a marked influence due to the change), and no decrease in Net Conversion (to verify no resulting loss in revenue). There was a practically significant decrease in Gross Conversion shown in the Effect Size test. This satisfies the first requirement of the launch criteria. There was no statistically significant change in Net Conversion in either the Effect Size test, or the Sign test. The confidence interval for Net Conversion: $[-0.01160,\ 0.001857]$, contains the negative of practically significant value -0.0075. Net Conversion might have decreased by an amount that would lower revenue by an amount not acceptable to Udacity. The second requirement of the launch criteria is not satisfied. Therefore, launch of the change is not recommended.

Follow-Up Experiment

A potential follow-up experiment with a similar goal, would be to track the amount of time spent in class during the free trial, and at a point prior to the initiation of the paid enrollment remind the students that have not kept pace with the 5 hour per week target that success in Udacity classes typically requires such a time commitment. The hypothesis would be to set clearer expectations for students upfront, thus reducing the number of frustrated students who continued with the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial. So as not to wait until too close to the 14 day mark, students with less than 5 hours spent in class after 7 days would have a pop-up when they enter the class website to remind them that Udacity classes typically require a commitment of at least 5 hours per week, and to this point they have not spent that amount of time in class. It would suggest that they commit themselves to that amount of time moving forward, and if they feel they cannot, it would give them the opportunity to opt-out of the free trial and the ability to cancel their commitment to the class. This would occur for students in the experimental group. The control group would not have the pop-up reminder. The unit of diversion for the experiment would be user-ids, as all students would be enrolled in the free trial at that point, and would be tracked by user-id. This would also serve as a good invariant metric for the experiment. An evaluation metric would Retention: the 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.