Experiment Design

Metric Choice

Invariant metrics: Number of cookies, Number of clicks, Click-through-probability
Evaluation metrics: Gross conversion, Net conversion
Explanation:
Number of cookies, Number of clicks and Click-through-probability have unit of analysis of cookies, which is also be the unit of diversion. Number of cookies and Number of clicks should be approximately equal in control group and experiment group. Click-through-probability should be invariant because it is not influenced by experiment since the change happened after click 'Start Free Trial'.
I didn't use Number of user-ids because it's number of users who enroll in the free trial. It's not an associate metric in this experiment, in which we concern number of users who enroll in paid version of courses.
Gross conversion, Net conversion and Retention are influenced by the change, since they all be measured after enrollment. (The change influences from clicking 'Start Free Trial' to enrollment.) Gross conversion and Net conversion have denominator of cookies, so cookie is their unit of analysis which is equal to unit of diversion. However, Retention's unit of analysis is user-id, which is larger than unit of diversion (cookies). In this case, a user-id could generate multiple cookies. Cookies from same user-id could be in control group or in experiment group, which might be a potential problem. And also because of unit of diversion is not equal to unit of analysis, analytical estimate of variability would be underestimated.
I used online calculator and found Retention needs at least 39115 samples (user-ids) per variance. With Enrollment per day is 660, we'll at least need 39115*2/660=120 days, and it is just calculated by underestimated analytical estimate, the empirical estimate is even larger, means more than 120 days is needed in this experiment. That's much more than Gross conversion and Net conversion which unit of diversion as cookies, which needs 18 days. So I didn't included Retention as evaluation metrics.
If the change in enrollment process has a expect result, it would reduce the number of user-ids enrolled. So the Gross conversion will decrease significantly. And ideally we hope it only reduces the number of user-ids who couldn't pass through 14days trial; the number of user-ids paying wouldn't be impacted. So Net conversion will not significantly change.
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

Standard deviation of Gross conversion is 0.0202 (analytical estimate)
Standard deviation of Net conversion is 0.0156 (analytical estimate)
Gross conversion and Net conversion both have denominator of cookies which is the unit of diversion. When the unit of diversion is equal to unit of analysis, the analytical

estimate of SE would be comparable to empirical estimate. We could calculate and use analytical estimate of SE.

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

I wouldn't use Bonferroni correction. The Gross conversion and Net conversion are depended. The assume independence of alpha-overall is too strict, and Bonferroni correction would even be more strict than alpha-overall.
I used the online calculator (http://www.evanmiller.org/ab-testing/sample-size.html) to calculate the samples needed in Gross conversion and Net conversion.
The number of samples of Gross conversion is 25835, the number of samples of Net conversion is 27413. They are the clicks of 'Start Free Trial', the pageviews should be multiplied by the click-through-probability of 0.08. Thus we got the least pageviews needed to power the experiment is 27413/0.08*2=685325

Duration vs. Exposure

Since the baseline values of unique cookies to view page per day is 40000, even if we divert all traffic to the experiment, we would need 18days (685325/40000=17.1). The risk is we couldn't run other experiments in the days when we run this experiment.

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

Experiment Analysis

Sanity Checks

The Number of cookies and Number of clicks on "Start Free Trial" should be approximately equal in control group and experiment group.
The 95% CI for Number of cookies is 0.4988-0.5012, the actual observed value is 0.5006, pass the sanity check.
The 95% CI for Number of clicks is 0.4959-0.5041, the actual observed value is 0.5005, pass the sanity check.
I also calculated sanity check on Click-through-probability:
Control group Control click-through-probablity: p = 28378/345543 = 0.0821
SE = sqrt(p(1-p)(1/345543))
M = SE * 1.96 = 0.0009
95% confident interval [0.0812, 0.0830]
Observed experiment CTR: p= 28325/344660 = 0.0822 Pass the 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

The 95% CI around the difference between the experiment and control groups on Gross conversion is: -0.0291—0.0120, significant (not include 0) and practical significant (dmin=0.01)

The 95% CI around the difference between the experiment and control groups on Net conversion is: -0.0116—0.0019, not significant (include 0) and not practical significant (include -0.0075)

Sign Tests

Sign test (http://graphpad.com/quickcalcs/binomial1.cfm) using the day-by-day data:

Gross conversion:

success(+) is 19,total is 23, two tailed p value is 0.0026, significant!

Net conversion:

Success(+) is 13, total is 23,two tailed p value is 0.6776, not significant!

Summary

I wouldn't use Bonferroni correction. The Gross conversion and Net conversion are depended. The assume independence of alpha-overall is too strict, and Bonferroni correction would even be more strict than alpha-overall.

The sign test has lower power than the effect size test, which is frequently the case for nonparametric test. That's the price you pay for not making any assumption.

In this case, Gross conversion and Net conversion are binomial; we also assume that d-hat is normal distribution because the n is large enough to meet n*p>5 and n*(1-p)>5. So we use binomial analytical estimate of SE and 1.96*SE to calculate margin.

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.

Recommendation

Udacity should not launch the experiment. Gross conversion is significantly reduced, which meets our expectation. However, Net conversion has CI included dmin, that indicate that the Net conversion would probably decrease larger than 0.0075. So we should be very careful with this change. I think we may run another experiment with longer time, which will give us more samples and consequently narrow CI to help us make a decision.

Make a recommendation and briefly describe your reasoning.

Follow-Up Experiment

The follow up experiment I'd like to test if the students turn to Free courses or they quit Udacity!
Actually we don't expect these students who stop when see the notice message in enrollment and leave Udacity, so it's necessary to test if the enrollment probability of Free courses increases when launch the change.
The Enrollment Probability in Free courses = user-ids of enrollment in Free courses/'Start Free Trail' clicks
The null hypothesis is the enrollment probability remains
The alternative hypothesis is the enrollment probability changes
The invariant metric is Pageviews and Clicks. (same to this experiment)
The evaluation metric is Enrollment Probability in Free courses
The unit of diversion is cookies, for users that do not enroll, their user-ids is not tracked in the experiment, even if they were signed in when they visited the course overview page.

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.