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

Since Udacity wants to improve the overall student experience and asks student whether they can devote their time more than 5 hours per week or not. This process is done before they enroll Udacity, so counting "Number of user-ids" which is crated "after" the enrollment is not effective. Students are asked questions when they click the "start free trial", so "number of clicks" is inevitable for invariant metrics. For the comparison, "number of cookies" is needed because if there's only one invariant metric, we can't decide whether the change is from the questions students are asked or not.

Evaluation Metrics

- Gross conversion
- Retention
- Net conversion

Gross conversion is, number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button. When students click the "Start free trial", they are asked questions whether they can devote enough time for Udacity. If they can devote enough time, they are encouraged to enroll. However if they can't, they are encouraged to access course materials. This will be only effective when they are asked this question. Therefore, this metric is necessary.

Retention is number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. This metric is also necessary to check whether the test is effective on improving students' experience or not.

Net conversion is number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button. This metric is also necessary because of the same reason described in Retention.

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.

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
Retention: 0.0549
Net conversion: 0.0156

As for Gross conversion and Net conversion, unit of diversion is equal to unit of analysis. Therefore, the analytical estimate would be comparable to the empirical variability.

As for Retention, unit of diversion is not equal to unit of analysis. Hence the empirical variability may be different from the analytical estimate. Therefore we perform both an analytical and empirical estimate for retention metric.

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.

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 Bonferroni correction during analysis phase. Since there are three metrics, I set α =0.05/3 $\fallingdotseq$0.02(in the calculation I used 0.02)
I selected Gross conversion, Retention, Net conversion as an evaluation metrics.
The pageviews of Gross conversion, Retention, Net conversion are 960125, 6062182 and 1465600 respectively. Therefore, to test the experiment, we need 6062182 pageviews.

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.)

Unique cookies per day are 40000. It means it will take 6062182/40000=152 days to finish the experiment. This number isn't feasible so I will use the number of Net conversion (1670550) instead. In this case, the duration will be 1465600/40000=37 days. It is still a bit long time, but much feasible than before.

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

The number of cookies in control group and experiment group is 345,543,344,660 respectively. The total number of cookies is 690,203.

Standard Error(SE)=$\sqrt{0.5*0.5*(1/690203)}$

m=SE*1.96

Lower bound=0.5-m=0.4988

Upper bound=0.5+m=0.5012

Observed value =345543/690203=0.5006

Lower Bound	Upper Bound	Observed
0.4988	0.5012	0.5006

Observed value is in between lower bound and upper bound. Therefore, number of cookies passes sanity check.

Number of clicks on "Start free trial"

The number of clicks in control group and experiment group is 28,378,28,325 respectively. The total number of cookies is 56,703.

Standard Error (SE)=$\sqrt{0.5*0.5*(1/56703)}$

m=SE*1.96

Lower bound=0.5-m=0.4959

Upper bound=0.5+m=0.5041

Observed value =28378/56703=0.5005

Lower Bound	Upper Bound	Observed
0.4959	0.5041	0.5005

Observed value is in between lower bound and upper bound. Therefore, number of clicks on "Start free trial" passes sanity 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.)

I used Bonferroni correction to calculate the effect test size. Therefore α =0.05/2 (I used two metrics: gross conversion and net conversion.)

$P_{pool}=(X_{cont}+X_{exp})/(N_{cont}+N_{exp})$
$SE_{pool}=sqrt(P_{pool}*(1- P_{pool})*(1/ N_{cont}+1/ N_{exp})$
d_hat= $X_{exp}/ N_{exp}- X_{cont}/ N_{cont}$
$m=2.24*SE_{pool}$
Lower Bound=d_hat-m
Upper Bound=d_hat+m

Gross conversion
X_{cont}=3785, X_{exp}=3423, N_{cont}=17293, N_{exp}=17260

Lower Bound	Upper Bound
-0.0303	-0.0108

Since the range between lower bound and upper bound doesn't include 0, it means that Gross conversion's effect test size is statistically significant.

Net conversion
X_{cont}=2033, X_{exp}=1945, N_{cont}=17293, N_{exp}=17260

Lower Bound	Upper Bound
-0.0126	0.0028

The range between lower bound and upper bound includes 0.Therefore net conversion's effect test size isn't statistically 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:
Number of success:4
Number of trials:23
Probability:0.5
p-value is 0.0026

p-value is less than 5%, so gross conversion is statistically significant.

Net conversion:
Number of success:10
Number of trials:23
Probability:0.5
p-value is 0.6776

p-value is far more than 5%, so net conversion isn't 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.

There are two evaluation metrics that I used in this A/B test. Since I used multiple metrics and the metrics are not so high, I used Bonferroni correction. When it comes to Gross conversion, both effect size hypothesis and the sign test show statistically significant. When it comes to Net conversion, neither effect size hypothesis and sign test shows statistically significant, meaning there's no discrepancies between them.

Recommendation

Make a recommendation and briefly describe your reasoning.

I recommend the update, even though there's discrepancies. According to Gross conversion, the number ratio of entering free trial has decreased and the number is statistically significant

from the both effect size test and sign test. The decreased number means that the number of students who think they can't dedicate their time for more than 5 hours has decreased.This is good for Udacity because Udacity can improve coaches' capacity to support students who are likely to complete the course.

Certainly the net conversion doesn't show any statistically significant result, but the expense for the update doesn't cost too much.(Only front-end change.) Therefore, it's worth updating.

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 want to suggest to set up free preliminary NanoDegree course. People who enter "free trial" have to finish pre-Nanodegree within 2 weeks. If they can't finish it, they are advised to access course materials(free). People who finish the pre-Nanodegree have a choice to enter the "Paid" nanodegree course. That is for everybody who wants to take nanodegree. This is because people who "THINK" they can devote their time for more than 5 hours per week may not actually spend time for more than 5 hours. Threfore, by setting pre-Nanodegree course, those people will realize that they can't devote 5 hours per week.

The hypothesis is this is that Udacity will enhance the number of motivated people who finish nanodegree and will help reduce the number of frustrated students and students who "thought" they can devote more than 5 hours. If this hypothesis hold true, Udacity can improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward.

Invariant metrics

・Number of unique cookies: Since the unit of diversion is cookie, unique cookies are divided evenly by control and experiment group. Therefore, unique cookies are good metric.

・Number of people who click "start free trial": I want to know whether setting up pre-Nanodegree is effective or not, this metric is significant. We can't know the difference of the change before and after "start free trial", the change do happen when students click the "start free trial". Therefore, this metric is necessary.

Evaluation metric

・Net conversion: number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button.

Those who finished the pre-Nanodegree realize that they are appropriate for taking nanodegree or not. So this ratio will change by setting pre-nanodegree. Therefore, this metric is necessary.