Experiment Design

**Metric Choice**

List which metrics you will use as invariant metrics and evaluation metrics here.

**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 to launch the experiment.

**Invariant metrics:**

Number of cookies:

A good population size metric. Cookies should be independent of the experiment change and should be randomly assigned to each group.

Number of clicks:

A good population size metric. Click is an event, it should be independent of the experiment change and should be randomly assigned to each group.

Click-through-probability:

A good invariant metric. It happened before the change triggered and should not be affected by the experiment.

**Evaluation metrics:**

Gross conversion:

A good evaluation metric. Since the one of the experiment targets was to find whether change would reduce students with not enough time left the free trail, and the number of cookies complete enrollments divided by the number of clicks would be a good evaluation. The gross conversion should decrease after the change.

Retention:

A good evaluation metric. Retention has all the characteristics of an evaluation metric. It is normalized by enrollments and measures the probability to pay given enrollment. This normalized probability makes for a ready comparison between experiment and control groups and is the most direct way of answering the second part of our hypothesis. However, if we take retention as evaluate metrics, it need 2,370,010\*2 = 4,740,020 samples to get the experiment, if we put all the traffic on the experiment, it need 4,740,020/ 40,000 = 118.5 days. It took too long to get the result and should be rejected.

Net conversion:

A good evaluation metric. As mention above, the number of cookies made a payment divided by the number of clicks would be a good evaluation for whether change would not significantly reduce the number of students continue past the free trial and eventually complete the course. The net conversion should remain the same after the change.

**Neither Invariant nor evaluation:**

Number of User ids:

Not a good invariant metrics. The number of users enrolled in the free trial is dependent on the experiment change. Besides, user ids could be slight different in control and environment group since the target unit of diversion is cookie.

Not a good evaluation metrics.  The number of user IDs is usable as evaluation metric because it would track the first part of the hypothesis, namely whether we will reduce the number of students to continue past the free trial, but since it isn't normalized, the gross conversion is better choice. In this case, I did not to use it.

I would pick the **Gross conversion** and **Net conversion as my evaluation metrics, each of them show one part of the experiment goal. If the Gross conversion decrease after the experiment change, it showed the change** reduce students with not enough time left the free trail**. If the Net conversion remain the same, it can show us** whether the change would not significantly reduce the number of students continue past the free trial and eventually complete the course**.**

**Measuring Standard Deviation**

List the standard deviation of each of your evaluation metrics.

**Gross conversion: 0.0202**

**Net conversion: 0.0156**

Do you expect the analytic estimates to be accurate? That is, for which metrics, if any, would you want to collect an empirical estimate of the variability if you had time?

To determine whether we should use analytic or empirical variability we compare the unit of diversion to the unit of analysis. If these match, we are confident that the two types of variability will yield similar values.

The denominator for both gross and net conversion is cookies therefore is the same as the unit of analysis. The courses learning was a continues process and there should not be too much people revisit and click the “start free trial” again, which suggest the analytic variability is close to the empirical one. If we had enough time, collecting more data to test that would be better.

**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.

No use of Bonferroni correction.

679284 page views would be needed.

Require sized of one group for Gross is 321237, for Net is 339642. Since we need consider both control and experiment groups, the result is the double of the highest, it should be 339642\*2 = 679284.

**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 take 80% of the traffic for the experiment. Which indicated it need 21 days for the experiment. The experiment contains low risk and run too long might increase expenses and interfere other experiments.

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

From risky aspect, the experiment would be risky for clients if either could harm someone or collect sensitive information. This experiment did none of above, in other words, it is not risky.

From duration aspect, take 50% of traffic would take approximately a month, and client would be fine. If we take less traffic and more days, it might arise problem for duration.

In this case, I would use the current fraction for 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.

All the invariant metrics passed.

The Number of cookies’ margin is 0.4988 to 0.5011, the observe value is 0.5006.

The Number of clicks’ margin is 0.4959 to 0.5041, the observe value is 0.5005.

The CTR rates’ difference margin is -0.0013 to 0.0013, the observe value is -0.0001.

**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.

The Gross conversion is both statistically and practically significant, the confidence interval is from -0.120 to -0.0291, since 0 is not in the interval, and the absolute boundary is 0.0291, larger than the d-min 0.01.

The Net conversion is neither statistically nor practically significant, the confidence interval is from -0.0116 to 0.0019. since 0 is in the interval and the boundary is within the d-min.

**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.

The gross conversion is statistically significant; the sign test result is 0.0026.

The net conversion is not statistically significant; the sign test result is 0.6776.

**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.

No use of Bonferroni correction.

For multiple metrics, if we would need only one of them to meet our criteria to launch, we would face the risk that a single metric could meet criteria by pure chance, by mistake. Bonferroni correction would decrease type I error (false positive) as mentioned above.

However, since we would need all the metrics to meet our criteria to launch, we would face the risk that a single metric could not meet criteria by pure chance, by mistake. We call that type II error (false negative). In this case, Bonferroni correction might even make it worse, since it potentially increased the type II error or made no difference. Therefore, we won’t use Bonferroni correction in this problem.

The effect size test and sign test result met.

**Recommendation**

Make a recommendation and briefly describe your reasoning.

The experiment wanted to find whether the change decreased the students left the free trial without enough time, while not decrease those use the free trial and past the course.

The Gross conversion showed us numbers of students try free trial decreased as we expected, but the Net conversion showed us numbers of students actual paid potentially decreased when we hope it keep the same. We can find the net conversion confident interval lower boundary is -0.0116 while the d-min is -0.0075. Due to these, my recommendation is not launch the experiments.

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.

To reduce the number of frustrated students, we need to find if there are any kind of incentive after enrollment for students to keep payment.

Students might cancel the enrollment because they thought the courses was useless, or the courses did not meet the course description, or they cannot get enough support during the enrollment. Take the tutorial support as an example, we can test if the students did not get a response within hours, or they find it is hard to get the point from the forums, or they feedback a low rate for tutors’ suggestion.

I would take the response time as my follow-up experiment.

**Null hypothesis**:

Reduce response time by half from tutor to users would reduce the early cancellation.

**Unit of Diversion:**

It should be user-id. The change take place after users enrolled in the courses.

**Invariant metrics**:

User-ids: A good population size metric. It should be independent of the experiment change and should be randomly assigned to each group.

Click-through-probability: A good invariant metric. It happened before the change triggered and should not be affected by the experiment.

Gross conversion: A good invariant metric. It happened before the change triggered and should not be affected by the experiment.

**Evaluation metrics**:

It should be retention. The change should increase the payment while keep the same enrollment. It would indicate the change is successful.