A/B Testing

Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

My choice of invariant metrics:
Number of cookies (# of unique cookies to view the course overview page)
Number of clicks (# of unique cookies to click the “start free trial” button)
Click-through-probability (# of unique cookies to click the “start free trial” button divided by number of unique cookies to view the course overview page)

My choice of evaluation metrics:
Gross conversion (# of user-ids to complete the checkout and enroll in the free trial divided by # of unique cookies to click “start free trial” button)
Net conversion (# of user-ids to remain enrolled past the 14-days boundary divided by the number of unique cookies to click the “start free trial”)

Number of cookies: The unit of diversion is based on cookies. This metric could be used to check the equivalences between the control & experiment groups and verify that the division has been performed correctly. Since it is measured before the screener pops up, it won’t be affected and will be a good invariant metric.

Number of user-ids: It isn’t selected as an invariant metric, because it is the count of user-ids enrolled in the course which is after the change appears and might be affected by the change. It isn’t selected as an evaluation metric neither, because it is not a normalized value and difficult to compare. Gross conversion rate would be a better metric to choose which considers both the # of user-ids as well as the # of cookies which click through.

Number of clicks: This should be invariant between control and experiment because in experiment group the change happens after users clicked the “start free trial” button. The clicks should be evenly distributed across two groups.

Click-through-probability: This is calculated by # of clicks divided by # of cookies. Both the numerator and denominator are invariant, this metric should be invariant between the control and experiment group as well. Also in experiment the change happens after users clicked the “start free trial” button. This metric shouldn’t be affected because it is measured before the change is triggered.

Gross conversion: the hypothesis of the experiment is that the screener will reduce the # of students to enroll the free trials, because those who don't have suggested time commitment will choose not to proceed. Gross conversion is exactly what we want to measure in order to evaluate the impact of the change. We expect that there will be a drop in gross conversion.

Retention: Retention isn't selected as an evaluation metric because it requires too many page views (4,741,212) to achieve the desired α, β and minimal detectable practical difference. It is not appropriate and feasible to conduct an experiment which lasts for 4 months.

Net conversion: This is a good metric to evaluate what proportion of students are still enrolled after the 14-day trial and make sure there won't be any negative impact by introducing this screener. As we expected if the change is effective to help students to make an informed decision, there might be a rise or at least no change in net conversion.

We should look into both gross conversion & net conversion to decide whether or not launching the experiment. The hypothesis was that this feature might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial due to lack of time—without significantly reducing the number of students to continue past the free trial and eventually complete the course.

If this hypothesis held true, there will be a decrease in gross conversion, and net conversion won't be negatively impacted at all. Thus Udacity could improve coaches' capacity to support students who are likely to complete the course without hurting its revenue. In this case, it is ok to launch the change.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics.
(These should be the answers from the "Calculating standard deviation" quiz.)

Given a sample size of 5000 cookies visiting the course overview page:

Gross Conversion SD = sqrt(0.20625*(1-0.20625)/(5000*0.08)) = **0.0202**
Net Conversion SD = sqrt(0.1093*(1-0.1093)/(5000*0.08)) = **0.0156**

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to 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.

The experiment is cookie-based diversion, and both gross conversion and net conversion use # of unique cookies as denominator. This means the unit of analysis and

the unit of diversion are the same, so the analytical estimate should be comparable to the empirical variability.

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. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

I won't use the Bonferroni correction during the analysis phase. First I want to set hypotheses for both metrics respectively and check whether or not both metrics meet my expectation significantly to make a decision, this already makes it less likely to detect false positives. Second the metrics are highly correlated which means it will be too conservative to use Bonferroni correction, if further correction is applied, this would end up making it too hard to detect the true positives.

The number of pageviews I need to power my experiment appropriately is **685,325.**

# of pageviews for each metric has been calculated separately and choose the one which is larger.

- Pageviews needed for gross conversion:
  20.625% base conversion rate, 1% min d. Samples needed: 25,835
  25835 / 0.08 * 2 = **645,876**
- Pageviews needed for net conversion:
  10.93125% base conversion rate, 0.75% min d. Samples needed: 27,413
  27413 / 0.08 * 2 = **685,325**

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. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

I'd like to divert 100% of traffic to this experiment. With daily traffic of 40,000 pageviews, it means it would take 685,325/40,000 = 17.13 $\approx$ 18 days.

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

I don't think there is any risk associated with the experiment. Firstly, it doesn't affect students who are currently enrolled and paying for the courses. Secondly, it doesn't change any of the content in the course overview page, users who see this change have already shown their big interest by click through the "start free trial" button. By adding the change to inform students that a reasonable amount of studying time is necessary (5

hours per week) is a nice-to-have step. However, I do recommend divert a small fraction of traffic first to check whether there is any bug for setting up the experiment.

Experiment Analysis

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. (These should be the answers from the "Sanity Checks" quiz.)

Number of cookies:
Confidence Interval [0.4988, 0.5012], observed: 0.5006

Calculation example:
Total # of pageviews in control group (Nctr): 345,543
Total # of pageviews in experiment group (Nexp): 344,660
Total # of pageview: 690,203
Probability of a cookie be diverted to any of the two groups: 0.5
SE = sqrt(0.5*(1-0.5)*(1/345543+1/344660) = 0.0006018
Margin of error (m) = SE * 1.96 = 0.0011796
Confidence Interval = [0.5-m, 0.5+m] = [0.4988,0.5012]
Observed Value = 344,660/690,203 = 0.5006

Number of clicks (Pass):
Confidence Interval [0.4959, 0.5041], observed: 0.5005

Calculation details:
Total # of clicks in control group (Nctr): 28,378
Total # of clicks in experiment group (Nexp): 28,323
Total # of clicks: 56,701
Probability of a cookie be diverted to any of the two groups: 0.5
SE = sqrt(0.5*(1-0.5)*(1/28378+1/28323) = 0.0021
Margin of error (m) = SE * 1.96 = 0.0041
Confidence Interval = [0.5-m, 0.5+m] = [0.4959,0.5041]
Observed Value = 28378/56701 = 0.5005

Click-through-probability on “Start free trial” (Pass):
Confidence Interval [-0.0013, 0.0013], observed: 0.0001

Calculation details:
Control CTP: Pcontrol = 28,378/345,543 = 0.082126
Experiment CTP: Pexp = 28,325/344,660 = 0.082182
Ppool = (28378+28325) / (345543+344660) = 0.082154
SEpool = sqrt(0.082154*(1-0.082154)*(1/345543+1/344660)) = 0.00066
Margin of error (m) = SEpool * 1.96 = 0.00130
Confidence Interval = [0-m, 0+m] = [-0.0013, 0.0013]
Observed Value = Pexp – Pcontrol = 0.082182 – 0.082126 = 0.0001

All the invariant metrics pass sanity check.

Result Analysis

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. (These should be the answers from the "Effect Size Tests" quiz.)

Gross conversion:
Confidence Interval [-0.0291, -0.0120]
Statistical Significance: Yes
Practical Significance: Yes

	Control	Experiment
Clicks #	17,293	17,260
Enrollments #	3,785	3,423
Gross conversion	0.2189	0.1983

Calculation details:
Ppool = (3785+3423) / (17293+17260) = 0.2086
SEpool = sqrt(0.2086*(1-0.2086)*(1/17293+1/17260)) = 0.0044
Margin of error (m) = SEpool * 1.96 = 0.00857
d (hat) = 0.1983 – 0.2189 = -0.02055
Confidence Interval = [-0.02055-0.00857, -0.0206+0.0086] = [-0.0291, -0.0120]
d_{min} = 0.01
-Statistically significant (CI doesn't contain zero)
-Practically significant (CI doesn't contain dmin value)

Net conversion:
Confidence Interval [-0.0116, 0.0019]
Statistical Significance: No
Practical Significance: No

	Control	Experiment
Clicks #	17,293	17,260
Payments #	2,033	1,945
Net conversion	0.1176	0.1127

Calculation details:
Ppool = (2033+1945) / (17293+17260) = 0.1151
SEpool = sqrt(0.1151*(1-0.1151)*(1/17293+1/17260)) = 0.00343
Margin of error (m) = SEpool * 1.96 = 0.00673
d (hat) = 0.1127 – 0.1176 = -0.00487
Confidence Interval = [-0.00487-0.0067, -0.00487+0.00673] = [-0.0116, 0.0019]
d_{min} = 0.0075
- not statistically significant (CI contains zero)
- not practically significant (CI contain d_{min} = +/- 0.0075)

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

The null hypothesis is: there is no significant difference in gross conversion/net conversion between two groups.

Gross conversion:
# of trials: 23
# of success: 19 (experiment gross conversion < control gross conversion)
p-value: 0.0026
The one-tail p-value = 0.0013 < 0.05(alpha), so we reject the null hypothesis and accept the alternative hypothesis that the gross conversions for experiment is significantly lower than the gross conversion of control group.

Net conversion:
# of trials: 23
# of success: 13 (experiment net conversion < control net conversion)
p-value: 0.6776
For a two-tail p-value = 0.6776 > 0.05 (alpha), so we accept the null hypothesis that the there is no significant difference between the net conversions between control & experiment groups.

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 didn’t use the Bonferroni correction. The Bonferroni correction is commonly used to adjust p value when making multiple comparisons, because as the number of tests increases, so does the likelihood of false positives. It is very common to use the correction when there is one universal null hypothesis (Ho) for all tests, and any metric with statistical significance would lead to reject the Ho and trigger the launch.

However, in the Udacity experiment, I choose to assess the statistical significance of individual test respectively and check whether or not both metrics are satisfied to trigger a launch. This approach already makes it less likely to detect false positives and even with a single false positive, it can’t govern a decision.

In addition, gross conversion and net conversion are highly correlated metrics, which means if further correction is applied, it would end up being too hard to detect the true positives and result in increasing likelihood of false negatives. Lowering the power of a test is not what we want, because a single false negative from any of the two tests would not trigger the launch. It is obviously not favored to have such a tradeoff by increasing false negatives in this experiment.

In conclusion, Bonferroni correction is not necessary in this experiment.

There aren't any discrepancies between the effect size hypothesis tests and the sign test. They are aligned in concluding that the change (screener) will significantly reduce the gross conversion but keep the net conversion unchanged statistically.

Recommendation

Make a recommendation and briefly describe your reasoning.

I won't launch the change right away. Though the feature meets our expectation to reduce gross conversion significantly and optimize the utilization of coach capacity to support dedicated students. However, when taking a closer look at the confidence interval of net conversion: [-0.0116, 0.0019], the lower bound of the CI (-0.0116) is smaller than the negative practical significance d_{min}= -0.0075. It means the drop of the net conversion may possibly reach a level that we do care about. And it's risky to launch the change since it may hurt our revenue.

I recommend:
First, doing a retrospective analysis, dig deeper into the average hours spent between two groups on Udacity by those students who dropped out after the free trial and by those who stay enrolled. In this way, we could have a better understanding of whether or not time commit is a critical issue. Also we can check if the screener is effective to encourage students to spend more time.

Second, it might be helpful to extend the experiment with larger sample size to see whether we could narrow down the CI of net conversion to exclude the practical significance.

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.

One possible scenario is that some students kept their time commitment and did spend more than 5+ hours to study but still found themselves not progressing as expected. So they felt very frustrated and thought the courses were just too difficult for them even they invested the necessary amount of time.

My hypothesis would be: encouraging students who lag behind the expected progress in the first week to seek help through Udacity services such as coach appointment, ask questions in forum or slack groups etc. would help them move forward, and consequently increase the retention rate. The change is to pop up a screener which can

encourage those students to ask questions in forum or book 1:1 appointment every time when they start viewing a course.

We need to do a filtering first, only select the user-ids which didn't reach our expected progress on the 7th day. The unit of diversion is based on user-ids.

The invariant metric is # of user-ids as well, since we want the equal number of lagged behind students to be diverted into two groups based on their user-id.

The evaluation metric 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. We will test whether the reminder feature could increase the retention rate of slow progressed students.