Design an A/B Test

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.)

- Invariant metrics: number of cookies, number of clicks, click-through-probability
- Evaluation metrics: gross conversion, retention, net conversion

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.

- Number of cookies: That is, the number of unique cookies to visit the course overview page. This number should not be affected by the A/B test due to that it occurs before the test stage. It should have similar distribution for both control and experiment groups. Therefore, it is invariant for both control and experiment groups, and cannot tell the difference between the two groups. That's why it's chosen as the invariant metrics, rather than the evaluation metrics.
- Number of user-ids: That is, the number of users who enroll in the free trail. This is not a suitable invariant metrics. Because the design of experiment may affect the number of users who enroll in the free trial. As a result, control and experiment groups may give different distribution or values for the number of user-ids. For the same reason, it is usable as evaluation metric. Namely, the experiment will reduce the number of students to continue past the free trail. It is not the best metric as it is not normalized.
- Number of clicks: that is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). Because this occurs before the trigger of the "Start free trail" button, the experiment should not have any effect on this number. Therefore, it is suitable to be invariant metrics, and not suitable for evaluation metrics.
- Click-through-probability: That is, number of unique cookies to click the "Start free trail" button divided by number of unique cookies to view the course overview page. This is a suitable invariant metrics. Because the click of "Start free trail" button is before the free trail screener, the experiment should not have any effect on this number. That is, for both control and experiments groups, this number should be similar. Therefore, it is a good invariant metrics and not suitable for evaluation metrics.
- Gross conversion: This is, number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trail" button. This is a suitable evaluation metrics. If the experiment really worked as expected, the experiment group should have a lower gross conversion than the control group due to that the experiment is expected to reduce the number of frustrated students who left the free trial because they didn't have enough time. Because the number of unique cookies to click the "Start free trail" button is invariant for the control and experiment groups, while the

number of user-ids to complete checkout and enroll in the free trail is different between the control and experiment group, this is a suitable metrics to evaluate the test.

- Retention: That is, number of user-ids to remain enrolled past the14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. This is also a suitable evaluation metrics. If the experiment worked as expected, it should decrease the number of user-ids to complete checkout and increase the number of user-ids to remain enrolled past the 14-day boundary, because it is expected to reduce the number of students who don't have enough time and would leave after the free trail. Therefore, this is a good metrics to evaluate the test.
- Net conversion: That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trail" button. This is also a suitable evaluation metrics. The number of unique cookies to click the "Start free trail" button should be similar for both control and experiment group, because the click occurs before the free trail screener. If the experiment worked as expected, the number of user-ids to remain enrolled past the 14-day boundary should increase, because it is expected to reduce the number of students who cannot continue after the free trail. Therefore, most students clicking the free trail button should continue to learn after the free trail if the experiment design works.

The evaluation metrics is expected to have lower gross conversion, higher retention and no decrease net conversion in order to launch the experiment.

Measuring Standard Deviation

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

Baseline Values

Unique cookies to view page per day:	40000
Unique cookies to click "Start free trial" per day:	3200
Enrollments per day:	660
Click-through-probability on "Start free trial":	0.08
Probability of enrolling, given click:	0.20625
Probability of payment, given enroll:	0.53
Probability of payment, given click	0.1093125

Given a sample size of 5000 cookies visiting the course overview page, calculate the estimate of standard deviation of evaluation metrics.

Gross conversion: This is, number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trail" button.

SD(Gross conversion) = $\sqrt{0.20625(1-0.20625)/(5000\times\frac{3200}{40000})} \approx 0.0202$

Retention: That is, number of user-ids to remain enrolled past the14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout.

SD(Retention) = $\sqrt{0.53(1-0.53)/(5000\times\frac{660}{40000})} \approx 0.0549$

Net conversion: That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the “Start free trail” button.

SD(Net conversion) = $\sqrt{0.1093125(1-0.1093125)/(5000\times\frac{3200}{40000})} \approx 0.0156$

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.

Because both gross and net conversions have denominators as the unique number of cookies to click the “Start free trail” button, which is the unit of diversion. That is, the unit of analysis is equal to the unit of diversion. So gross and net conversions would be comparable to the empirical variability. For the retention, the denominator is number of user-ids to complete checkout, which is not equal to the unit of diversion. So its analytic estimate would be different from the empirical one.

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 will not use the Bonferroni correlation in the analysis phase. Because Bonferroni correction is used for adjusting probability values due to the increased risk of Type I error when making multiple statistical tests. That is, it is suitable for the case that we were to launch the experiment when any metric would match our expectations. Here, we would launch the experiment if all metrics match our expectations. Therefore, the Bonferroni correlation is not applied in this analysis.

Use alpha=0.05 and beta=0.2, the sample size needed are:

- Gross conversion: 25,835
- Retention: 39,115
- Net conversion: 27,413

We need to consider the largest value, which is the one for retention. Note that, the probability of payment given enroll is 0.53, the probability of enroll given click is 0.20625, and the click-through-probability is 0.08. Therefore, 39115/0.20625/0.08=2370606. We need to double this value because we need both control and experiment groups. So the number of pageviews would be 2370606*2 = 4,741,212. Given daily unique cookies of 40000 and 100% diversion, it would take about 119 days, which is too long.

If we do not include retention, the largest value of sample size would be 27,413. Consider the click-through probability of 0.08, the pageviews would be 27413/0.08*2=685,325. Given daily unique cookies of 40000 and 100% diversion, it only takes 18 days, which is acceptable. So I would go with the 685,325 pageviews with evaluation metrics of gross and net conversions.

Note: the sample size is calculated using this website. Baseline conversion is the corresponding probability in the baseline table, and minimum detectable effect is the d_{min}.

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.)

Use same reasoning as above, given daily unique cookies of 40000 and 100% diversion, it only takes 18 days with evaluation metrics of gross and net conversions. That is, $685{,}325/40000 \approx 17.13$.

For safety purpose, we can divert 60% of traffic to this experiment. Then it would take 29 days to run the experiment. That is, $685{,}325/40000/0.6 \approx 28.56$.

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

Due to the property of this experiment, it is not very risky in terms of revenue. Considering we only use 60% of the traffic and it only takes about one month, it should be safe for Udacity.

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.)

For a count (such as the first three metrics), you should calculate a confidence interval around the fraction of events you expect to be assigned to the control group, and the observed value should be the actual fraction that was assigned to the control group.

For any other type of metric, (such as the last four metrics), you should construct a confidence interval for a difference in proportions, then check whether the difference between group values falls within that confidence level.

	Lower bound	Upper bound	Observed	Passes
Number of cookies	0.4988	0.5012	0.5006	passed
Number of clicks	0.4959	0.5041	0.5005	passed
Click-through-probability	0.0812	0.0830	0.0822	passed

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.**

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.)

A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)

	Lower bound	Upper bound	Statistical significance	Practical significance
Gross conversion	-0.0291	-0.0120	Yes	Yes
Net conversion	-0.0116	0.0019	No	No

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.)

Gross conversion: p-value is 0.0026, less than 0.05, statistically significant.
Net conversion: p-value is 0.6776, greater than 0.05, statistically not significant.

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 did not use the Bonferroni correction. Because Bonferroni correction is used for adjusting probability values due to the increased risk of Type I error when making multiple statistical tests. That is, it is suitable for the case that we were to launch the experiment when any metric would match our expectations. Here, we would launch the experiment if all metrics match our expectations. Therefore, the Bonferroni correlation is not applied in this analysis.

Both the effect size hypothesis tests and the sign tests show that the change has significant influence on the gross conversion, but not on the net conversion.

Recommendation

Make a recommendation and briefly describe your reasoning.

I would not recommend the change. The hypothesis states that "…without significantly reducing the number of students to continue past the free trial and eventually complete the course". Therefore, the hypothesis asks the net conversion not decrease. The change of the net conversion indeed is not significant, but the confidence interval does include the negative of the practical significance boundary. That is, the net conversion could go down by an amount that would matter to the business. So it is not acceptable according to our hypothesis. Therefore, the change is not recommended.

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.

I suspect that some students may still choose the "Start free trail" even they are suggested to access the course materials for free. Therefore, based on the previous experiment, I would add another message window after students click "Start free trail" button and are informed to better access the course materials for free. That is, if the students clicked "start free trail", 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. By far, this is similar as the previous experiment. After the message appears, if the students didn't choose the "free trail button", everything will go as usual. But if the students still clicked the "start free trail" button after the suggestion message appeared, we can make a change that the second message would appear and indicate that the students may just want to access course materials for free due to the limited committed time.

My hypothesis would be the second suggestion message could effectively reduce the number of students that enrolled the free trail and cannot continue to finish the program. I suppose the first message may not bring the students' attention, while a second message can make the students realize the situation effectively.

We can use the same metrics as previous experiments. For evaluation, that would be gross and net conversions. The unit of division is same as before, a cookie. For invariant metrics, that would be number of clicks, number of unique cookies, and click-through-probability. The reasoning is same as before. Because we only care about whether the change can reduce the number of frustrated students who left the free trial because they didn't have enough time—without significantly reduce the number of students to continue past the free trial and eventually complete the course.