P7: Design an A/B Test

Carvus Byrd IV

Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here.

Invariant Metrics	Evaluation Metrics
Number of Cookies Number of Clicks Click-Through-Probability	Gross Conversion Retention* Net Conversion

*Later removed as an evaluation metric

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

Metric	Classification	Reasoning for Classification
Number of Cookies	Invariant	Independant from experiment because the visits occur before user reaches experiment.
Number of Clicks		Independant from experiment because the visits occur before user reaches experiment.
Click-Through-Probability		Independant from experiment because the visits occur before user reaches experiment.
Gross Conversion	Evaluation	Directly dependent on the effect of the experiment. Displays whether we managed to decrease the cost of enrollments that aren't likely to become steady paying customers.
Net Conversion		Directly dependent on the effect of the experiment. Displays financial impact of change. Net Conversion does not have a denominator and therefore does not normalize.
Retention	Evaluation (Not Used)	Directly dependent on the effect of the experiment. Displays financial impact of change.
Number of User-IDs	Neither	The number of users who enroll in the free trial is dependent on the experiment so it's not a good invariant metric. It's also not a good evaluation metric because the number of visitors may be different between the experiment and control groups.

I will look for Gross Conversions to have a practically and statistically significant decrease as it will show us whether we lower our costs by introducing the screener.

I will look for Net Conversions to **not** have a statistically significant decrease as it will show how the change affects our revenues.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics.

Metric	Standard Deviation
Gross Conversion	0.0202
Retention*	0.0549
Net Conversion	0.0156

*Power requirements made this metric unsuitable to evaluate in this experiment

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

We can use the analytical estimate of the variance only because Gross Conversion and Net Conversion both use the number of cookies as their denominator and are probabilities, which tend to be binomial in large samples.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately

I did not use the Bonferroni correction.

I originally wanted to use Retention but the power needed would not be realistic for an experiment (almost 7 times the power needed). Using only Gross Conversion and Net Conversion will only require about 685,325 pageviews of power.

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment.

My initial plan would be to split the traffic 50/50 to ensure we account for any type of seasonality that could occur. This split would need about 35 days to reach required power levels. The project instructions detail that they experiment should not last more than a few weeks. Therefore, I can split the data 90% experiment and 10% control to have the test only run about 20 days.

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

I believe there is little risk in the experiment as there is no sensitive data involved and it does not affect current customers. The only major risk is if the experiment negatively affects new enrollments (potential customers) and is a great reason to keep a small percentage (10%) within the control group.

Experiment Analysis

Sanity Checks

	Pageviews (Cookies)	Clicks	CTR
Probability	0.5	0.5	0.0821
SE	0.0006	0.0021	0.0005
M	0.0012	0.0041	0.0009
Lower CI	0.4988	0.4959	0.0812
Upper CI	0.5012	0.5041	0.0830
Observed	0.5006	0.5005	0.4998
Pass?	Yes	Yes	Yes

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. Do not proceed to the rest of the analysis unless all sanity checks pass.

All metrics passed.

Result Analysis

Effect Size Tests

	Gross Conversions	Net Conversions
r_pool	0.2086	0.1151
d_min	0.0100	0.0075
r_hat_cont	0.2189	0.1176
r_hat_exp	0.1983	0.1127
d_hat	-0.0206	-0.0049
SE_pool	0.0044	0.0034
M	0.0086	0.0067
Lower CI	-0.0291	-0.0116
Upper CI	-0.0120	0.0019
Statistically Significant?	Yes	No
Practically Significant	Yes	No

Sign Tests

	Gross Conversions	Net Conversions
Diff < 0	19	13
Diff Total	23	23
P-value	.0026	.6776
Statistically Significant?	Yes	No

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

I decided not to use a Bonferroni correction as we are only testing one variation, both of our metrics need to be significant before we will launch the change. Since both (or all) of our metrics have to be significant, the Bonferroni correction would be too conservative. Our risk for a false positive launch is already low enough.

It would only be appropriate to use the Bonferroni correction if it was decided to do post-experiment segmentation on the results and would be satisfied with only certain “sub” metrics meeting significance standards before launching (i.e. Perhaps we found that the metrics were significant on weekend days but not weekdays. Do we need all metrics to be significant before launching?).

There were no discrepancies between the Effect Size Tests and the Sign Tests as both indicated a change in Gross Conversions but not Net Conversions.

Recommendation

Make a recommendation and briefly describe your reasoning.

My recommendation is to experiment further. This test does show we have achieved a reduction (statistically and practically) in the amount of users committing to a free trial, Gross Conversions (hopefully because of our efforts to help users understand time commitments). However, the test also showed the change in Net Conversions (users who remain for two weeks and become customers) to be statistically and practically insignificant. The confidence interval for Net Conversions includes negative numbers within its range and overlaps with its practical significance mark. Therefore, there is a risk that the introduction of the trial screener may lead to a decrease in revenue.

I would propose an additional experiment to better understand the changes, or lack of changes, in Net Conversions. This could be through a redesign of the experiment or through increasing the power of the experiment.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

My idea for a follow up experiment would be very similar to the original approach except for one key detail: length of free trial. I believe Udacity offers a great service from start to finish and they need to better expose the whole process to new customers.

As I am finishing my 2nd Nanodegree, I usually bank on the expectation that each project will take me about one month. During that month I am exposed to each value added service Udacity provides. If a user had a trial period of 30 days (about one month), I believe they would have more time to understand if they can fit the workload into their schedule, become better exposed to Udacity's services (including the opportunity to submit a project and receive feedback), and perhaps feel a certain expectation to finish the degree given that they have already completed one project.

Therefore, my null hypothesis is that extending the free trial from 14 days to 30 days will not result in a significant difference between the control and experiment samples. My alternate hypothesis is that extending the free trial from 14 days to 30 days will result in a significant difference between the control and experiment samples.

The unit of diversion will be the number of clicks (The number of unique cookies to click the "30 Day Free Trial" button).

Invariant Metrics	Evaluation Metrics
Number of Cookies Number of Clicks Click-Through-Probability	Gross Conversion Net Conversion

Invariants Metrics are chosen for the same reasoning as above "Independant from experiment because the visits occur before user reaches experiment." Also, Number of User-IDs and Retention are removed.

Metric	Classification	Reasoning for Classification
Gross Conversion	Evaluation	Directly dependent on the effect of the experiment. Displays whether we managed to decrease the cost of enrollments that aren't likely to become steady paying customers.
Net Conversion	Evaluation	Directly dependent on the effect of the experiment. Displays positive financial impact of change.

The launch criteria would be exactly the same as above:

I will look for Gross Conversions to have a practically and statistically significant decrease as it will show us whether we lower our costs by introducing the screener.

I will look for Net Conversions to **not** have a statistically significant decrease as it will show how the change affects our revenues.

Resources

Binomial Test
http://vassarstats.net/binomialX.html

Testing Work
https://docs.google.com/spreadsheets/d/1Fpo5sUJ7hBuMkfyIwS8_DmA68CmXkgn3RG_xara2oSw/edit?usp=sharing