P7: Ying Wu

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.)

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.

The customer funnel on website has 4 parts:

A. # of unique cookies view

|

B. # of unique cookies click

|

C. # of users enroll

|

D. # of users converts (stay for over 14 days)

The experiment is adding a prompt between part B and C of funnel. The hypothesis is that by setting clearer expectations for potential students, you’ll have less students drop out from the free trial without significantly reducing the number of students who convert and eventually finish the class.

# of unique cookies view (Number of cookies) – part A:

* Invariant metric: Yes – occurs before experiment so no change is expected
* Evaluation metric: No – occurs before experiment, don’t expect any changes

# of unique cookies click (Number of clicks) – part B:

* Invariant metric: Yes – occurs before experiment so no change is expected
* Evaluation metric: No – occurs before experiment, don’t expect any changes

# of users enroll (Number of user-ids) – part C:

* Invariant metric: No – this number is not independent of the experiment
* Evaluation metric: No – this number will change but without comparing to conversions, the difference in user enrollment is not meaningful.

Click-through-probability – (part B) / (part A):

* Invariant metric: Yes – occurs before experiment so no change is expected
* Evaluation metric: No – occurs before experiment, don’t expect any changes

Gross conversion – (part C) / (part B):

* Invariant metric: No – this number is not independent of the experiment since number of users enrolling is expected to change from experiment
* Evaluation metric: Yes – this measures the effect of adding the prompt on enrollment. If the hypothesis is true, then Gross conversions will decrease since an option has been added for user to access without paying so fewer users will enroll.

Retention – (part D) / (part C):

* Invariant metric: No – this number is not independent of the experiment since number of users enrolling and number of users converting are expected to change from experiment
* Evaluation metric: Yes – this measures dropout rate. If the hypothesis is true, then retention should increase since the users who do enroll should have sufficient time to finish the coursework.

Net Conversion – (part D) / (part B):

* Invariant metric: No – this number is not independent of the experiment since number of users converting is expected to change from experiment
* Evaluation metric: Yes – If this hypothesis is true, then this metric should not have a statistically significant decrease. The explanation would be that that the decrease in users from Gross conversions should be balanced out by the increase in Retention.

**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 standard deviation calculation is: sqrt(1/N \* p \* (1-p)) where N is quantity of unit of analysis and p is probability of event

Gross conversion std dev: 0.0202 I would expect the analytic estimate and empirical variability to be comparable since the unit of analysis (cookies) match the unit of diversion for this experiment.

Retention std dev: 0.0549 I would expect the analytic estimate and empirical variability to be not comparable since the unit of analysis (user-id) does not match the unit of diversion for this experiment (cookies).

Net Conversion std dev: 0.0156 I would expect the analytic estimate and empirical variability to be comparable since the unit of analysis (cookies) match the unit of diversion for this experiment.

**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.)

Bonferroni will not be used since very few metrics are being tested and metrics are correlated so bonferonni would be overly conservative.

Gross conversion: 20.625% baseline conversion and 1% d­min at α = 0.05 and β = 0.20 gives minimum sample size of 25,835 per group. Doubling this value for treatment and control means that 51,670 users need to click enroll and with a 8% click through probability, a total of 645,875 users would need to start in the funnel.

Retention: 53% baseline conversion and 1% d­min at α = 0.05 and β = 0.20 gives minimum sample size of 39,115 per group. Doubling this value for treatment and control means that 78,230 users need to click enroll and with 1.65% of users who view enrolling, a total of 4,741,212 users would need to start in the funnel.

Net conversion: 10.931% baseline conversion and 0.75% d­min at α = 0.05 and β = 0.20 gives minimum sample size of 27,413 per group. Doubling this value for treatment and control means that 54,826 users need to click enroll and with a 8% click through probability, a total of 685,325 users would need to start in the funnel.

Retention requires a much higher sample size to achieve desired power so we will drop that as an evaluation metric. The sample size we would need then is over 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.)

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

Given that there is very little risk to the user (no additional data collected), very few things being changed thus little engineering/implementation risk (adding a prompt), and change is not expected to substantially affect business, I would divert 100% of the traffic for this experiment. With 40k visits/day, we would need 18 days to reach the required sample size of 685,325.

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.**

Number of unique cookies view:

0.5 +/- 1.96 \* sqrt(0.5\*0.5/( 345543+344660) = [0.49882, 0.50118]  
Actual: 0.500639667 – PASS

Number of unique cookies clicks:

0.5 +/- 1.96 \* sqrt(0.5\*0.5/(28378+28325)) = [0.495884, 0.504116]

Actual: 0.500467347 – PASS

Click-through-probability:

28378/345543 = 0.82126

0.82126 +/- sqrt(0.82126 \* (1-0.82126) / 345543) = [0.08121036, 0.083041267]

Actual: 0.082182441 – 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:

Confidence interval is: [-0.02912032, -0.011989429]

This range is higher than practical significance (0.01) as well as being statistically significant (not encompassing 0)

Net conversion:

Confidence interval is [-0.01160431, 0.001856864]

This range is within practical significance (0.0075) as well as being not statistically significant (encompasses 0)

**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:

Negative gross conversion is observed in 19 of 23 days. This result in a two-tailed p-value of 0.0026. Since this value is less than our α of 0.05, the sign test tells us that we see a statistically significant difference in Gross conversions.

Net conversion:

Negative net conversions is observed in 13 of 23 days. This results in a two-tailed p-value of 0.6776. Since this value is greater than our α of 0.05, the sign test tells us that we are not seeing a statistically significant difference in net conversion.

**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.

We will not use Bonferroni correction. Both net and total conversion metrics must match our expectations before proceeding with launch. Thus, family wise error rate control what we are interested in. Bonferroni correction is often used when we are looking across a large number of metrics and we are looking for any metric to match expectations before launching. In this latter case, we would be looking to control the false discovery rate.

**Recommendation**

Make a recommendation and briefly describe your reasoning.

Based on the results, I would not launch the changes. The expectations that less students would sign up was confirmed by a statistically significant decrease in Gross conversions. Although net conversions showed no statistically significant differences, the confidence interval was centered on -0.005 with a 95% confidence interval from -0.011 to 0.002. The lower bound of this confidence interval exceeds the negative practical significance boundary (-0.0075). This suggests that net conversions could be decreasing and this would have an adverse revenue impact.

Additional experiments or larger sample size could be run as follow-up to assess if the confidence interval for net conversions would get smaller and if net conversions will increase.

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 believe the obvious next experiment would be changing the wording on the prompt to adjust number of hours recommended as well as using different units (such as average hour commitment per day or per month). The metrics measured would be the same as this experiment with the hypothesis that we could make a better prompt with the same goal of keeping net conversions about the same and decreasing gross conversions (lowering student dropout rate).

A bigger change could be tracking user interaction time within the trial period and sending a reminder email if the user has not viewed at least 5 lessons within the first week. The hypothesis would be that users might have signed up but forgot so a quick reminder halfway through the trial period could remind the user to continue with the trial. For this experiment, the unit of diversion would be user-id and metrics we would want to evaluate are lessons watched in second week versus first week (expect to increase in experiment), percentage of users who enroll in trial that pay (expect to stay the same or increase in experiment). Invariant metrics would include number of users and number of lessons watched in the first week per user since both of these should be unaffected by the change.

I would run this experiment on a small subset of users first since a large increase in emails sent could trigger spam filters and run the experiment for at least 4 weeks since the trial takes two weeks and with 4 weeks we could also see if there are month effects (such as possibility that salaries will be paid at the end of the month affecting outcome).

Success would be defined as a statistically significant increase in videos watched in second week versus first week along with statistically significant increase in % of users converting.

References

<https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval> <https://en.wikipedia.org/wiki/Binomial_distribution#Normal_approximation>

<http://www.evanmiller.org/ab-testing/sample-size.html>

<http://graphpad.com/quickcalcs/binomial1.cfm>