Udacity Mock A/B Testing

Experiment Design

Metric Choice

Due to the obvious mutual exclusivity between evaluation and invariant metrics, reasons below are only given for choosing them as one kind, but omitted for not choosing them as the other.

Evaluation metrics:

- **Gross conversion** can be used as a proxy to measure the first half of the hypothesis ("reducing the number of frustrated students who left the free trial"), since the experiment tries to divert students who don't have enough time away from the free trial in the first place, thus preventing them from getting frustrated later on.
- **Net conversion** basically paraphrases the second half of the hypothesis ("without significantly reducing the number of students to continue past the free trial and eventually complete the course"). Hopefully this metric won't drop significantly.

In order to launch the experiment, we would look for a significantly decreased gross conversion, and an absence of significantly decreased net conversion.

Invariant metrics:

- **Number of cookies**, as a population sizing metric, should be comparable for the two groups, because cookie is the unit of diversion.
- **Number of clicks** and **click-through-probability** are not expected to change, as the experiment should only impact what happens after the students click the "start free trial" button.

Leftout metrics:

- **Number of user-ids** should not be invariant, since a decrease of enrolled number of user-ids should be expected. And as a count, it's not as a good evaluation metric as the gross conversion, which also accounts for the number of unique cookies clicking the trial button.
- **Retention** would have been chosen as an evaluation metric if the time needed had been reasonable. But with 4,741,213 unique cookies of pageviews required, it would have taken 149 days to collect the data, which is too long to be practical.

Measuring Standard Deviation

SE(gross conversion) = 0.02023060414

SE(net conversion) = 0.01560154458

First a quick justification of the calculation: The two standard deviations of the sample proportions (standard errors of the true proportions) were calculated based on binomial/normal distributions, because of the two exclusive outcomes (enroll or not, and pay or not) i.i.d qualifying binomial distributions, and large enough sample sizes justifying the normal approximations (both 5k * 0.08 * 0.21 and 5k * 0.08 * 0.11 are way bigger than the threshold 5).

Since the unit of analysis matches the unit of diversion (both are number of unique cookies), analytic estimates should be closely comparable to empirical variabilities. Nevertheless, it might still be worth doing empirical estimates as sanity checks.

Sizing

Number of Samples vs. Power

We will need 685,325 unique cookies of pageviews without using the Bonferroni correction.

Duration vs. Exposure

Since the experiment won't be of high risk (no one is really in harm's way and time commitment to learning does not qualify as sensitive data), the fraction of traffic won't need be strictly limited. Although Udacity could potentially risk a decline of revenues, if that occasion materializes, the experiment can be quickly terminated to prevent further losses, since we will be monitering the net conversion closely.

However, Udacity might still want to run other experiments concurrently, so 80% seems to be a reasonable fraction to divert to the current experiment, with 40% going to each branch.

With 40,000 unique cookies of daily pageviews as the baseline, the length of the experiment works out to be 22 days, within the range of "a few weeks".

Experiment Analysis

Sanity Checks

	95% CI	Actual Value
Number of cookies	(0.4988, 0.5012)	0.5006
Number of clicks	(0.4959, 0.5041)	0.5005
Click-through-probability	(-0.0012, 0.0013)	0.0000566

All of the invariant metrics have passed sanity checks.

Result Analysis

Effect Size Tests

	95% CI	Statistically Significant	Practically Significant
Gross conversion	(-0.0291, -0.0120)	Yes (<0)	Yes (<-0.01)

Net conversion	(-0.0116, 0.0019)	No	No

Sign Tests

	P-value	Statistically Significant
Gross conversion	0.0026	Yes (<0.05)
Net conversion	0.6776	No

Summary

I have not used the Bonferroni correction, because it "controls for false positives at the expense of power (with increasing false negatives)", which is too conservative for our case. Because our launch criteria are based on both metrics, increased false negatives on either of them would unnecessarily inhibit our decision to launch the experiment.

And there is no apparent discrepancy between the effect size hypothesis tests and the sign tests.

Recommendation

Gross conversion has significantly decreased (both statistically and practically), and net conversion has not significantly changed statistically, seemingly suggesting to launch the experiment.

However, there is a wrench in the works: the confidence interval of net conversion captures the practical significance level, which means there is a possible chance of a practically significant decrease.

In light of this, we either need to iterate with further experiments of more power, or pass on launching the it altogether.

Follow–Up Experiment

If we want to reduce the number of frustrated students who cancel early in the course, we might first need to conduct some user experience researches, focus groups, or surveys, to identify possible factors leading to early cancellation. Suppose we find that a lot of students who cancel early think the courses could do better by explaining concepts in greater details.

Then we can feature an experiment group exposed to deeper explained contents and a control group for comparison.

The hypothesis would be that, with finer explanation of the contents, there might be fewer frustrated students who cancel early. Although we did not have time to evaluate retention in the previous experiment, it could be a good time to do so this time, as the metric directly measures the new hypothesis.

In order to squeeze the new experiment into a more realistic timeframe, we might need to refine the metric of "retention" by adjusting the threshold from 14 days to some other value, and/or increase α, β, or/and dmin.

And the unit of diversion would be user-ids, since we only care about users who are already enrolled, and it matches the unit of analysis, which would make our analytic estimates of variabilities more accurate.

References

Inference for Proportions