Udacity Course Free Trial Screener

Experiment Background

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.

Udacity found that many students dropped out of the free trial due to insufficient study time. Our aim is to discourage enrollment in the free trial by students who lack sufficient time, as they are likely to discontinue later. However, we aim to avoid decreasing the number of students who complete the free trial and proceed to finish the course.

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.

Experiment Design

Unit of Diversion

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.

Metric Choice

Invariant metrics are used for sanity checks and will remain equivalent between control and experiment groups, including population sizing metrics and other actual invariants.

Evaluation metrics are what we care about, and they must be observed for consideration in the decision to launch the experiment.

d_min: The practical significance boundary for each metric, that is, the difference that would have to be observed before that was a meaningful change for the business. They are given as absolute changes.

Invariant metrics:

- **Number of cookies:** number of unique cookies to view the course overview page.
- **Number of clicks:** number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is triggered).
- **Click-through-probability:** number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.

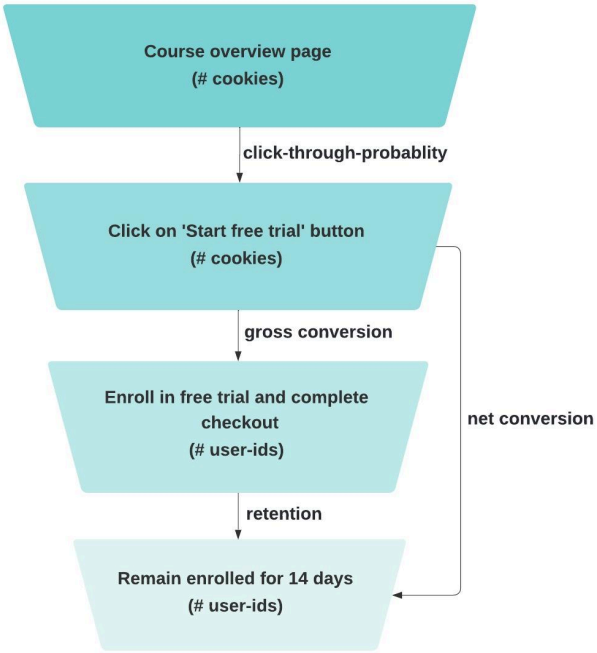
Evaluation metrics:

- **Gross conversion:** 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)
- **Retention:** 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)
- **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. (d_min=0.0075)

I am aiming for the following outcomes:

- A significant decrease in Gross Conversion
- A significant increase in Retention
- A significant increase in Net Conversion

The funnel below shows the process.



Measuring Standard Deviation

Based on the rough estimates of the baseline values for the metrics, make an analytic estimate of standard deviation for each evaluation metric, given a sample size of 5000 cookies visiting the course overview page.

Udacity provided the baseline values:

Unique cookies to view course overview page per day:	40000
Unique cookies to click "Start free trial" per day:	3200
Enrollments per day:	660
Click-through-probability on "Start free trial":	0.08
Probability of enrolling, given click:	0.20625
Probability of payment, given enroll:	0.53
Probability of payment, given click	0.1093125

The distribution of Gross conversion, Retention, Net conversion: binomial distribution.

So, the standard deviation: sqrt[p*(1-p)/N]

- **Gross conversion:**

  N = 5000*0.08 = 400, p = 0.20625

  standard deviation = 0.0202

  Unit of diversion: cookie

  Unit of analysis: cookie

The unit of diversion and the unit of analysis are the same, so the analytically computed variability is likely to be very close to the empirically computed variability.

- **Retention:**

  N = 5000*0.08*0.20625 = 82.5, p = 0.53

  standard deviation = 0.0549

  Unit of diversion: cookie

  Unit of analysis: user-id

  They are different. So the variability is going to be much higher. And the analytical estimate will be an underestimate of the empirical variance. It is worth doing an empirical estimate if there is time.

  But it seems that the unit of analysis is bigger than the unit of diversion, which may be inappropriate.

- **Net conversion:**

  N = 5000*0.08 = 400, p = 0.1093125

  standard deviation = 0.0156

  Unit of diversion: cookie

  Unit of analysis: cookie

  They are the same. So the analytically computed variability is likely to be very close to the empirically computed variability.

Sizing

Number of Samples vs. Power

Use an alpha of 0.05 and a beta of 0.2. (We will not use Bonferroni correction during the analysis phase.)

Pageviews total (across both groups) needed to collect to adequately power the experiment:

Using the code, here are the results:

```
import numpy as np
from scipy.stats import norm
import math

def get_z_star(alpha):
    return -norm.ppf(alpha / 2)

def get_beta(z_star, s, d_min, N):
    SE = s / np.sqrt(N)
    return norm.cdf(z_star * SE, loc=d_min, scale=SE)

def required_size(s, d_min, Ns=None, alpha=0.05, beta=0.2):
    if Ns is None:
        Ns = range(1, 200001)
    for N in Ns:
        if get_beta(get_z_star(alpha), s, d_min, N) <= beta:
            return N
    return -1
```

- **Gross conversion:**

  p = 0.20625

  s = math.sqrt(p*(1-p)*(1/1+1/1))

  required_size(s=s, d_min=0.01)

  N = 25699 (cookies)

  25699 * 2 / 0.08 = 642475 (pageviews)

- **Retention:**

  p = 0.53

  s = math.sqrt(p*(1-p)*(1/1+1/1))

  required_size(s=s, d_min=0.01)

  N = 39104 (user-ids)

  39104 * 2 / 0.20625 / 0.08 = 4739879 (pageviews)

- **Net conversion:**

  p = 0.1093125

  s = math.sqrt(p*(1-p)*(1/1+1/1))

  required_size(s=s, d_min=0.0075)

  N = 27172 (cookies)

  27172 * 2 / 0.08 = 679300 (pageviews)

In order to test for all three metrics, we would require the maximum number of pageviews: 4739879.

These metrics are correlated and all tend to move at the same time. In this case, Bonferroni correction is too conservative. So we will not use Bonferroni correction during the analysis phase.

Duration vs. Exposure

We know that the number of unique cookies to view the course overview page per day is 40000. If we divert all of the pageviews, it would take 119 days to run the experiment, which is way too long.

So we should reconsider the decision about using 4739879 pageviews and use 679300 instead.

We will not use Retention as an evaluation metric. For Gross Conversion and Net Conversion, we can use 100% of the traffic. The experiment will take 17 days.

The duration is short. So we can divert 80% of Udacity's traffic each day (assuming there were no other experiments we wanted to run simultaneously), and the experiment will last 22 days.

Also, the change is not risky enough, so there is no need to limit the number of users exposed.

Experiment Execution and Data Collection

The experiment data can be found in **this spreadsheet**. There are two sheets within the spreadsheet - one for the experiment group, and one for the control group.

Experiment Analysis

Sanity Checks

Each pageview and click is randomly assigned to the control or experiment group with probability 0.5. So the number for the control group and the probability for the control group will follow the binomial distribution. We can construct a binomial confidence interval. Because the total sample size is large, it is definitely enough to assume a normal distribution.

Standard error of binomial distribution with probability 0.5 of success: SE = sqrt[0.5*0.5/N]

Based on the experiment data, we can get:

	Pageviews	Clicks
N_control	345543	28378
N_experiment	344660	28325
N_total	690203	56703
p	0.5	0.5
p^=N_control / N_total	0.5006396669	0.5004673474

- **Number of cookies:**

  SE = sqrt[0.5*0.5/690203] = 0.0006

  m = SE * 1.96 = 0.0012

  lower bound = 0.5 - m = 0.4988

upper bound = 0.5 + m = 0.5012

95% confidence interval around 0.5: [0.4988, 0.5012]

Check whether observed fraction is within interval: Yes

- **Number of clicks:**

  SE = sqrt[0.5*0.5/56703] = 0.0021

  m = SE *1.96 = 0.0041

  lower bound = 0.5 - m = 0.4959

  upper bound = 0.5 + m = 0.5041

  95% confidence interval around 0.5: [0.4959, 0.5041]

  Check whether observed fraction is within interval: Yes

We can clearly see that the probability for both Pageviews and Clicks falls within the confidence interval. Therefore both metrics pass the sanity check.

To perform sanity check for the Click-through-probability, we would expect that the difference between the two groups would be zero.

Based on the experiment data, we can get:

	Click Through Probability
d	0
p_control	0.08212581357
p_experiment	0.08218244067
d^	0.00005662709159

- **Click-through-probability:**

  p_pooled = (28378+28325) / (345543+344660) = 0.0822

  SE_pooled = sqrt[p_pooled*(1-p_pooled)*(1/345543+1/344660)] = 0.0007

  m = 1.96 * SE_pooled = 0.0013

  lower bound = 0 - m = -0.0013

  upper bound = 0 + m = 0.0013

  95% confidence interval around 0: [-0.0013, 0.0013]

  Check whether observed difference is within interval: Yes

We can clearly see that the observed difference falls in the confidence interval, therefore the Click-through-probability passes the sanity check.

Result Analysis

Effect Size Tests

Calculate the 95% confidence interval around the difference between the experiment and control groups for each of our evaluation metrics, and check whether each metric is statistically and/or practically significant.

Previously, we had chosen Gross Conversion and Net Conversion as our final evaluation metrics.

d = probability for experiment - probability for control

Based on the experiment data, we can get:

	Control	Experiment
Clicks	17293	17260
Enrollments	3785	3423
Payments	2033	1945
p^ (Gross Conversion)	0.2188746892	0.1983198146
p^ (Net Conversion)	0.1175620193	0.1126882966

- **Gross conversion:**

  d_min = 0.01

  d^ = 0.1983-0.2189 = -0.0206

  p_pooled = (3785+3423) / (17293+17260) = 0.2086

  SE_pooled = sqrt[(p_pooled)*(1-p_pooled)*(1/17293+1/17260)] = 0.0044

  m = 1.96 * SE_pooled = 0.0086

  lower bound = d^ - m = -0.0291

  upper bound = d^ + m = -0.0120

  95% confidence interval around d^: [-0.0291, -0.0120]

  The confidence interval does not include 0, so Gross Conversion is statistically significant. Also, the confidence interval does not include negative d_min, so it is practically significant.

- **Net conversion:**

  d_min = 0.0075

  d^ = 0.1127 - 0.1176 = -0.0049

  p_pooled = (2033+1945) / (17293+17260) = 0.1151

  SE_pooled = sqrt[p_pooled*(1-p_pooled)*(1/17293+1/17260)] = 0.0034

  m = 1.96 * SE_pooled = 0.0067

  lower bound = d^ - m = -0.0116

  upper bound = d^ + m = 0.0019

  95% confidence interval around d^: [-0.0116, 0.0019]

The confidence interval includes 0, so Net Conversion is not statistically significant. Also, the confidence interval includes negative d_min, which means that it is not practically significant.

Sign Tests

In order to double check, we need to do a sign test using the day-by-day data for each evaluation metric. This is to check whether the signs of the difference of the metric between the experiment and control groups agree with the confidence interval for the difference.

We cannot assume a normal distribution, since 23 days is not enough for the binomial distribution to closely approximate a normal distribution. So, we do the sign test calculation using **this online calculator**.

If there is no difference, then 50% chance of negative change on each day.

- **Gross conversion:**

  # days: 23

  # days with negative change: 19

  The two-tail p-value: 0.0026

  Since this is less than the chosen alpha of 0.05, the sign test agrees with the hypothesis test. This result is unlikely to come about by chance. The result is statistically significant.

- **Net conversion:**

  # days: 23

  # days with negative change: 13

  The two-tail p-value: 0.6776

  Since this is larger than the chosen alpha of 0.05, the sign test agrees with the hypothesis test. It is not statistically significant.

Summary

For two metrics, the chance of at least 1 false positive: 1 - 0.95*0.95 = 0.0975 (assume independence)

However, the two evaluation metrics are correlated and tend to move at the same time.

Therefore, 0.0975 is an overestimate of the probability of at least 1 false positive.

This is acceptable. So I didn't use the Bonferroni correction. In this case, Bonferroni correction is too conservative.

Recommendation

This experiment is designed to understand whether the screener can help filter out students who don’t have enough time to devote to the course, thereby reducing the number of students who left the free trial because they didn’t have enough time.

Meanwhile, the experiment is designed to determine whether the screener is effective without significantly reducing the number of students to continue past the free trial and eventually complete the course.

Our results indicate a significant reduction in Gross Conversion with the implementation of the screener, thereby supporting our first goal. The screener will help reduce the enrollment, which can improve coaches' capacity to support students who are likely to complete the course.

Although there are no significant changes in Net Conversion, we cannot conclude that the number of users staying enrolled for 14 days is the same between the experiment and control groups, as there is insufficient evidence to confirm a consistent number of students making payments. It is likely that fewer students will end up completing the course, which is not what we want.

I would not recommend launching this screener.