A/B Testing Final Project Instructions

Experiment Overview: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the course overview page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.
In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead. This screenshot shows what the experiment looks like.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could 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. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

Experiment Design

The null hypothesis is the gross conversion doesn’t change, the net conversion does not change either.
I used the Bonferroni correction, since I was analyzing multiple metrics, gross and net conversions, at the same time. Since the number of metrics I use is only 2, and they are not positively correlated, the Bonferroni correction should not be substantially conservative.

Metric Choice

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. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

Explanations / Reasons:

Number of cookies: That is, number of unique cookies to view the course overview page. (d_{min}=3000) #invariant. The change is only triggered after the viewing of the course overview page, so this should stay the same.

Number of user-ids: That is, number of users who enroll in the free trial. (d_{min}=50) # decrease, #unused. If the hypothesis is true, then this number should decrease. However, since it is not normalized to a total population, it would difficult to compare this number in practice.

Number of clicks: That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (d_{min}=240) #invariant. This number should stay the same since it is counted before the change is triggered.

Click-through-probability: That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. (d_{min}=0.01) #invariant. Both the numerator and denominator of CTP are counted before the change is triggered.

Gross conversion: That 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. (d_{min}= 0.01) #should decrease, **evaluation.** The hypothesis is that, with the information that the course would require over 5 hours of work per day, those who don't have this much time, would choose not to complete checkout after clicking the "start free trial" button. So this gross conversion rate should drop. This is a good evaluation metric since it shows that it would reduce the burden of coaches.

Retention: That 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. (d_{min}=0.01) #should increase, **evaluation.** The hypothesis is that the number of user-ids that remain rolled after 14-day (numerator for retention) would not significantly drop, while the number of user-ids

who complete checkout (denominator for retention) should drop due to more information about study hours requirement. So do the math, the retention should increase as a result.
Retention was originally chosen by me as an evaluation metric, but later the calculations show that this one requires too big of the size. So in the end, retention was **not used** in the final experiment design.
Net conversion: That 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. (d_{min}= 0.0075) #should not be significantly reduced, **evaluation.** The hypothesis is that the number of user-ids that remain rolled after 14-day (numerator for retention) would not significantly drop, while the number of cookies to click the “start free trial” button should stay the same since it happens before the change is triggered. So the Net conversion should not decrease significantly. This is a good evaluation metric since it shows the change would not decrease the revenue.

Measuring Standard Deviation

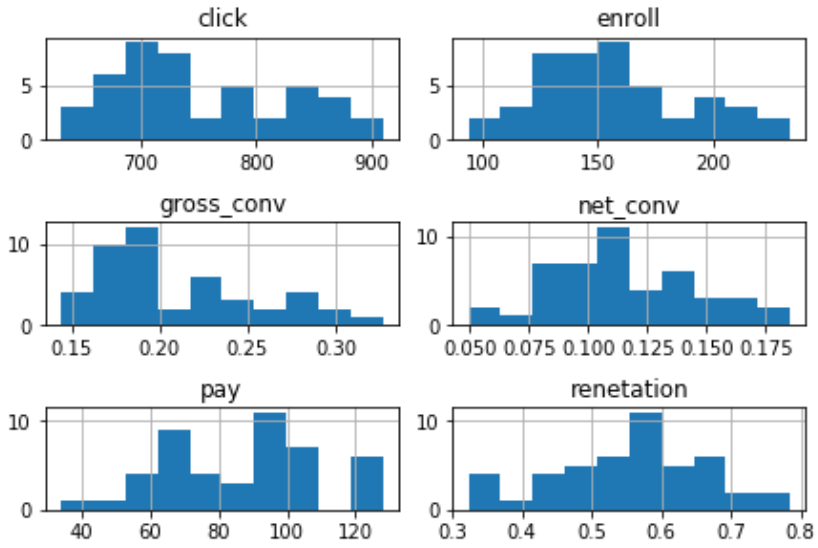
Given sample size of 5000 cookies:	Enroll or not enroll, pay or not pay, both follow binomial distribution,N*P > 10, so SE = sqrt(p(1-p)/N).				
Evaluations:	numerator	denominator	probability P	N	standard deviation
Gross conversion	enroll	click	0.20625	400	0.02024
Retention	pay	enroll	0.53	82.5	0.05498
Net conversion	pay	click	0.1093125	400	0.01568

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.

Empirical deviations are larger than the analytical ones listed above, explanations:

(1) The distribution isn’t normal either, the sample size of evaluations is only 46, the larger the sample size, the tighter the standard deviation.
(2) The claim of being binomial isn't always as true as people might hope since the rate may not be constant.

Empirical estimate of deviations (sample std, n-1):
gross_conv 0.0465
renetation 0.1122
net_conv 0.0306



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.

I use the Bonferroni correction.

alpha = 0.5, considering two tailed Bonferroni correction, alpha = 0.5/2/2 = 0.125, corresponding Z score = 2.24. alpha was originally calculated as 0.5/3/2, but later the retention data was not used since it requires too large sample size.

beta = 0.2, power = 0.8, corresponding to Z score = 0.84

Hence, use the equations below:

(0.84 + 2.24) * SE = d

SE = sqrt(2* p*(1-p)/n)

N in pageviews= n in clicks*2 *ratio, since sample size needs to be doubled since need two, one for control, one for experiment

Final size in unit of pageviews = max[776525.0, 821000.0] = 821000

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.

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

It will take 26 days if we use 80% of the traffic to do the test. I think the risky level is not that high since the change is only for one feature and is not technically challenging.

Experiment Analysis

Pageviews: Number of unique cookies to view the course overview page that day.
Clicks: Number of unique cookies to click the course overview page that day.
Enrollments: Number of user-ids to enroll in the free trial that day.
Payments: Number of user-ids who who enrolled on that day to remain enrolled for 14 days and thus make a payment. (Note that the date for this column is the start date, that is, the date of enrollment, rather than the date of the payment. The payment happened 14 days later. Because of this, the enrollments and payments are tracked for 14 fewer days than the other columns.)

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.
Number of counts, Number of clicks on “start free trial”, and CTP on “start free trial”, Sanity Check Results: all pass
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.**

	Ratio Expected	**Ratio Observed**	**Lower CI**	**Upper CI**	**Sanity Pass?**
Number of counts (Ratio of Control/Total)	.5	.5006	.4988	.5012	Pass
Number of counts (Ratio of Control/Total)	.5	.5005	.4959	.5041	Pass
CTP on “start free trial”	.0821	.0822	.0812	.0830	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.

	val_con	val_exp	val_avg	v_diff	SE	margin	ci_low	ci_upp	statistic_significance	practical_significance
gross	0.218875	0.198320	0.208607	-0.020555	0.004372	0.009793	-0.030347	-0.010762	True	True
net	0.117562	0.112688	0.115127	-0.004874	0.003434	0.007692	-0.012566	0.002819	False	False

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.

I used the Bonferroni correction. The two-tail alpha value should be 0.25

Gross conversion: The two-tail P value is 0.0026 This is the chance of observing either 19 or more successes, or 4 or fewer successes, in 23 trials. Since p-value < alpha = 0.25, the result is statistically significant.

Net conversion: The two-tail P value is 0.6776 This is the chance of observing either 13 or more successes, or 10 or fewer successes, in 23 trials. Since p-value > alpha = 0.25, the result is not 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.

I used the Bonferroni correction, since I was analyzing multiple metrics, gross and net conversions, at the same time. Since the number of metrics I use is only 2, and they are not positively correlated, the Bonferroni correction should not be substantially conservative.

Cite Wiki: Statistical hypothesis testing is based on rejecting the null hypothesis if the likelihood of the observed data under the null hypothesis is low. If multiple hypotheses are tested, the chance of a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis (i.e., making a Type I error) increases.[4]
The Bonferroni correction compensates for that increase by testing each individual hypothesis at a significance level of alpha/m, where alpha is the desired overall alpha level and m is the number of hypotheses.

There is no discrepancy between the effect size test and the sign test.

Recommendation

Make a recommendation and briefly describe your reasoning.

The goal of the experiment is to improve students experience as well as reduce coaches' load. So we are expected to observe a decrease in number of students/accounts who enroll in the free trials, but doesn't decrease the number of students who choose to stay after 14 days of trial and make payments. This corresponds to a decrease in the gross conversion, and a not significantly drop in net conversion.

The null hypothesis is the gross conversion doesn't change, the net conversion does not change either. The observed results reject the gross conversion null since the change of gross conversion is both statistically and practically significant, and retain the net conversion null since the change of net conversion is neither statistically or practically significant.
So I recommend implement the change.

Follow-Up Experiment

Follow-Up Experiment: How to Reduce Early Cancellations

If you wanted to reduce the number of frustrated students who cancel early in the course, what experiment would you try? Give a brief description of the change you would make, what your hypothesis would be about the effect of the change, what metrics you would want to measure, and what unit of diversion you would use. Include an explanation of each of your choices.

Generally speaking, there are some common reasons that students got frustrated by a online class: students didn't take enough prerequisite courses; the course materials are too hard; the class doesn't have Q & A hours, or no TA hours, etc. Obviously the project description doesn't contain enough information about the current state of this online class, so I would just assume that there is no online teaching assistant provided. In this case, I would suggest the change of adding an online teaching assistant who would held an online question and answer hour each week with a chatting window open to all students registered, i.e. those who enrolled in the free trial. The idea is that this would help the students with the coursework by providing a more interactive way of learning with relatively less effort of the coaches. Since it is a public chatting hour so students have access to other students questions, and the coaches do not have to waste time in answering the same questions over and over. We can also even save the chat log files so it would be accessible to students all the time.

Experimental Design:
Change I would make:
Add an online Q&A hour each week and make it visible to students after they click the "start free trial" button

Expected results:
Reduce the number of students cancel earlier, hence increase the number of students who make payments after the free trials.

Hypothesis:
Null: the number of students who make payments would stay the same.

Metrics:

The unit of diversion is user_id or cookie. 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.

Evaluation metrics:

Evaluations:	numerator	denominator	Expected effect
Gross conversion	enroll	click	increase
Net conversion	pay	click	increase

Invariant metrics:
Number of cookies, Number of clicks, Click-through-probability