Experiment Overview: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the home 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

Metric Choice

Invariant metrics are used to make sure that there is consistency between control and experiment group. When there is consistency between these two groups then we can compare two groups together.

Evaluation metrics are used to answer our research question.

Number of cookies, **number of clicks** and **click through probability** are chosen as invariant metrics.

We expect the number of cookies in the experiment and control groups to be the same to be able to compare them, so the first invariant metric that we chose was the number of cookies, also this metric is the unit of diversion in our analysis.

Since the experiment and the control group both should have the same or comparable number of unique cookies from users who click the "Start free trial" button, I chose Number of clicks as one of the invariant metrics.

Click through probability is also chosen as another invariant metric. Click through probability is actually the ratio of the number of clicks to the number of cookies. Even though we chose both of these metrics as invariant metrics, still we should use CTP(click through probability) as another invariant metric. The reason is that in reality we might not have a complete 50-50 split for the number of cookies in experiment and control groups, by considering this ratio we are actually removing the effect of this small deviation.

Gross conversion, **net conversion** and **retention** are chosen as evaluation metrics.

Gross conversion is defined as the number of user-IDs to complete checkout and enroll in the free trial divided by the number of unique cookies to click the "Start free trial" button. If the experiment group shows statistically significant decrease in Gross conversion rate, then this new feature could be helpful in reducing the number of students who proceed to the free trial.

Net conversion is defined as the 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 shows how many students remain enrolled after the free trial. If the experiment doesn't show significant decrease in net conversion then we would be interested to launch the new feature.

Retention is defined as the 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. If this ratio increase statistically significantly for the experiment group, we would be interested to launch this new feature.

In the experiment group people know about the time requirement for taking the courses so some of them may become unwilling to enroll for the courses because the new feature asks them about the duration of their availability. However in the controlled group the retention rate might be less due to the fact that they are not informed about the time requirement for completing the program.

Number of user IDs is actually the number of user IDs who enroll in the program. We also have gross conversion as another metric which is the ratio of the number of enrollments to the number clicks. So since gross conversion better meets our expectation, we wouldn't need to use user ID as an evaluation metric.

We would launch the experiment if we see statistically significant change in gross conversion and retention and no significant decrease in net conversion.

Measuring Standard Deviation

For estimating the standard deviation, first we need the following probabilities:

Probability of enrolling, given click	0.20625
Probability of payment, given enroll	0.53
Probability of payment, given click	0.1093125

Formula for the Standard Deviation: sqrt(p(1-p)/n)

Standard deviation for gross conversion:

sqrt((0.20625*(1-0.20625))/(5000*(3200/40000))) = 0.020231

Standard deviation for retention:
sqrt((0.53*(1-0.53))/(5000*(660/40000))) = 0.054949

Standard deviation for net conversion:
sqrt((0.1093125*(1-0.1093125)/(5000*(3200/40000))) = 0.015602

These values are analytic estimates of standard deviation, we can also calculate the empirical estimate for standard deviation. For those metrics that the unit of diversion equals the unit of analysis (denominator of the metric), we can expect analytic estimate to be accurate.

In this project the unit of diversion is cookies. Each cookie that visits the page will be diverted to either the control or experiment groups.

For gross conversion and net conversion, the unit of analysis for both evaluation metrics is still cookies, specifically, "the number of cookies that click on the 'start free trial' button" Since both the unit of diversion and the unit of analysis are the same, the empirical and analytic estimates should be similar.

However, given the constraints of this experiment, it doesn't make sense to use retention at all. The analytic estimate of the variability is already too high to feasibly use retention as an evaluation metric (the duration is simply too long), and getting a new, probably higher, empirical estimate is simply going to make the experiment even longer. So in this case, we don't have to get an empirical estimate, because we already know that retention doesn't make sense for the length of time we're willing to commit.

Sizing

Number of Samples vs. Power

For this part, we decided not to use Bonferroni correction, because we want all the evaluation metrics to match our expectations.

The online A/B test sampling size calculator to calculate the number of samples we needed. alpha=0.05, 1-beta=0.8 and baseline values are used for determining the sampling size.

Gross Conversion:

d_{min}	0.01
Baseline value	20.625%
Practical significant level	20%
Alpha	5%

Using the values from the above table and also using the online calculator we would have:

Sample Size= 25835

For gross conversion, the unit of analysis is cookies clicking the "start free trial" button.

For the total number of page views we should use the following formula:

Sample_size / (Cookies to click “start free trial” per day / Page views per day)

The total page views: 25835/(3200/40000) = 322938

The total number of page views for experiment and control group:
322938*2 = 645876

Net conversion:

d_{min}	0.01
Baseline value	10.93125%
Practical significant level	20%
Alpha	5%

Using the values from the above table and also using the online calculator we would have:

Sample Size= 27413

For net conversion, the unit of analysis is cookies clicking the "start free trial" button.

For the total number of page views we should use the following formula:

Sample_size / (Cookies to click “start free trial” per day / Page views per day)

The total page views: 27413/(3200/40000) = 342663

The total number of page views for experiment and control group:
342663*2 = 685326

The largest number should be selected as the number of page views. So we would choose 685326.

Duration vs. Exposure

Now we should consider the risks associated with the experiment. The data that we are dealing with is not sensitive data, and there is no risk of privacy involved in this project, because the users would not disclose any personal information about themselves, and they just declare whether they can assign a specific amount of time for learning or not. Also the participants of this experiment would not encounter any physical, psychological and emotional, social and economic harm. So we can safely use 100% of the traffic for this experiment, because no participant would be hurt by using 100% of the traffic.

After determining the largest number for the required page views, we would calculate the number of days needed to do the experiment with using 100% of traffic:

635326/40000 = 17.13, so considering using 100% of traffic around 18 days is needed to complete this experiment.

Experiment Analysis

Sanity Checks

We determine three invariant metrics earlier: number of cookies, number of clicks and click through probability. Now we want to determine the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check.

- **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

Control pageviews = 345543

Experiment pageviews=344660

Control clicks= 28378

Experiment clicks=28325

Enrollment experiment = 3423

Enrollment control = 378

Number of Cookies:
Probability of success = 0.5
SE(Standard Error) = sqrt (0.5*0.5 / ([Control Pageviews] + [Experiment Pageviews]))
m (Margin of Error) = 1.96 * SE = sqrt (0.5*0.5 / (345543+ 344660))= 0.001179608

Lower CI Bound = 0.5 - M = 0.498820392
Upper CI Bound = 0.5 + M = 0.501179608

Observed = [Control Pageviews] / ([Control Pageviews] + [Experiment Pageviews])

Observed =0.500639667

Since observed value is in the 95% of confidence interval [0.4988, 0.5012], it will pass the sanity check.

Number of Clicks:

Probability of success = 0.5
SE(Standard Error) = sqrt (0.5*0.5 / ([Control Pageviews] + [Experiment Pageviews]))

SE= sqrt (0.5*0.5 / (28378 + 28325))= 0.0021
m (Margin of Error) = 1.96 * SE = 1.96 * 0.0021 = 0.004115504
Lower CI Bound = 0.5 – m = 0.495884496
Upper CI Bound = 0.5 + m = 0.504115504
Observed = [Control Clicks] / ([Control Clicks] + [Experiment Clicks])

Observed = 0.5004673474

Since observed value is in the 95% of confidence interval [0.4959, 0.5041], it will pass the sanity check.

Click Through Probability:

X_{cont} (Control pageviews) = 345543

X_{exp} (Experiment pageviews) = 344660

N_{cont} (Control clicks)= 28378

N_{exp} (Experiment clicks) = 28325

$P_{pool} = \frac{Xcont+Xexp}{Ncont+Nexp} = 0.082154091$

SE_{pool} = 0.000661061

P_{exp} = 28325 / 344660 = 0.082182441

P_{cont} = 28378/ 345543 = 0.082125814

Upper bound= 1.96*SE_{pool} = 0.001295679

Lower bound = -1.96*SE_{pool} = -0.001295679

d(observed)= $P_{exp} – P_{cont}$ = 0.00005663

Since observed value is in the 95% of confidence interval [-0.0013, 0.0013], it will pass the sanity check.

Result Analysis

Effect Size Tests

Gross conversion:
In order to do the calculation for the effect size test, the data after 2 Nov is removed since it does not have enrolment data that we need to calculate gross conversion.

N_{cont} (Clicks control) = 17293

N_{exp} (Clicks experiment) = 17260

X_{cont} (Enrolment control) = 3785

X_{exp} (Enrolment Experiment) = 3423

$P_{pool} = (X_{exp} + X_{cont}) / (N_{exp} + N_{cont}) = 0.208607067$

$SE_{pool} = sqrt(P_{pool} * (1 - P_{pool}) * (1 / N_{cont} + 1 / N_{exp})) = 0.004371675$

Control Gross Conversion = 0.218874689
Experiment Gross Conversion = 0.198319815
m = SE_{pool}* 1.96 = 0.008568483
d hat = $X_{exp}/N_{exp} – X_{cont}/N_{cont}$ = (3423/17260) – (3785/17293) = -0.020554875
Lower CI Bound: d hat - m = -0.029123358
Upper CI Bound: d hat + m = -0.011986392

Since the limit range doesn't include 0(direction of change is clear) and also "d hat" is greater than d_{min} = 0.01 we conclude that difference in Gross Conversion to be both statistically and practically significant.

Net conversion:

In order to do the calculation for the effect size test, the data after 2 Nov is removed since it does not have payment data that we need to calculate net conversion.

N_{cont} (Clicks control) = 17293

N_{exp} (Clicks experiment)= 17260

X_{cont} (payments control) = 2033

X_{exp} (payments Experiment) = 1945

$P_{pool} = (X_{exp} + X_{cont}) / (N_{exp} + N_{cont}) = 0.115127485$

$SE_{pool} = sqrt(P_{pool} * (1 - P_{pool}) * (1 / N_{cont} + 1 / N_{exp})) = 0.003434134$

Control Gross Conversion = 0.117562019
Experiment Gross Conversion = 0.112688297

$m = SE_{pool} * 1.96 = 0.006730902$
$\hat{d} = X_{exp}/N_{exp} - X_{cont}/N_{cont} = -0.004873723$

Lower CI Bound: d hat - m = -0.011604624
Upper CI Bound: d hat + m = 0.001857179

Since the limit range includes 0(direction of change is clear) and also "d hat" is less than $d_{min}=0.0075$ we conclude that difference in Gross Conversion not to be neither statistically nor practically significant.

The interval does not include d_{min} (practical significance boundary) so it is recommended to launch the experiment. I can be confident at 95% of confidence level that the true change is large enough to be worth launching

Sign Tests:

Gross conversion:

Number of days= 23

Number of days with positive change= 0

Binomial p=0.5 (we have either positive or negative change)

According to the online calculator: p-value=0.0026

Since the p-value is less the chosen alpha of 0.05, the sign test agrees with the hypothesis test. This result is unlikely to come about by chance and the difference is statistically significant according to both the sign test and the effect size test

Net conversion:

Number of days= 23

Number of days with positive change= 10

Binomial p=0.5 (we have either positive or negative change)

Acording to the online calculator: p-value=0.6776

Since the p-value is greater the chosen alpha of 0.05, the difference is not statically significant.

Summary

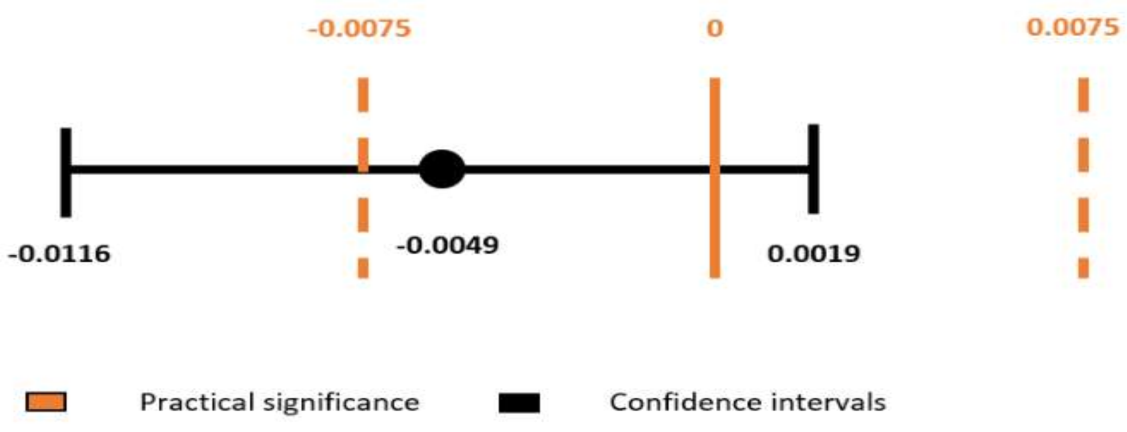
In this experiment we are interested to test whether adding the new feature to Udacity website would have positive effects, which is setting clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trail because of the lack of time, without significantly reducing the number of students to continue past the free trial.

As a result, in this experiment there are two types of students who could be screened out, students who cancel the free trial and students who continue past the free trial. The hypothesis for the experiment is that the first type of students would be reduced and the second type of students would not be significantly reduced. The first hypothesis can be tested with gross conversion and the second one can be tested with net conversion.

Bonferroni correction is not used for this experiment, because all the evaluation metrics including gross conversion and net conversion should match our expectations in order to be able to launch the new feature. Actually in this case we risk not to launch if at least of one the metrics fail to reject the null hypothesis when the null hypothesis is not the true effect. So we won't reduce this risk by using Bonferroni correction since we want all the metrics to meet our expectations.

Recommendation

After doing effect size test and sign test we conclude that according to both of these tests, there is statistically significant decrease in the case of gross conversion, and no statistically significant decrease in the case of net conversion.
However for net conversion, the confidence interval does include the negative of the practical significance boundary. So it's possible that this number went down by an amount that would matter to the business.
In this case, we can run the experiment with more power(requires more sample) and this can shrink the CI that we can reach statistical significance. However, the better approach would be changing the intervention, since it is unlikely to change the overall trend in the net conversion metric.



Follow-Up Experiment

As a follow up experiment, it is recommended to add a follow up prompt when the students want to cancel the subscription. When they click on the "cancel subscription" button, a message box would pop up which encourages students to do “Introduction to programming” course before deciding to cancel their subscription. This could work as an incentive to continue doing Nanodegree.

This follow up experiment would be done during the free trial period, and it is designed to reduce the number of frustrated students during the free trial. So the hypothesis would be as follows:

H0: The new feature has no significant increase on the Retention Rate of enrolled students.
H1 : The new feature has significant increase on the Retention Rate of enrolled students.

In this experiment we want to test whether this new feature would make statistically significant increase in Retention Rate or not. The invariant metric that we would use for this experiment is the user ID and the evaluation metric would be retention because this is an experiment which is done after the registration. The unit of diversion in this case is user ID, because the experiment is done after user log-in. We would launch the new feature if we observe statistically significant increase the retention rate.