Udacity - Data Analyst Nanodegree

Project 7: A/B Testing

Created by: Layne Newhouse

Re-submitted: March 17, 2017

Experiment Design

Metric Choice

Of the metrics to choose from, the number of cookies, number of clicks, and click-through-probability all precede any experimental changes and therefore can not be used as evaluation metrics but can be used as invariant metrics. The number of clicks and click-through-probability are very similar metrics with the exception that click-through-probability normalizes to the size of the control and experiment group. For this reason, the number of cookies and click-through-probability will be chosen as the invariant metrics for this experiment. The number of use user-ids, gross conversion, retention, and net conversion all proceed the experiment change and therefore can be considered when looking for evaluation metrics (for this reason they cannot be used as invariant metrics).

To properly assess evaluation metrics, we must consider the hypothesis, which is that we wish to see a statistically significant reduction in the number of students that enroll in the free trial in conjunction with no change, or a positive change, in the number of students that continue past the free trial. The change in enrolled users can be observed through the number of user-ids as well as gross conversion. Because gross conversion is a rate, it normalizes the data across the control and experiment group and therefore is the preferred metric for this part of the hypothesis. We will be looking for a statistically negative change in this evaluation metric. Retention and net conversion both address the second part of the hypothesis which is looking at the change in enrolled students past the 14-day free trial period. Evaluating both potential metrics, we come across an issue sizing when using retention since the unit of analysis is the number of user-ids to complete checkout as opposed to the number of clicks on the “Start free trial” button. For this reason, we will be using net conversion as our second evaluation metric and the expectation as per our hypothesis is that this will not have a statistically negative change, that is there will not be a reduction in the number of students continuing past the free trail period.

Measuring Standard Deviation

Gross conversion

$$\hat{P}_{GC} = \frac{enrollments}{clicks} = \frac{660}{3200} = 0.2063$$

$$SE_{GC} = \sqrt{\frac{\hat{P}_{GC}(1-\hat{P}_{GC})}{N}} = \sqrt{\frac{0.2063(1-0.2063)}{5000(\frac{3200}{40000})}} = 0.0202$$

Net conversion

$$\hat{P}_{NC} = \frac{\hat{P}_{payment|enrollment} * \ enrollments}{clicks} = \frac{0.53(660)}{3200} = 0.1093$$

$$SE_{NC} = \sqrt{\frac{\hat{P}_{NC}(1-\hat{P}_{NC})}{N}} = \sqrt{\frac{0.1093(1-0.1093)}{5000(\frac{3200}{40000})}} = 0.0156$$

For the evaluation metrics, I expect the analytical estimate to be comparable to the empirical variability since the unit of analysis is the same as the unit of divergence, which in both cases is a cookie.

Sizing

Number of Samples vs. Power

Using http://www.evanmiller.org/ab-testing/sample-size.html the following sample sizes were obtained. The Bonferroni correction will be used during this analysis.

$$\propto_{individual} = \frac{\propto_{overall}}{n_{metrics}} = \frac{0.05}{2} = 0.025 \qquad \beta = 0.2$$

Gross conversion

$$Baseline\ conversion = 0.2063 \qquad d_{min} = 0.01$$

$$n_{GC} = 31{,}435_{clicks} * \left(\frac{40{,}000_{pageviews}}{3200\ _{clicks}}\right) * 2 = 785{,}876_{pageviews}$$

Net conversion

$$Baseline\ conversion = 0.1093 \qquad d_{min} = 0.0075$$

$$n_{NC} = 33{,}335_{clicks} * \left(\frac{40{,}000_{pageviews}}{3200\ _{clicks}}\right) * 2 = 833{,}376_{pageviews}$$

*Note: we multiply the total pageviews by two to obtain adequate pageviews for both the control and the experiment groups

Therefore, to power the experiment properly we will need 833,376 pageviews.

Duration vs. Exposure

There are a few considerations in the assessment of risk, two of which are: is there a chance that anyone gets hurt because of the experiment and are we dealing with sensitive information such as political attitudes, personal disease history, sexual preferences etc. Since our experiment does not hurt people or anything to do with sensitive information we can consider it low risk and can expose all of the traffic to the experiment which would result in completing the experiment in 21 days. Given the ability to ramp up the exposure to the project I would consider beginning with diverting a smaller fraction of the traffic and increasing this number to 1 such that any bugs or glitches in the experiment would be caught early on.

Experiment Analysis

Sanity Checks

Number of cookies (pageviews)

$$\hat{P} = \frac{N_{cont}}{N_{total}} = \frac{345{,}543}{345{,}543 + 344{,}660} = 0.5006$$

$$SE = \sqrt{\frac{P(1-P)}{N_{total}}} = \sqrt{\frac{0.5(1-0.5)}{345{,}543 + 344{,}660}} = 0.0006$$

$$m = Z * SE = 1.96(0.0006) = 0.0012$$
$$CI = P \pm m = [0.4988, 0.5012]$$

Since $\hat{P}$ is within the confidence interval, the sanity check for this metric passes.

Click-through-probability

$$\hat{P}_{cont} = \frac{x_{cont}}{N_{cont}} = \frac{28378}{345543} = 0.0821$$

$$SE_{cont} = \sqrt{\frac{\hat{P}_{cont}(1-\hat{P}_{cont})}{N_{cont}}} = 0.00047$$

$$m = Z * SE = 1.96(0.00047) = 0.009$$
$$CI = \hat{P}_{cont} \pm m = [0.0812, 0.0830]$$

$$\hat{P}_{exp} = \frac{x_{exp}}{N_{exp}} = \frac{28325}{344660} = 0.0822$$

Since $\hat{P}_{exp}$ is within the confidence interval, the sanity check for this metric passes.

Result Analysis

Effect Size Tests

Gross conversion

$$\hat{d} = r_{exp} - r_{cont} = \frac{x_{exp}}{N_{exp}} - \frac{x_{cont}}{N_{cont}} = \frac{3426}{17{,}260} - \frac{3785}{17{,}293} = -0.0206$$

$$\hat{P}_{pool} = \frac{x_{exp} + x_{cont}}{N_{exp} + N_{cont}} = 0.2086$$

$$SE_{pool} = \sqrt{\hat{P}_{pool}(1 - \hat{P}_{pool})\left(\frac{1}{N_{exp}} + \frac{1}{N_{cont}}\right)} = 0.0044$$

$$m = Z * SE = 1.96(0.0044) = 0.0086$$

$$CI = \hat{d} \pm m = [-0.0292, -0.0120]$$

Statistically significance is achieved in this metric since the confidence interval does not include 0. It is also practically significant since $|-0.0120| > (d_{min} = 0.01)$.

Net conversion

$$\hat{d} = r_{exp} - r_{cont} = \frac{x_{exp}}{N_{exp}} - \frac{x_{cont}}{N_{cont}} = \frac{1945}{17{,}260} - \frac{2033}{17{,}293} = -0.0049$$

$$\hat{P}_{pool} = \frac{x_{exp} + x_{cont}}{N_{exp} + N_{cont}} = 0.1151$$

$$SE_{pool} = \sqrt{\hat{P}_{pool}(1 - \hat{P}_{pool})\left(\frac{1}{N_{exp}} + \frac{1}{N_{cont}}\right)} = 0.0034$$

$$m = Z * SE = 1.96(0.0034) = 0.0067$$

$$CI = \hat{d} \pm m = [-0.0116, 0.0018]$$

Neither statistically significance or practical significance is achieved in this metric since the confidence interval includes 0.

Sign Tests

Using https://graphpad.com/quickcalcs/binomial1.cfm the following p-values were obtained.

Gross conversion

$$\#\ (+)\ \Delta Enrollments\ Days = 19\ (successes)$$
$$\#\ Total\ Enrollment\ Days = 23$$
$$P = 0.5$$

Using graphpad, we obtain a two-tail p-value of 0.0026. Since this is less than α = 0.05 the Gross conversion sign test is statistically significant.

Net conversion

$$\#\ (+)\ \Delta Enrollments\ Days = 13\ (successes)$$
$$\#\ Total\ Enrollment\ Days = 23$$
$$P = 0.5$$

Using graphpad, we obtain a two-tail p-value of 0.6776. Since this is greater than α = 0.05 the Net conversion sign test is not statistically significant.

Summary

To propose the recommendation, we will need the both evaluation metrics, net and gross conversion, to match our expectations (we look for a decrease in gross conversion and for a no decrease in the net conversion). Because of this the Bonferroni correction must be used in order for us to propose the recommendation using an overall alpha (α) of 0.05. The experiment showed that there was a significant change in the gross conversion rate as this metric went down by at least the practical significance boundary. The minimum change needed for this metric was 0.01 and the observed confidence interval for the change was $[-0.0292, -0.0120]$. The net conversion has shown no significant change with the confidence interval of $[-0.0116, 0.0018]$. The minimum difference considered significant for this metric is 0.0075. There was no discrepancy between the hypothesis test and the sign test and the

Recommendation

This experiment has shown a significant change in the gross conversion rate as predicted by the alternative hypothesis. Although net conversion does not show a significant decrease the confidence interval does include the negative of the practical significance boundary. Therefore, it is possible that the number went by an amount that would matter to the business. For this reason, I would not recommend implementing the proposed change in this experiment but instead perform a follow up experiment that could be designed and implemented in attempt to increase net conversion following this experiments change. (See next section for details.)

Follow-Up Experiment

In this experiment, Udacity will test a change where if a student spends too long on a specific quiz questions a small ‘Need Help?’ button will float into the top right side of their screen. If this button is clicked it will display a pop-up showing some of the most active forum posts related to this question, which they can click on and read through.

The hypothesis is that this pop-up will help students in the free trial period understand how to properly take advantage of the forums and get past questions that they are having trouble with, which may be causing frustration. If this hypothesis is held true, Udacity could improve the overall student experience and increase the number of students likely to complete the course.

The unit of diversion is user-id since the users will already be enrolled in the 14-day free trail period.

The invariant metric would be unique user sign-ins per day and the evaluation metric would be number of user-ids that continue past the 14-day free trial period.