Design an A/B test: Free Trial Screener
Yuchen Yeh, December 2016

I. Experiment Design

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead.This screenshot shows what the experiment looks like.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

This spreadsheet contains rough estimates of the baseline values for these metrics, and the experiment data for analysis is here.

II. Metric Choice

Invariant Metrics:
Number of cookies
Number of clicks
Click-through-probability

Evaluation Metrics:
Gross conversion
Net conversion

- **Number of cookies**: This is an invariant metric but not an evaluation metric, as the number of cookies should be the same in the control group and the experiment group to carry out the experiment.
- **Number of user-ids:** This is not an invariant metric as adding a screener message does affect the number of users who enroll in the free trial. Also, this metric is not used for

evaluation because it is not not normalized but it could potentially be an evaluation metric to track the first part of the hypothesis.

- **Number of clicks:** This is an invariant metric not an evaluation metric, as the number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger) should be the same in the control group and the experiment group to carry out the experiment.
- **Click-through-probability:** This is an invariant metric but not an evaluation metric, as adding a screener message does not affect the number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.
- **Gross conversion:** This is not an invariant metric as adding a screener message does affect the ratio of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button.
- **Retention:** This is not an invariant metric as adding a screener message does affect the number of user-ids to remain enrolled past the 14-day boundary divided by number of user-ids to complete checkout. However, this is not an evaluation metric as it takes a long time to evaluate the result.
- **Net conversion:** This is not an invariant metric as adding a screener message does affect the number of user-ids to remain enrolled past the 14-day boundary divided by the number of unique cookies to click the "Start free trial" button.

It is expected that the gross conversion in the experiment group would be less when compared with the control group, as the screener message intends to prevent non-qualified students from enrolling. However, net conversion in the experiment group should not any significant decrease in the comparison of the control group. In order to launch the experiment, net conversion in the experiment group should show statistically significant changes at a confidence level of 95%. And gross conversion should not see a significant decrease.

III. Variability

Measuring Standard Deviation

Gross conversion: standard deviation of 0.0202
(SQRT(0.20625*(1-0.20625)/(5000*3200/40000))

Net conversion: standard deviation of 0.0156 (SQRT(0.1093*(1-0.1093)/(5000*3200/40000))

Both of the evaluation metrics chosen have the same the denominator (the unit of analysis), which is number of cookie. The unit of diversion is also unique cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. Therefore, the analytic variability is likely to match the empirical variability, and it is best to use empirical variability for the evaluation metrics.

IV. Sizing

Number of Samples vs. Power

No, we need both evaluation metrics to show significant changes. Therefore, Bonferroni correction is not going to be used during the analysis phase. Bonferroni correction should be used when only one or some of metrics are needed for making a decision.

Use an alpha of 0.05 and a beta of 0.2.

Probability of enrolling, given click:
20.63% base conversion rate, 1% min d.
Samples needed: 25,839
Pageviews: 25,839/(3200/40000) = 322,988

Probability of payment, given click:
10.93% base conversion rate, 0.75% min d.
Samples needed: 27,411
Pageviews: 27,411/(3200/40000) = 342,637.5

Since the samples are needed for both the experiment group and control group, the required number of samples is 685,275 (doubling of 342,637.5).

Duration vs. Exposure

Since this experiment doesn't deal with sensitive data and the experiment still allows free trial enrollments after seeing the screener, this is a low risk experiment for Udacity.

The fraction of sample that would be diverted for this experiment is 1 to allow Udacity to see the result of the experiment in a few weeks.

Give the the required samples of 685,275 and 100% of daily traffic diverted for this experiment, 18 days are required to run this experiment.

V. Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check.

	Lower Bound	Upper Bound	Observed	Pass
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks on “Start free trial”	0.4959	0.5041	0.5005	Yes
Click-through-probability on “Start free trial”	0.0812	0.0830	0.0822	Yes

VI. Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant.

Control Gross Conversion = 0.218874689
Experiment Gross Conversion = 0.198319815
D hat = -0.02055

	Lower Bound	Upper Bound
Gross conversion	-0.0291	-0.0120

Given 1% Minimum Detectable Effect(Dmin), the lower bound of the confidence interval is -0.0291 and the upper bound of the confidence interval is -0.0120. The confidence interval is significantly lower than -1% Dmin so it is practically significant. Also, it is statistically significant since the confidence interval does not include 0, which means there is a significant change.

Control Net Conversion = 0.536
Experiment Net Conversion = 0.5682
D hat =- 0.0049

	Lower Bound	Upper Bound
Net conversion	-0.0116	0.0019

The confidence interval is way below the Given 0.75% Minimum Detectable Effect, so it is not practically significant. Also, the confidence interval includes 0, which means there is not a statistically significant change.

VII. Sign Tests

For each evaluation metric, do a sign test using the day-by-day breakdown.

Gross conversion: p value of 0.0026 (4 successes of 23 trials)
Net conversion: p value of 0.6776 (10 successes of 23 trials)

It shows gross conversion is significantly lower in the experiment group, but net conversion doesn't show any significant changes.

VIII. Summary

Bonferroni correction is not used during the analysis phase, as we need both evaluation metrics (gross conversion and net conversion) at the same time to inform the decision. The Bonferroni correction is designed to reduce the risk that one metric is deemed significant by mistake, in the situation where we need just one metric to meet expectations in order to launch an experiment.

The effect size tests and the sign tests both show the experiment doesn't shows significant changes in net conversions, although we see a significant decrease in gross conversions.
The effect size tests summarise the difference during/throughout the experiment, but sign tests show the daily difference for the experiment.

IX. Recommendation

Based on this experiment, we are not going to launch the change. It is within the expectation that adding a free trial screener significantly decreased the number of students enrolling to free trial as result of lack of enough time for the course. However, this experiment sees a negative practical significance boundary, which means net conversions could go down by an amount that could affect the business. The business did not expect gross conversion to have a significant decrease. Therefore, it is not an acceptable risk to launch the free trial screener.

X. Follow-Up Experiment

To reduce the number of frustrated students that abandon early, a follow-up experiment is to add an incentive of 50% discount on the first month immediately after enrollment. The hypothesis is that the conversion rate from free trial students to payment students should see an increase.

Invariant metrics is:

- Number of user-ids: That is, number of users who enroll in the free trial. We need to have the same number of user ids in the experiment group and the control group.

Evaluation metrics is:

- Retention: number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout.

The unit of diversion is user-ids. The number of users who enroll in the free trial will be assigned evenly in the experiment group and the control group to test if the incentive does help to reduce the number of frustrated students that abandon early.