Template Format

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: page views, 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.

Metric	invariant	evaluation
Page views: number of unique cookies that view course page views per day.	Yes: not affected by experiment changes because these occur before the new pop-up.	No: not affected by experiment.
Enrollment: number of user-ids who enroll per day.	No: this is affected by experiment because the pop-up change occurs before user decides to enroll or not.	No: since I'm using gross conversion (which is enrollments normalized as a fraction of clicks), I won't use enrollment counts.
Clicks: number of unique cookies to click on "start free trial" button	Yes: this occurs before the pop-up that we are testing, so it should not be affected by the experiment.	No: not affected by experiment.
Click-through probability: clicks / page views.	Yes: this occurs before the pop-up, so it should be unaffected by the experiment.	No: not affected by experiment.
Gross conversion: enrollment / clicks.	No: enrollment occurs after experiment groups see the new pop-up, so it can be affected by the experiment.	Yes: this ratio tracks what fraction of users get past the enrollment step given that they already reached the "click" step. The pop-up event occurs between these

		two stages, so the experiment affects this ratio. If the pop-up results in a significant decrease in gross conversion, then we should launch the experiment. The pop-up is intended to reduce the number of students who cannot make the time commitment from enrolling.
Retention: payment after free trial / enrollment	No: payment occurs after the pop-up, so control and experiment groups are not the same.	Yes: the enrollment stage is followed by the payment stage, both of which are after the pop-up event. This measures how likely users who sign up are happy enough to stay enrolled and pay. If the pop-up does not decrease retention, we can launch this change.
Net conversion: payment after free trial / clicks to “start free trial”.	No: payment occurs after the pop-up, so control and experiment groups are not the same.	Yes: payment occurs after the pop-up event, so it is affected by the experiment. This tracks how likely users who click on “start free trial” are happy enough to stay enrolled and pay. The pop-up may prevent some students from enrolling, and may reduce payments. We want net conversion not to significantly decrease (i.e. stay the same or even increase) for us 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.)

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.

The unit of diversion is a cookie. If the unit of analysis of the evaluation metric (the unit for the denominator of a ratio) is the same as the unit of diversion, we can expect the analytic standard deviation to be similar to the empirical standard deviation. If not, the analytic standard deviation may be an underestimate.

For the evaluation metrics, the units of analysis are either cookies or user ids.

Evaluation Metric	Standard Deviation (analytic estimate assuming 5000 page views, 400 clicks, and 82.5 enrollments)	Compare to empirical standard deviation
Gross conversion: enrollment / clicks.	0.0202	The unit of analysis is a cookie. This is not the same as the unit of diversion. It would be better to use an empirical estimate of standard deviation.
Retention: payment after free trial / enrollment	0.0549	The unit of analysis is a user-id, which is not the same as the unit of diversion. It would be better to use an empirical estimate of standard deviation.
Net conversion: payment after free trial / clicks to “start free trial”.	0.0156	The unit of analysis is a cookie. This is not the same as the unit of diversion. It would be better to use an empirical estimate of standard deviation.

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 do not use the Bonferroni, because it would make the test for significant difference too conservative.

With 2 metrics, we need 685,325 page views in order to have an overall alpha of 0.05 and a beta of 0.20.

I dropped the retention metric (payments / enrollments) because it requires about 4.7 million page views to get the desired practical significance, alpha and beta. This is because there are relatively few enrollments (it takes 60 page views for 1 enrollment).

Of the two remaining evaluation metrics, gross conversion requires 645,875 page views and net conversion requires 685,325 page views, which is the largest of the two.

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.)

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

We need 685,325 page views to gather data for the experiment and control groups combined. We expect 40,000 page views each day. If we direct 100% of the total web traffic to the experiment, we can complete the experiment in 18 days.

Since this experiment does not collect additional personal information, and does not affect the health or safety of users, it is a fairly low risk 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.)

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.**

I'm assuming that the unit of diversion is unique page views, so subjects are randomly assigned to the control or experiment group based on their unique cookie for the page view. So I expect the invariant metrics of page views, clicks, and click-through probability (clicks / page views) to be randomly assigned to control and experiment groups. So I expect 50% of the observations to be in the control group for page views and clicks. For click-through probability, I expect the difference between experiment and control groups to be zero.

Evaluation Metric	Lower bound	Upper bound	observed	Passes sanity check?
Page views	0.4988	0.5012	0.5006	Yes, because 0.5000 is within this range.
Clicks	0.4959	0.5041	0.5005	Yes, because 0.5000 is within this range.
Click-through probability	-0.0013	0.0013	0.0001	Yes, because the observed difference of 0.0010 is within the 95% confidence interval.

Since all invariant metrics pass their sanity checks, the control and experiment groups appear to be properly randomized.

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.)

Not using Bonferroni correction:

Evaluation Metric	Lower bound	Upper bound	Statistically significant?	Practically significant?
Gross conversion	-0.0291	-0.0120	Yes, entire range is < 0.	Yes, magnitude at least > 0.01
Net conversion	-0.0116	0.0019	No, range includes 0.	No, range overlaps with the range for practical significance (-0.0075 to 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.)

Not using the Bonferroni correction, the alpha for each metric is 0.05.

Evaluation Metric	(Number of days where experiment > control) / (total days)	P-value	Statistically significant?

Gross conversion	4 / 23	0.0026	Yes, p-value is less than alpha of 0.05.
Net conversion	10 / 23	0.6776	No, p-value > 0.05.

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 do not use the Bonferroni, because it would make the test for significant difference too conservative. The Bonferroni correction reduces the alpha, which widens the confidence interval, and makes it less likely for us to conclude that there was a significant difference between experiment and control groups. Since we require both the gross conversion to decrease and the net conversion not to change significantly, we will only launch the change when both metrics satisfy their desired results. Bonferroni correction is used to make the analysis more conservative when the metrics when the significance of at least one metric is enough to launch the change. Since we require both metrics to meet expectations in order to launch, it is better not to reduce the alpha using the correction.

The observed difference for the gross conversion was statistically significant and practically significant. The change appears to reduce the gross conversion. The net conversion is not significantly different, but the lower range does include the practical significance level of -0.0075. This means it's possible that the change did reduce the net conversion, which would keep us from launching this change.

The hypothesis tests and sign tests both agreed that the gross conversion showed a statistically significant difference, whereas the net conversion did not. If the hypothesis tests and sign tests did not agree, it could be due to outliers on a small proportion of days. For instance, if on one day, there is a large difference between groups, but on the other days, the experimental and control groups alternate fairly evenly, a hypothesis test might show a significant difference whereas the sign test would not.

Recommendation

Make a recommendation and briefly describe your reasoning.

Given the results, I would recommend not making the change, since not all of the evaluation metrics satisfy the desired results. Even though gross conversion likely decreases with the change, there is a risk that the net conversion does not stay the same, but decreases instead.

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 am seeking a change that does not reduce enrollments, but still aims to increase retention. For all users who enroll, we could include the message about the minimum time commitment some time after enrollment and before the free trial expires. For instance, 7 days after enrollment, users receive an email that reminds them of the recommended time commitment.

I choose retention (payments / enrollments) as the evaluation metric, since the change occurs after the enrollment stage and before the payment stage.

Since the unit of analysis is user ids from enrollments, I choose the unit of diversion to also be user ids from enrollments.

I choose user ids from enrollments to be an invariant metric, since user ids from enrollments will be randomly assigned to control or experiment groups.

The null hypothesis is that there is no difference in the retention. The alternative hypothesis is that there is a statistically significant difference between the experiment group and control group.

References

Udacity A / B Testing course lectures.