A/B Testing

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.

Invariant Metrics:

1. Number of cookies: This metrics is used as a invariant metrics because it's evenly distributed in control and experiment group.
The number of cookies is not sufficient to measure the result of the experiment so, it can't be used as an evaluation metric.

2. Number of clicks: This metric is used as an invariant metrics because the event happens before the experiment and the metrics don't change.
Number of clicks cannot be used to measure the effect of the experiment and hence, it's not a evaluation metric.

3. Click through probability: This metric is used as an invariant metrics because the event happens before the experiment so, It's unaffected.
It measures the event before the screen gets shown so, It can't be used to measure the effect of experiment. Hence, it cannot be used as a evaluation metric.

Evaluation Metrics:

1. Gross conversion: This is the good evaluation metrics because we get to know about how much students actually enrolled after seeing the screener. This help to reduce the cost for the number of students who are not intrested and don't have much time for completing nanodegree. Also enrollment happens after the screener is shown so, it cant't be an invariant metric.

2. Retention: This is a good evaluation metric because It keeps the record of all students who make atleast one payment and also it follows the experiment and is quite affected by it.
The enrollment happens after the screener is shown and hence, it cant't be an invariant metric.
3. Net conversion: It captures the change in behaviour due to experiments so, it's an evaluation metric.
The enrollment happens after the screener is shown and hence, it cant't be an invariant metric.

Number of user_ids: Since we want to take into consideration the total number of cookies that clicked the button So, ratio of this metrics and the number of clicks which is a gross conversion, is a good choice of evaluation metrics since this ratio can normalize to different sized experiment and control groups. Also this ratio can marginalizes variances in the emprical count of user id's.
Also the number of user who enrolls in a free trial is dependent on the experiment so it's not a good invariant metrics.

The goal is to 1. decrease the enrollment by unprepared students 2. without decreasing the number of students who complete the free trial and make payment.

Gross Conversion metric should pass should pass both the statistical and practical significance test. Since the student who enrolls for free trial are aware of the time minimum time committment. So the Gross conversion of experimental group is lower than the Gross control of the Control group. So, goal conversion addresses goal 1.

We are hoping that Net conversion of experimental group is not significantly lower than that of the Control group. Since we are expected to see less students enroll in free trial in the experimental group but if our hypothesis is true then less students after free trial will drop.
So, Net conversion addresses goal 2 by significantly increase in the number of students or not significantly decrease in the number of students to continue past the free trial and completes the course to increase revenue.

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.

STD = sqrt((p*(1-p))/N)
= sqrt((0.20625*(1-0.20625))/400)

GROSS CONVERSION:

N = 5000*(3200/40000)
= 400

p = 0.20625 (Given)

= 0.0202306

RETENTION:

p = 0.53 (given)
N = 5000*0.08*0.20625 = 82.5
std = 0.0549

NET CONVERSION:

p = 0.1093125 (given)
N = 5000*0.08 = 400
std = 0.01560155

Since the quantity on the denomitor is number of unique cookies to clicks button for both the metrics so, both the analytical and emprocal variability matches.
In retention the number of user enrolled in the free trial is far away from the numbe of cookies so, analytic and imprical variability can be different.

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.)
No, I don't use the Bonferroni correction because the metrics used in the experiment are highly correlated.
For net conversion:
baseline rate= 10.93125%
minimum detectable effect = 0.75%
sample size = 27413
Number of page views needed = 685325

similarly,
For Gross conversion:
Number of page views needed = 645875

For Retention:
Number of page views needed = 4741212

since the number of pageviews in Retention(4741212) is much greater than Gross Conversion(645875) and net conversion(685325) so, retention will take greater time.
So, net conversiona nd gross conversion are choosen to proceed further.

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?

Since the number of page views is greater, 100% of traffic is used to speed up the data collection process. 18 days are required for the experiment to run (685325/40000).
The experiment is not at all risky because the data that is collected is not at all confidential. There is no harm to user. All the data being collected is not confidential and is collected through user.

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.**

NOTE: denoting confidence interval as C

Number of cookies:
C -> (0.4988, 0.5011)
observed value: 0.5006

Number of clicks:
C->(0.495884, 0.504115)
observed value: 0.5004673

Click through probability:
C->(-0.001295, 0.001295)
observed value: 0.000056627

All the 3 metrics pass through the sanity checks.

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:
C-> (-0.02912, -0.0119)
The metric is statistically significant since it doesn't conatins zero in its interval and is also practical significant since it's upper bound is smaller than -0.01.

Net Conversion:
C->(-0.0116, 0.0019)
The metrics is not statistically significan since it does not includes zero in its interval.
Also it's not pracically significant as its not 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.)

For the Gross Conversion rate:
0.0026

For the net Conversion rate;
0.6776

Only Gross Conversion rate passes the cutt off so it's 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.

In statistics, Bonferroni correction is used when we want multiple corrections and when multiple independent test are performed simultanously.
For this experiment both gross conversion and net conversion needs to be significant. As bonferrino is much conservative so it's not used.
If we have just one metric from many to make a decision then Bonferroni will be suitable for use.
But in our hypothesis we have two metrics so Benferroni isn't a good choice.

No discrepancies were found between effect size test and the sign test.

Recommendation

Make a recommendation and briefly describe your reasoning.

Net conversion is insignificant both practically and statistically. It's confidence interval includes negavtive number, so launching this is not a good idea because it' is possible that this number went down by an amount that would hurt the business by decreasing the revenue. Also it is very risky to just depend upon reduction of free trials or decrease in Gross conversion.

Whereas Gross Conversion is practically and statiscally significant because it lowers cost by discouraging the trial sign up. Gross Conversion decrease enrollment by unprepared students.

So, my recommendation is not to launch the experiment.

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.

The follow-up experiment could be to reward more discount to the student on the completetion of nanodegree if the students exceeds specification on more than half of his projects.
The hypothesis is to provide reward greater than 50% on completion according to the number of projects which exceeds specification.
Also this will provide students the motivation and students will denote more hardwork for completing their nanodegrees.

Unit of diversion: user_ids
evaluation metrics: number of payments divided by the number of unique user-ids

invariant metric: number of user_ids
since we want to keep the count of students who kept enrolled for greater time.
user_id will be a good invariant metric because it is unique for every user once enrolled.

As number of students who enroll should not change across experimental or control groups so it's a good choice for an invariant metric, while the number of payments / completed free trials divided by the number of user-ids should change over experimental and the control group so, it is a good evaluation metric.

References:

https://en.wikipedia.org/wiki/Sign_test

Udacity A/B tesing course

https://en.wikipedia.org/wiki/Effect_size