Data Analyst Nanodegree
Project 7 - A/B Testing
Thomas Lindstrom-Vautrin

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, 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.

Number of Cookies - This makes an ideal invariant metric because in this situation we collect the cookies before the crucial point in the experiment meaning that this should be independent of the experiment. It makes a poor evaluation metric precisely because it is not likely to change since it is determined before the crucial part of the experiment.

Number of User IDs - This cannot be used as an invariant metric because user ids are collected after the person signs up for the service and this may be affected by the experiment that we run. It can be used as an evaluation metric because it is likely to be different in the experiment and control groups. We expect to have fewer user ids in the experiment group since our warning might discourage people from signing up.

Number of Clicks - This makes an ideal invariant metric since the click occurs before the critical moment in the experiment and should not depend on the experiment. It makes a poor evaluation metric precisely because it is not likely to change since it is determined before the crucial part of the experiment.

Click-Through Probability - Again since the actual clicking occurs before the crucial moment in the experiment the click through probability should be unaffected by whether it is in the control or experimental group so this makes for a good invariant metric and for the same reason it makes a poor evaluation metric.

Gross Conversion - This makes for a good evaluation metric because we can judge if the experiment affects the probability of enrollment given the click. Perhaps after seeing the experimental message people are less likely to enroll. It makes a poor invariant metric since we expect it to change from the control to the experiment group.

Retention - This makes for a good evaluation metric because it is dependent on whether or not it is in the experiment or control group. We are expecting those who enroll in the experiment to be more likely to make an eventual payment since they have been warned about the workload. Thus it will not make a good invariant metric since it is likely to change from control to experiment.

Net Conversion - This is another good evaluation metric since it is dependent on the experiment. It amounts to the Gross Conversion times the Retention and will reveal whether the experiment actually gets us more, less or the same amount of paying customers in the end. Thus it will not make a good invariant metric since it is likely to change from control to experiment.

Also, state what results you will look for in your evaluation metrics in order to launch the experiment.
I will focus on Gross Conversion and Net Conversion as evaluation metrics since Retention can be thought of as just Net Conversion divided by Gross Conversion. In order to run the experiment there should be a practically significant decrease in Gross Conversion and no chance for a practically significant decrease in Net Conversion.

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: 0.0202
Net Conversion: 0.0156

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.

Both of the evaluation metrics I've selected are based on the click which uses cookies. Our unit of diversion is cookies, so since the denominator of the evaluation is the same as the unit of diversion we can expect the analytic estimate and the empirical variability to be comparable.

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 be using the Bonferroni correction.

Clicks needed per group for Gross Conversion:
BCR =20.625%
dmin = 1%
Alpha = 5%
1 - beta = 80%
Clicks required per group: 25,835

Clicks needed per group for Net Conversion:
BCR =10.93125%
dmin = 0.75%
Alpha = 5%
1 - beta = 80%
Clicks required per group: 27,413

TOTAL CLICKS REQUIRED PER GROUP: 27,413

There are .08 clicks for every pageview.

TOTAL PAGEVIEWS REQUIRED PER GROUP: 27,413 / .08 = 342,662.5

There are two groups: the control and the experiment.

TOTAL PAGEVIEWS REQUIRED: 685,325

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 100% of the traffic and this would take 18 days.

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

This is not a risky experiment. We are not dealing with particularly sensitive data, only whether or not someone has seen the warning and whether or not they choose to sign up and then whether or not they end up paying for the course. There is no danger of harm, financial risk or people being provided with dangerous misinformation. People are unlikely to get confused or upset by simply adding in a warning message that the workload for the course is significant. There is no reason not to use a large portion of the traffic in order to minimize the length of the experiment. In fact, there is no reason not to divert all traffic to this 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.)

Pageview (Number of Cookies) Sanity Check:

Sum of Control Group Pageviews: 345543
Sum of Experiment Group Pageviews: 344660
Total Sum of Pageviews: 690203
Probability in Control Group: 0.5
Probability in Experimental Group: 0.5
SE = SQRT(0.5*0.5/(345543+344660)) = 0.00060184074
Z-score: 1.96
Margin of error = z-score*SE = 0.00117960785
Confidence Interval = [0.5-m, 0.5+m] = [0.4988, 0.5012]
Observed Value: 0.5006
Verdict: PASSED!

Number of Clicks Sanity Check:

Sum of Control Group Clicks: 28378
Sum of Experiment Group Clicks: 28325
Total Sum of Clicks: 56703
Probability in Control Group: 0.5
Probability in Experimental Group: 0.5
SE = SQRT(0.5*0.5/(28378+28325)) = 0.00209974707
Z-score: 1.96
Margin of error = z-score*SE = 0.00411550427
Confidence Interval = [0.5-m, 0.5+m] = [0.4959, 0.5041]
Observed Value: 0.5005
Verdict: PASSED!

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:

Sum of Control Group Clicks by Nov 2: 17293
Sum of Experiment Group Clicks by Nov 2: 17260
Sum of Control Group Enrollments: 3785
Sum of Experiment Group Enrollments: 3423
Control Group Gross Conversion: 0.218874689
Experiment Group Gross Conversion: 0.1983198146
P-hat Pooled = (3785+3423)/(17293+17260) = 0.2086070674
SE = SQRT(0.2086*(1-0.2086)*(1/17293 + 1/17260)) = 0.00437162085
Z-score = 1.96
Margin of error = z-score*SE = 0.00856837686
D.hat = p.exp - p.cont = -0.0205548744
Confidence Interval = [D.hat-m, D.hat+m] = [-0.0291, -0.0120]
Dmin (practical significance) = +/- 0.01
Verdict: Statistically Significant! (CI does not contain 0); Practically Significant! (CI does not contain dmin)

Net Conversion:

Sum of Control Group Clicks by Nov 2: 17293
Sum of Experiment Group Clicks by Nov 2: 17260
Sum of Control Group Payments: 2033
Sum of Experiment Group Payments: 1945
Control Group Net Conversion: 0.11756201931
Experiment Group Net Conversion: 0.11268829664
P-hat Pooled = (2033+1945)/(17293+17260) = 0.11512748531
SE = SQRT(0.1151*(1-0.1151)*(1/17293 + 1/17260)) = 0.00343377688
Z-score = 1.96
Margin of error = z-score*SE = 0.00673020269
D.hat = p.exp - p.cont = -0.00487372267
Confidence Interval = [D.hat-m, D.hat+m] = [-0.01160392536, 0.00185648002]
Dmin (practical significance) = +/- 0.0075
Verdict: NOT Statistically Significant! (CI does not contain 0); NOT Practically Significant! (CI does not contain dmin)

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:

Successes (exp-cont positive): 4
Trials: 23

P-value (two tailed for probability 50%): 0.0026
Lower Alpha-value = 0.025
P-value > Alpha-value, **so practical and statistical significance!**

Net Conversion:
Successes (exp-cont positive): 10
Trials: 23
P-value (two tailed for probability 50%): 0.6776
Lower Alpha-value = 0.025
P-value > Alpha-value, **so NO practical or statistical significance.**

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.

The Bonferroni correction is used to guard against individual metrics incorrectly rejecting the null hypothesis which is more likely when we are considering a set of statistical inferences simultaneously. This is problematic if you require only one of your metrics to meet some criteria in order to launch. However we face the opposite problem. Our problem is that we need both of our metrics to meet some criteria in order to launch and we run a bigger risk of failing to meet the criteria by chance in one metric and thus not launching than we do of meeting the criteria by chance. Thus I have elected not to use the Bonferroni correction. There are no discrepancies between the effect size hypothesis tests and the sign tests. Both yielded the same result.

Recommendation
Make a recommendation and briefly describe your reasoning.

We should not launch the experiment since it did not meet our launch criteria. The Gross Conversion did indeed decrease in a practically significant way in the experiment group however the Net Conversion revealed a potential decrease since the confidence interval contains the negative practical significance, meaning our Net Conversion may actually be outside of the practical significance boundary for staying the same, which is what we were hoping to achieve with the Net Conversion (neither an increase nor a decrease).

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.

An interesting experiment to run would be daily email (or phone) reminders which keep track of the work you've already done and how many hours of work you have left in the two weeks leading up to payment. Ideally, this will help people stay on track with their work for the course and not be overwhelmed at the end of the week. Hopefully this will increase retention (probability of payment given enrollment).

Null Hypothesis: The daily reminders do not increase retention in a practically significant manner.

Unit of Diversion: The user id can now be used as a unit of diversion since we are interested in sending email reminders to accounts which have already signed up.

Invariant Metric: Number of user ids can be used as an invariant metric since sign-up occurs before daily reminders are received.

Evaluation Metric: Retention, probability of payment given enrollment. If retention in the experiment group demonstrates a practically significant increase we can roll out this experimental feature.