Experiment Overview: Free Trial Screener

Background:

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.

Hypothesis:

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.

Impact:

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.

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

- **Number of cookies:** That is, number of unique cookies to view the course overview page. (d_{min}=3000)
- **Number of user-ids:** That is, number of users who enroll in the free trial. (d_{min}=50)
- **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)
- **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)

- **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)
- **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)
- **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)

Invariant Metrics:

- **Number of cookies:** That is, number of unique cookies to view the course overview page. (d_{min}=3000)
- **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)
- **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)

Reasoning: For the three metrics above, they all happened before they see the changes we are planning to launch. Users behavior should not change due to the change before or after. So I think these are the invariant metrics that won't be affected before or after the change is launched.

Evaluation Metrics:

- **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)
- **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)
- **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)

Reasoning:
Once again, here is the hypothesis: 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.

- **Gross conversion:** with the change, users will see the warning first before they decide enroll in the course. So users will have a clear expectation on how many time they need to commit for the course. Once they decide to sign up, the hope is they already have the planned time and would likely continue the course and eventually stay after 14 days of trial. Here if users don't have 5 hours or more to commit, they probably will switch to "access course material" only. The gross conversion rate after the change will likely not the same as before the change. Mostly, it probably will be lower as right now for students who don't have time commit yet, they will do option 2. We want to test if this is statistically significant and practically significant.
- **Retention**: for students who choose to enroll, they saw the warning before they do so. So we hope these are the more prepared students, and thus are more likely stay after the 14 days free trial. We would expect the

retention rate after the change will be higher than before. We would like to see if it's statistically significant and practically significant.

- **Net conversion**: Similar reasoning like the retention. We would expect the net conversion will be higher than before after the change. We'd like to see if it's statistically significant and practically significant.

Measuring Variability

Description	Value
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

*** Given a sample size of 5000 cookies visiting the course overview page*

For each metric, estimate its standard deviation analytically.

- **Gross conversion = # of enrol / # of click**
  - Binomial
  - Probability = 0.20625
  - N = 5000 pageview * probability of click (0.08) = 400
  - SD = sqrt(p*(1-p)/N) = sqrt(0.20625*(1-0.206225)/400) = 0.0202
- **Retention = enroll past 14/completed check out**
  - Binomial
  - Probability = 0.53 (from baseline file)
  - N = 5000 pageview * probability of click (0.08) * Probability of enrolling, given click (0.20625) = 82.5
  - SD = sqrt(p*(1-p)/N) = sqrt(0.53*(1-.53)/82.5) = 0.0549
- **Net conversion = paid / # of click start free trial**
  - Binomial
  - Probability = 0.1093125 (from based file)
  - N = 5000 pageview * probability of click (0.08) = 400
  - SD = sqrt(p*(1-p)/N) = sqrt(0.1093*(1-0.1093)/400) = 0.0156

To understand whether the analytical estimates of standard deviation are accurate, i.e. whether it matches the empirical standard deviation, we consider whether or not the unit of analysis and unit of diversion matches up.

- **Gross conversion:** In analytic estimate, the unit of analysis is # of people who click the "start free trial", and the unit diversion is # of cookies click. They are similar but not exactly the same as for the unit of diversion, some people can use different cookies to visit the same page, such as different device or different browser. But the two units are fairly correlated. So the analytic estimate is mostly accurate. If we test the empirical estimate if have time, but not highly needed.
- **Retention**: The unit of the analysis is # of people who enrolled the free trial, but the unit of diversion is # of user_id who did so. They are almost the same, as people who enrolled in the class will likely use the same user_id to continue their course. So the analytical estimates should match the empirical one well given the unit of analysis and unit of diversion have very strong match.
- **Net conversion:** The unit of analysis here is the same as the one in **Gross conversion**. So same logic apply here.

Sizing

Choosing Number of Samples given Power

Use the following link to estimate sample size. **Use an alpha of 0.05 and a beta of 0.2.**
Link: http://www.evanmiller.org/ab-testing/sample-size.html

- **Gross conversion**
  - Alpha = 5% /2 = 2%
  - Beta = 0.2, so 1-beta = 80%
  - Probability = 0.20625 (Baseline conversion rate)
  - Minimum Detectable Effect: 0.01
  - **Sample size = 33,014 per variation**
- **Retention**
  - Alpha = 5%/2 = 2%
  - Beta = 0.2, so 1-beta = 80%
  - Probability = 0.53 (Baseline conversion rate)
  - Minimum Detectable Effect: 0.01
  - **Sample size = 50,013 per variation**
- **Net conversion**
  - Alpha = 5%/2 = 2%
  - Beta = 0.2, so 1-beta = 80%
  - Probability = 0.1093 (Baseline conversion rate)
  - Minimum Detectable Effect: 0.0075
  - **Sample size = 35,013 per variation**

For Gross conversion and Net conversion, the unit of diversion is # of click. The probability of click is 0.08. Also, we have two groups here, one is control group, one is experiment group.
Formula: Pageview * probability of click / # of group = sample size. Then we solve for pageview.

For Retention, the unit of diversion is # of enrol. The probability of enroll, given click is 0.20625. Probability of click is 0.08. We have two groups here, one is control group, one is experiment group.
Formula: Pageview * probability of click * probability of enroll given click / # of group = sample size. Then we can solve for pageview for Retention.

- **Gross conversion** pageview = 33,014 / 0.08 * 2 = 825,350
- **Retention** pageview = 50,013 / 0.08 / 0.20625 * 2 = 6,062,182
- **Net conversion** = 35,013 / 0.08 * 2 = 875,325

But we know the site will only have 40,000 pageview per day. If we keep retention metric, we need 6,062,182 pageviews, which takes 152 days. This is too long for the test. So we have to take down this metic. Only keep Gross conversion and net conversion in our experiment. So we take the larger one, which is 875,325.

We need Bonferroni correction since Gross conversion and Net conversion metric are not independent.
And the sample size is the larger one from the above, which is 875,325.

Choosing Duration vs. Exposure

Exposure:
I'd like to diver 60% of the traffic to the experiment, which means 30% to the experiment group and 30% to the control group and leaving the 40% unchanged. The change will affect how people sign up for the courses, and eventually affect how revenue is generated due to the sign-up. So i think it's risky enough that we wouldn't want to run on all traffic.

Duration Calculation:
Percentage I choose: 60%
Maximum pageview per day for the site: 40,000 (from the baseline file)
Duration = 876,325 / 40,000 / 0.6 = 37 days.

If 37 days seems too long for the test, and we are fine with diver 100% of the traffic to the experiment, we can shrink down the duration to **22 days.**

Analysis

Sanity Checks

Invariant metric:

- **Number of cookies (count):**
  - Control group total(N1): 345,543
  - Experiment group total(N2): 344,660
  - Probability in each group is: 0.5
  - SD = sqrt(P*(1-P)/(N1 + N2)) = 0.00060184
  - Lower bound = 0.5 - 1.96 * SD = 0.4988
  - Upper bound = 0.5 + 1.96 * SD = 0.5012
  - P^ = N1/(N1+N2) = 0.5006
  - Since P^ is inside the 95% CI, so number of cookies metric pass the sanity check
- **Number of clicks (count):**
  - Control group total(N1): 28,378
  - Experiment group total(N2): 28,325
  - Probability in each group is: 0.5
  - SD = sqrt(P*(1-P)/(N1 + N2)) = 0.0021
  - Lower bound = 0.5 - 1.96 * SD = 0.4959
  - Upper bound = 0.5 + 1.96 * SD = 0.5041
  - P^ = N1/(N1+N2) = 0.5005
  - Since P^ is inside the 95% CI, so number of cookies metric pass the sanity check
- **Click-through-probability**
  - Control group click-through-probability: 0.08212581357
  - Experiment group click-through-probability: 0.08218244067
  - SD_con = sqrt(P*(1-P)/(N1 + N2)) = 0.0004670682766
  - Lower bound = 0.08213 - 1.96 * SD = 0.08121035975
  - Upper bound = 0.08213 + 1.96 * SD = 0.0830412674
  - Since the experiment group click-through-probability is inside the 95% CI, number of cookies metric pass the sanity check.

Check for Practical and Statistical Significance

Next, calculate a confidence interval for the difference between the experiment and control groups, and check whether each metric is statistically and/or practically significance.

Result Analysis

Effect Size Tests

Evaluation metrics: before doing Bonferroni Correction.

- **Gross conversion:**
  - Control group # of click (N1): 17,293
  - Experiment group # of click (N2): 17,260
  - Control group # of enroll (X1): 3,785
  - Experiment group # of enroll (X2): 3,423
  - P1 = 3785/17293 = 0.2189
  - P2 = 3423/17260 = 0.1983
  - P_pool = (X1 + X2)/(N1 + N2) = 0.2086
  - SE_pool = sqrt(p_pool * (1-p_pool)*(1/N1 + 1/N2)) = 0.0044
  - P_different = P2 - P1 = -0.0206
  - Lower bound: P_different -1.96 * SE_pool = -0.0291
  - Upper bound: P_different + 1.96 * SE_pool = -0.0120
  - Since zero is not inside the 95% CI, so it's statistically significant saying that there is a 95% CI the control and experimental group results are not the same.
  - Also, the dmin = 0.01. -dmin = -0.01, they are not within the 95% CI, so we can see that the test is also practically significant.
- **Net conversion:**
  - Control group # of click (N1): 17,293
  - Experiment group # of click (N2): 17,260
  - Control group # of payments (X1): 2,033
  - Experiment group # of payments (X2): 1,945
  - P1 = 3785/17293 = 0.1176
  - P2 = 3423/17260 = 0.1127
  - P_pool = (X1 + X2)/(N1 + N2) = 0.1151
  - SE_pool = sqrt(p_pool * (1-p_pool)*(1/N1 + 1/N2)) = 0.0034
  - P_different = P2 - P1 = -0.0049
  - Lower bound: P_different -1.96 * SE_pool = -0.0116
  - Upper bound: P_different + 1.96 * SE_pool = 0.0019
  - Since zero is inside the 95% CI, so it's not statistically significant, which means we can't reject control and experiment group are the same.
  - Also, the dmin = 0.0075. -dmin = -0.0075, they are within the 95% CI, so we can see that the test is also not practically significant.

Evaluation metrics: after doing Bonferroni Correction. 5% /2 = 2.5%, so z-score = 2.24

- **Gross conversion:**
  - Control group # of click (N1): 17,293
  - Experiment group # of click (N2): 17,260
  - Control group # of enroll (X1): 3,785

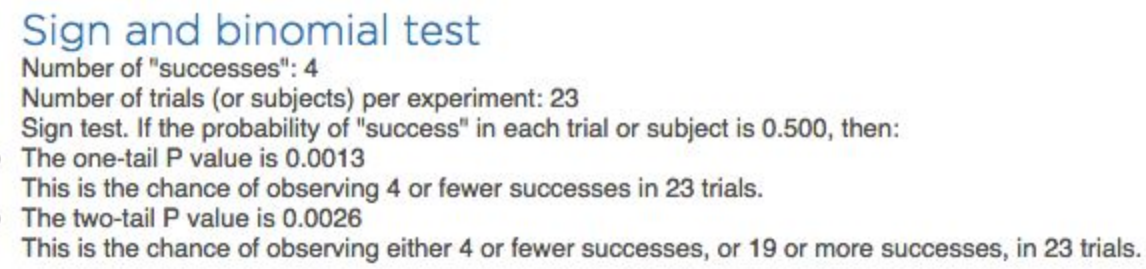
    - Experiment group # of enroll (X2): 3,423
    - P1 = 3785/17293 = 0.2189
    - P2 = 3423/17260 = 0.1983
    - P_pool = (X1 + X2)/(N1 + N2) = 0.2086
    - SE_pool = sqrt(p_pool * (1-p_pool)*(1/N1 + 1/N2)) = 0.0044
    - P_different = P2 - P1 = -0.0206
    - Lower bound: P_different -2.24 * SE_pool = -0.0303
    - Upper bound: P_different + 2.24 * SE_pool = -0.0108
    - Since zero is not inside the 95% CI, so it's statistically significant saying that there is a 95% CI the control and experiment group results are not the same.
    - Also, the dmin = 0.01. -dmin = -0.01, they are not within the 95% CI, so we can see that the test is also practically significant.
- **Net conversion:**
    - Control group # of click (N1): 17,293
    - Experiment group # of click (N2): 17,260
    - Control group # of payments (X1): 2,033
    - Experiment group # of payments (X2): 1,945
    - P1 = 3785/17293 = 0.1176
    - P2 = 3423/17260 = 0.1127
    - P_pool = (X1 + X2)/(N1 + N2) = 0.1151
    - SE_pool = sqrt(p_pool * (1-p_pool)*(1/N1 + 1/N2)) = 0.0034
    - P_different = P2 - P1 = -0.0049
    - Lower bound: P_different -2.24 * SE_pool = -0.0126
    - Upper bound: P_different + 2.24 * SE_pool = 0.0028
    - Since zero is inside the 95% CI, so it's not statistically significant, which means we can't reject control and experiment group are the same.
    - Also, the dmin = 0.0075. -dmin = -0.0075, they are within the 95% CI, so we can see that the test is also not practically significant.

Run Sign Tests

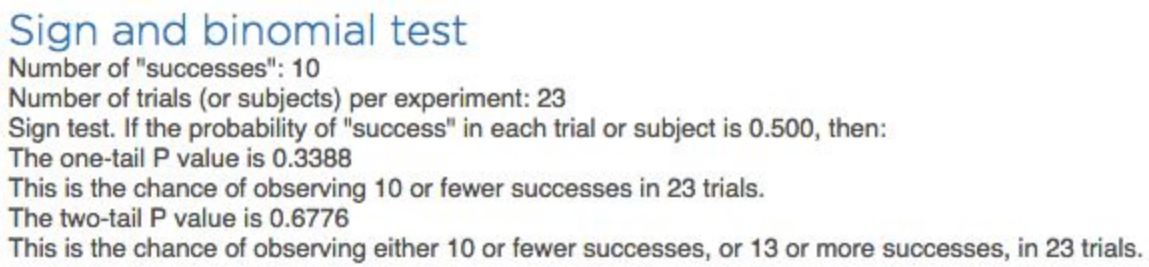
Still want to run a Bonferroni correction. So we should compare the p-value to 5%/2 = 0.025.

User this link to do the sign test: https://www.graphpad.com/quickcalcs/binomial1.cfm

- **Gross conversion:** compared the day-by-day enrollment/click rate, we see we have 4 days that experiment group is higher than the control group. There are 23 days of data. The p-value is: 0.0026, which is less than 0.025. It's statistically significant.



- **Net conversion:** compared the day-by-day enrollment/click rate, we see we have 10 days that experiment group is higher than the control group. There are 23 days of data. The p-value is: 0.6776, which is larger than 0.025. It's not statistically significant.



I will run a Bonferroni correction since we can't assume the two metrics are independent. When two metrics are not independent, we might see high false positive. Let's do it in conservative way. But both the effect size hypothesis and sign tests indicate that gross conversion is statistically and practically significant, but net conversion is not. That's the gross conversion rate is tend to be lower after the change, while the net conversion rate doesn't change.

Make a Recommendation

I would not recommend to launch this feature to add "5 or more hours" recommandation. In the A/B test result we see that the gross conversion drop statistically significant after the change, which makes sense since for students they aren't committed yet, they aren't likely to sign-up for the trail. But we hope this will increase the net conversion rate. However, the result shows that we can't statistically prove that net conversion rate will change after the new feature. So the change won't provide us any benefit in driving more students stay in the course after the trial expired, meaning getting more paid users. So I don't think we should launch this feature.

A/B Testing Steps Summary:

1. Define the hypothesis
2. Choose invariant metrics an evaluation metrics
3. Calculate the standard deviation for the evaluation metrics
4. Calculate sample size, and # of pageviews needed
5. Calculate duration and exposure
6. Sanity check for invariant metrics, make sure they don't change statistically significant
7. Check effect size statistically significant and practically significant
8. Do size test (non-parametric test)
9. Conclusion!