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.)

I use Number of cookies, Number of clicks and Click-though-probability as my invariant metrics.
I use Gross conversion and Net conversion as my evaluation metrics.

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: I use this as my invariant metric because I need to make sure both my control version and variation version have the same number of cookies. Otherwise, they may have different results just because they have different number of page views.

Number of user-ids: I don't use this as my invariant metric because this is different across my control version and variation version. In my variation version, some students may give up free trial. I also don't use this as my evaluation metric because this is not what I really care about. I really care about whether my new feature will influence the number of students who continue to use coach support and finish their class with a certificate.

Number of clicks: I use this as my invariant metric because I need to make sure both my control version and variation version have the same number of clicks. Otherwise, they have different results maybe just because they have different number of clicks rather than my new feature in variation version.

Click-through-probability: I use this as my invariant metric because I want to keep this same for my control version and variation version. So I can see whether the difference between the results are from my new features in variation version.

Gross conversion: I use this as my evaluation metric. Since my control version and variation version have the same number of clicks, if their gross conversion are significantly different, maybe the new feature is the reason. If gross conversion in variation version is significantly smaller than the gross conversion in control version, this means our new feature significantly reduces the number of frustrated students who left the free trial because they didn't have enough time. I would expect gross conversion in variation version is significantly smaller than the one in control version.

Retention: I don't use this as my invariant metric or evaluation metric because this doesn't influence my experiment.
Net conversion: I use this as my evaluation metric. Since my control version and variation version have the same number of clicks, if their net conversion are significantly different, maybe the new feature is the reason. If the net conversion in control version and variation version are not significantly different, this means that our new feature doesn't significantly reduce the number of students to continue past the free trial and eventually complete the course. I would expect the net conversion in control version and variation version are not significantly different.

Expectation: To launch the experiment, I would expect gross conversion in variation version is significantly smaller than gross conversion in control version, while net conversion in control version and variation version are not significantly different.

Measuring Standard Deviation

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

Standard deviation of Gross conversion: 0.02023
Standard deviation of Net conversion: 0.01560

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.

For my evaluation metrics Gross conversion and Net conversion, the unit of analyses are cookies as our unit of diversion. So I would presume that the analytical estimate will probably be accurate and we don't need to use empirical variance.

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 correction during my analysis phase.
I will need at least 685325 pageviews for my experiment.

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 would run this experiment on all traffic and I would need about 18 days to run the experiment.

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

Since this is the only experiment Udacity wants to run, and this experiment will take a non-trivial amount of time, I think this experiment its not very risky and we can get the results within 18 days which is not very long time if we run the experiment on all traffic.

Besides, we would expect our new feature will have some negative effect on Gross conversion, so we don't want the experiment lasts too long, because for a long time, students tend to quit. So this will mitigate the effect of our new feature.

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 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.**

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

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.)

Metrics	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.)

Metrics	p-value	Statistical significance
Gross conversion	**0.0026**	**Yes**
Net conversion	**0.6776**	**No**

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 Bonferroni correction. Because in our case, we need both gross conversion and net conversion are significantly different in order to launch the experiment. Both "sign tests" and "effect size tests" are designed for single metrics. So it doesn't make sense to use Bonferroni correction for "sign tests" and "effect size tests".

In Effect Size tests, if I don't use Bonferroni correction, "Gross conversion" is statistical significant and "Net conversion" is not; If I use Bonferroni correction, "Gross conversion" is still statistical significant and "Net conversion" is still not. So there are no discrepancies.

In Sign Tests, if I don't use Bonferroni correction, “Gross conversion” is statistical significant and “Net conversion” is not. If I use Bonferroni correction, both “Gross conversion” and “Net conversion” are not statistical significant. There are discrepancies here. I think it’s because Bonferroni correction is overly conservative in Sign tests.

Recommendation

Make a recommendation and briefly describe your reasoning.

I would not recommend launch this change from our A/B testing. Because this change indeed has some statistical significant negative influences on Gross conversion, in other words, this change will reduce the students who get enrolled in the free trial, which confirms our hypothesis. This change doesn’t have statistical significant influence on Net conversion, in other words, it won’t significantly reduce the number of students who keep remain enrolled in the course. But we can see that the confidence interval for Net conversion can include negative part. That means this feature may reduce the number of students who keep remain enrolled in the course, which is not what we want. So I would not recommend to launch this experiment.

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 want to design an experiment in which if the student clicks “Start free trial”, they will be asked to answer several questions. These questions are what we hope they know before they take the course. If they answer some fraction of these questions, they would be taken through the checkout process as usual. If they don’t perform well in the test, a message would appear indicating that this course would use these knowledge and suggesting they learn these knowledge first. In this way, I hope it can reduce the number of frustrated students who cancel early in the course.

The hypothesis is that this change will let students know whether we have enough knowledge to take this course, thus reducing the number of frustrated students who left the free trial because they don’t have the ability to finish this course. And this change won’t significantly reduce the number of students who eventually complete the course.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

I will use Number of cookies, Number of clicks and Click-through-probability as my invariant metrics because I want to keep these same for both control version and experiment version. I will use Gross conversion and Net conversion as my valuation metrics. Because I want to see whether the change will influence these metrics in my experiment version.