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: number of enrollments can depend on number of users who visit the course page, and number of visits per day won't be perfectly equal between control and experiment groups for each day.
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	Yes: this ratio tracks what fraction of users get past the enrollment step given that

	affected by the experiment.	they already reached the "click" step. The pop-up event occurs between these two stages, so the experiment affects this ratio. Since the pop-up intends to only encourage students to enroll if they can make the minimum time commitment, I expect the pop-up to decrease gross conversion. If the pop-up changes gross conversion by 0.01 or more, this will be a significant change for the business.
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. The pop-up is intended to increase the fraction of enrolled students who are happy enough to pay. If the pop-up changes retention by 0.01 or more, this will be significant enough for the business.
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 would ideally increase the net conversion by setting time commitment

		expectations for enrolled students, such that they are happy enough to pay. It is also possible that the pop-up discourages students from enrolling (and therefore never pay), when they would otherwise have enrolled and also paid. If the pop-up changes net conversion by 0.0075, this is a significant change for the business.

For the evaluation metrics, I want to check if the number of observations required to reach a desired alpha and beta can be gathered in a reasonable number of days. If it takes too long to gather the necessary data, I will need to find other metrics for which the data can be gathered in a reasonable amount of time.

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.

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	Since this is a ratio, the distribution of the data may not be normal, and the actual empirical SD

		may be larger.
Retention: payment after free trial / enrollment	0.0549	Since this is a ratio, the distribution of the data may not be normal, and the actual empirical SD may be larger.
Net conversion: payment after free trial / clicks to “start free trial”.	0.0156	Since this is a ratio, the distribution of the data may not be normal, and the actual empirical SD may be larger.

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 use the Bonferroni correction to choose an individual alpha for each metric as the overall alpha divided by number of metrics. With 2 metrics and an overall alpha of 0.05, I need 791,500 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 6 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 746,100 page views and net conversion requires 791,500 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?

If I want to run the experiment for 40 days or less, and need 791,500 page views for the experiment and control groups, then I need 50% of the total traffic to be diverted towards the experiment and control groups. This would take 40 days. If I divert less than 50%, the experiment will take too long to run before we had a result. Since about half of this 50% is the control group, the remaining half (25% of the total traffic) includes the experimental group. This is still somewhat risky. If there is a significant bug, this will affect 25% of the total web traffic.

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	Passes sanity check?
Page views	0.4988	0.5012	Yes, because 0.50 is within this range.
Clicks	0.4959	0.5041	Yes, because 0.50 is within this range.
Click-through probability	-0.0013	0.0013	Yes, because the observed difference of 0.001 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.)
Using Bonferroni correction:

Evaluation Metric	Lower bound	Upper bound	Statistically significant?	Practically significant?
Gross conversion	-0.0303	-0.0108	Yes, entire range is < 0.	Yes, magnitude at least > 0.01
Net conversion	-0.0126	0.0028	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.)

Using the Bonferroni correction, the individual alpha for 2 metrics and overall alpha of 0.05 is 0.025.

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 0.025.
Net conversion	10 / 23	0.6776	No, p-value > 0.025.

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 used the Bonferroni correction to keep the overall alpha at 0.05. If I did not, then the overall alpha for the two metrics would be 0.0975 (which is 1 - (1- 0.05)^2). If I had not used the correction, I could have reduced the number of page views needed from 791,500 to 685,325, and reduced the number of days to run the experiment from 40 days to 35 days.

The observed difference for the gross conversion was large enough to be statistically significant and practically significant with both the Bonferroni correction or without. However, with the correction, the differences were closer to being not significantly different from zero, and closer to not being practically significant.

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 the change reduces gross conversion and therefore reduces revenue. We do not know if the pop-up would increase retention, because this metric requires so many page views (5.4 million) that it would take too long to run an experiment (about 9 months).

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 net conversion (payments / page views) as the evaluation metric, since the number of page views needed is small enough that I can run the experiment in 40 days or less. Ideally, I would want to use retention, but the number of page views required makes the experiment duration nearly 9 months. I would no longer use gross conversion as an evaluation metric, enrollments occur before the change that the experiment introduces.

Since net conversion unit of analysis is number of unique page views per day, I choose the unit of diversion to be the same. This way, the analytic standard deviation is more likely to be close to the empirical standard deviation (instead of being an under-estimate).

The null hypothesis is that there is no difference in the net conversion. 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.