

# 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 238 days to collect data, which is too long to implement.

## Measuring Standard Deviation

**SE(gross conversion)** = 0.02023060414

**SE(net conversion)** = 0.01560154458

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 does not seem to be especially of high risk (only students who don't have enough time are encouraged not to enroll in the free trial, with no material change of user experiences and infrastructures), the fraction of traffic won't need be strictly limited.

However, Udacity might still want to run other experiments concurrently and probably won't want to draw too much publicity on the experiment, so 50% seems to be a reasonable fraction to divert to the current experiment, with 25% going to each branch.

With 40,000 unique cookies of daily pageviews as the baseline, the length of the experiment works out to be 35 days.

## 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 is too conservative when metrics are correlated and it's rather unreasonable to assume that gross conversion and net conversion are independent.

And there does not seem to be any discrepancies between the effect size hypothesis tests and the sign tests.

## Recommendation

Since gross conversion has indeed significantly decreased (both statistically and practically), and net conversion has not significantly changed, which supports our hypothesis, I would definitely recommend launching the experiment.

## 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. Then we should design an experiment implementing changes targeting those factors with a control group in comparison. Without the preliminary steps, we would not be able to know which changes to test in the experiment.

The hypothesis would be that, with the changes dealing with those factors, retention would increase. Although we did not have time to evaluate it in the previous experiment, it could be a good time to do so, as the metric directly measures our hypothesis.

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](#)