Project 7: Design an A/B Test

Experiment Design

Metric Choice

Invariant metrics are number of cookies, number of clicks, and click-through-probability.
Evaluation metrics are gross conversion and net conversion.

There are 3 metrics that are not impacted by the test: number of cookies, number of clicks, and click-through-probability. They are therefore chosen as invariant metrics. To make it detailed, number of cookies and number of clicks are chosen because we want to make sure the populations in control group and experiment group are close. Click-through-probability is chosen as an invariant metric since it is not expected to differ in the control group and the experiment group. These three metrics are used for our sanity check.

4 other metrics are then our potential evaluation metrics.

Gross conversion is chosen as an evaluation metric because we want to check if the experiment would lead to a lower enrollment rate, which can be described by gross conversion. If it is true, then this experiment does help to reduce students who enrolled in the free trial.

Net conversion is chosen as the other evaluation metric because we want to check if the experiment would not impact net conversion. If it is true, then this experiment does not reduce students who continued past the free trial.

However, number of use-ids is not chosen since it is a raw count, which is affected by the number of cookies in each group. It is difficult to justify the difference of it between two groups if the numbers of cookies in two groups are different. The reason that retention is dropped is because it takes too long to gather enough data to draw the conclusion. Details would be explained later in part “Sizing”.

Measuring Standard Deviation

The standard deviation of gross conversion is 0.02023.
The standard deviation of net conversion is 0.01560.

These metrics are probabilities and can be considered as binomial distributed. Given a large size of data, the average values of these metrics should become normal distributed and the analytical standard deviations should be accurate. The other issue to think about is independent sampling. In our case, these two metrics have the same denominator, number of cookies. At the same time, the unit of diversion is a cookie, consistent with the denominator. Therefore, we can trust the analytical estimates.

Sizing

Number of Samples vs. Power

I do not use the Bonferroni correction during the analysis phase. The number of pageviews I need to power the experiment is 685325.

If I use retention as an evaluation metric, then the number of pageviews I need would become 4737818. Given this number, we would need at least 119 days to do the test, even if 100% of traffic is diverted to this experiment. The time spread of 119 days is too long for our test. This is why retention is not chosen as an evaluation metric.

Duration vs. Exposure

I would divert 70% of traffic to this experiment. Given this, I would need 25 days to run the experiment.

I use 70% for the fraction because I think this test is of minimal risk. It won't put any physical or emotional concerns to users, and no additional sensitive information is asked to provide. I choose not to use 100% just to the time spread long enough to avoid biased results resulted from seasonality.

Experiment Analysis

Sanity Checks

For number of cookies and number of clicks, I calculated the confidence intervals around the fraction of events expected to be assigned to the control group. The observed values are the actual fraction that was assigned to the control group.

95% confidence interval for number of cookies: [0.4988, 0.5012]
95% confidence interval for number of clicks: [0.4959, 0.5041]
95% confidence interval for click-through-probability: [-0.001296, 0.001296]

The observed value of number of cookies is 0.5006, the observed value of number of clicks is 0.50047, and the observed value of click-through-probability is 0.00005663. Therefore, all metrics pass the sanity check.

Result Analysis

Effect Size Tests

95% confidence interval for gross conversion: [-0.02912, -0.01199]
95% confidence interval for net conversion: [-0.01160, 0.001857]

Since the 95% confidence interval for gross conversion does not include 0 and -0.01, it is statistically and practically significant.

Since the 95% confidence interval for net conversion does include 0, it is not significant statistically or practically.

Sign Tests

The p-value of the sign test for gross conversion: 0.0026
The p-value of the sign test for net conversion: 0.6776

The result for gross conversion is statistically significant, while the result for net conversion is not statistically significant.

Summary

I did not use the Bonferroni correction. Actually the correlation between our two metrics is 0.4138, which does not show strong correlation. However, we need both metrics to meet the criteria to launch: gross conversion is lower in experiment group and net conversion is unchanged in experiment group. The Bonferroni correction is not useful in our case. It is useful to reduce the risk when we would launch if any metric meets the criteria. If we apply the Bonferroni correction in our case, we would increase the risk of failing to launch because some metrics are rejected by mistake.

The effect size hypothesis tests and the sign tests have consistent results.

Recommendation

I recommend not launching this experiment. Although it clearly lowered gross conversion and did not significantly change net conversion, most (86.2%) of the confidence interval for net conversion is negative. Moreover, the lower boundary of the confidence interval for net conversion is lower than -0.01, which is the negative of net conversion practical significance boundary. Therefore, additional test is needed to decide whether net conversion is really unchanged in the experiment or not.

Follow-Up Experiment

If I wanted to reduce the number of frustrated students who cancel early in the course, I would try an experiment to send emails to students asking them if they want to receive weekly emails about their progress summary to keep track of their performance comparing to the average.

The hypothesis is that if students accepted to receive these emails, they are more aware of their progress and have a clearer mindset to adjust their efforts weekly hence they are less likely to cancel early in the course.

Number of user-ids would be used as an invariant metric because we want to make sure the populations in control group and experiment group are close.
Cancel rate would be used as an evaluation metric. Cancel rate can be defined as the number of user-ids that cancelled within first month of the course divided by the number of user-ids that enrolled in the course. We want check if this rate would be lower in the experiment group.

We also want to track the cancel rate of students who did not opt in receiving emails in the experiment. This is to check whether the sample of students who chose to opt in is biased or not.

The unit of diversion would be user-id, because this experiment is based on students who have enrolled in the course.

Reference:

1. http://graphpad.com/quickcalcs/binomial1.cfm
2. https://en.wikipedia.org/wiki/Bonferroni_correction
3. http://napitupulu-jon.appspot.com/posts/variability-abtesting-udacity.html
4. http://www.evanmiller.org/ab-testing/sample-size.html