Template Format

This template can be used to organize your answers to the final project. Items that should be copied from your answers to the quizzes should be given in blue.

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, 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.

-Each of the invariant metrics: number of cookies; number of clicks; and click through probability were chosen to be invariant metrics because they should not show any sort of material difference between the control and experiment group. In other words, in execution of the experiment, these metrics should not be materially different in either group of data. If we see an impact in either group of data in these metrics, it is a tip off to let us know something might have gone wrong and we need to go back to our experiment and reevaluate.

The evaluation metrics used in the experiment are net conversion and gross conversion. We use these as the evaluation metrics because we do expect there to be some sort of difference between the experiment group and the control group, either positively or negatively. It might turn out there is no effect, but our hypothesis is testing whether there will be an effect on those two metrics.

In order to launch the experiment, we are looking for the evaluation metrics to show statistical and practical significance. For gross conversion, the practical significance minimum is 1%, with statistical significance of 5%. For net conversion, the practical significance minimum is .75% with a statistical significance of 5%. If our 95% confidence intervals for our evaluation metrics do not include the practical significance level, our practical significance requirement to launch the experiment will have been met. If the 95% confidence intervals for our evaluation metrics do not include 0, the statistical significance threshold to launch the experiment will have been met.

We are looking to reduce the gross conversion, while not adversely impacting net conversion. We want our gross conversion to hit a significance minimum of 1%. We want this number to be

negative, which would show a decrease in the number of students enrolling because of the experiment's intervention. Our hypothesis contends this would decrease the number of frustrated students who enroll, while not adversely affecting net conversion, or the ratio of students who end up making a payment after the 14-day free trial. Our practical significance is .75% for net conversion. We would not want to see a percentage decrease that high in net conversion, along with a corresponding minimum 1% decrease in gross conversion in order to ponder going live with the experiment.

Neither retention, nor number of user ids were used in the experiment. The reason for this is how long it would take to run the experiment. The general guideline is you don't want an experiment to take longer than 30 days. If retention was one of the evaluation metrics, we would need a much higher number of pageviews to generate enrollment figures large enough to power an experiment that could be executed in less than 30 days. We have no guarantee we could get a significantly higher number of pageviews in order for that to happen. Because of that, retention is not a metric we can use as an evaluation metric.

Number of user IDs is not used as an evaluation metric because it is not normalized. It could help us in determining the first part of the experiment, which is whether we reduce the number of enrollments. Therefore, it could be used as an evaluation metric, but since it is not normalized, gross conversion is a better metric to use.

Measuring Standard Deviation

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

Gross Conversion: .0202

Net Conversion: .0156

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to 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.

-I think the analytic estimate would be comparable to the empirical variability due to the unit of diversion and unit of analysis being the same (cookie). Because of this, there is little reason to calculate the empirical variability, when we can safely assume it will look almost the same as the analytical estimate.

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 your experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

I will not use the Bonferroni correction. The number of pageviews needed to power the experiment appropriately is 685,300.

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 divert 62.5% of Udacity's traffic over a 28 day period to run the experient.

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

-In order to get enough pageviews in a good time frame, we would need to divert a minimum of 62.5% of the traffic. The reason for diverting this much of the traffic is we would like to get the amount of pageviews needed to power the experiment within 30 days. This limits the risk of running into seasonal, or other, factors which could materially change our data. This experiment is not risky for Udacity given the fact we are not collecting sensitive information, such as health or financial data. If we were collecting such data, it might change our opinion on how much of our traffic we want to divert to our experiment group.

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.)

Number of cookies:

LB: .4988, UB: .5012, OBS: .5006, passes check

Number of clicks:

LB: .4959, UB: .5041, OBS: .5005, passes check

Click through probability:

LB: -.0013, UB: .0013, OBS: .0001, passes check

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.)

Gross Conversion:

LB: -.0291, UB: -.0120, both statistically and practically significant

Net Conversion:

LB: -.0116, UB: .0018, Neither statistically or practically significant

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: .0026, Is statistically significant

Net Conversion:

P-Value: .6776, Not statistically 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 I did not want to get an overly conservative estimate of the effect of the experiment. We will not launch the experiment if we do not see statistical and practical significance in both conversion rates(we are hoping to see a decrease in gross conversion and no effect on net conversion). The Bonferroni correction does not assist in this situation. When we will only launch the experiment if all metrics meet significance, the error type we face is type II (false negative, where we have failed to reject the null hypothesis when we should have). In other words, we want all of our metrics to match the expected outcomes, but at least one of the metrics does not meet our expectations and we do not acknowledge this happening. Due to launching only when all metrics meet significance, we are already pretty well protected from launching something we shouldn't have. There isn't any need to add the Bonferroni Correction.

However, if we were going to launch this experiment if any of our evaluation metrics met significance, when they were not actually significant, the Bonferroni Correction would prove useful, because it would prevent a false positive, or type I error. Reducing the p-value would reduce the chance our metrics meet statistical significance, therefore reducing the chance we launch the experiment when we shouldn't have. Because the risk is much greater of misjudging the significance of one metric as opposed to all of them, using the Bonferroni Correction would be appropriate in this scenario.

There were not any discrepancies between the effect size and the sign tests. Gross conversion was statistically significant in both tests, while the net conversion was not significant in both.

Recommendation

Make a recommendation and briefly describe your reasoning.

-My recommendation would be to not launch the experiment. We did observe the expected result for gross conversion, which was a small decline that did meet both our practical significance and statistical significance levels. We had a practical significance level of 1%. We did reach this result, but in the negative direction, which is what we wanted. We were hoping our experimental intervention would have the effect of reducing gross conversion, and in doing so reduce the number of frustrated students to enter the free trial period. The result for net conversion was not what we were hoping for. We did not meet our practical significance level or our statistical significance level. However, the adverse practical significance level is within our confidence interval. Therefore, there is a chance we could be incurring a significant risk if we did actually launch the experiment. We are able to declare this experiment does accomplish what we want in the first part of the hypothesis, but there is a chance it accomplishes the opposite of what we want in the second part of the hypothesis. Due to the possibility of this risk, and the

results not meeting the expectations for both parts of our hypothesis, I recommend the experiment should not be launched.

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.

Follow up Experiment: The follow up experiment I would run involves the course material. When students click on a course overview page, I would provide two possible options: 1) a couple paragraph overview of the course, 2) a full detailed syllabus breakdown of what is covered in each unit of the course, in addition to expected learning outcomes. To create a control and experiment group, I would funnel cookies that click on a course overview page to one of the two options. The control group would see what is written on the course overview page right now. The experiment group would see the full detailed syllabus breakdown. The success of the intervention would be determined in the same way as the initial experiment: fewer clicks on the start free trial button, which lead to enrollment, but the same or a slightly higher percentage of those enrolling completing the free trial and making a payment.

Hypothesis: By putting a detailed course syllabus on each course overview page, instead of a couple paragraph overview, students who see the detailed course syllabus, and are not fully interested in the course material, would be more likely to get frustrated by the course and are less inclined to enroll. Serious students will see the detailed course syllabus on the course overview page and be more inclined to not only enroll, but also make a payment. Gross conversion will decrease in a practically and statistically significant way, but net conversion will not be affected, or slightly increase, in a practically and statistically significant way.

Metrics to Measure: Invariant metrics (number of cookies, number of clicks, click through probability). Evaluation metrics (gross conversion, net conversion). Unit of diversion (cookie).

My reasoning for using the same metrics is we are trying to solve the same problem as the above experiment: “How can we reduce the number of frustrated students who cancel early in the course without significantly reducing the number of students to continue past the free trial and eventually complete the course?” Therefore, it makes sense to use the same metrics to evaluate the success of the follow up experiment intervention. Also, using retention as an evaluation metric would simply take too long for the experiment to complete, which is the same reason we do not use it in the initial experiment.

-The follow up experiment I would run is changing the number of hours specified in the pop-up box. I would try different values for that number. My hypothesis would be if we move the number higher, the net conversion would go up, but gross conversion would go down and vice versa for moving the number lower. This would lead to more students enrolling who truly want to develop their careers. The invariant metrics I would measure are the same as the experiment conducted for this project: number of cookies, number of clicks, and click through probability. The evaluation metrics I would use are gross conversion and net conversion. My unit of diversion would be a cookie. My metrics would be the same as the above experiment because it is largely related to it, and we are testing a different way of figuring out how to separate the students who will truly be dedicated to the program from the those whose initial interest does not match the necessary dedication to complete the program.