Online Free Trial Screener A/B Testing

Experiment Overview

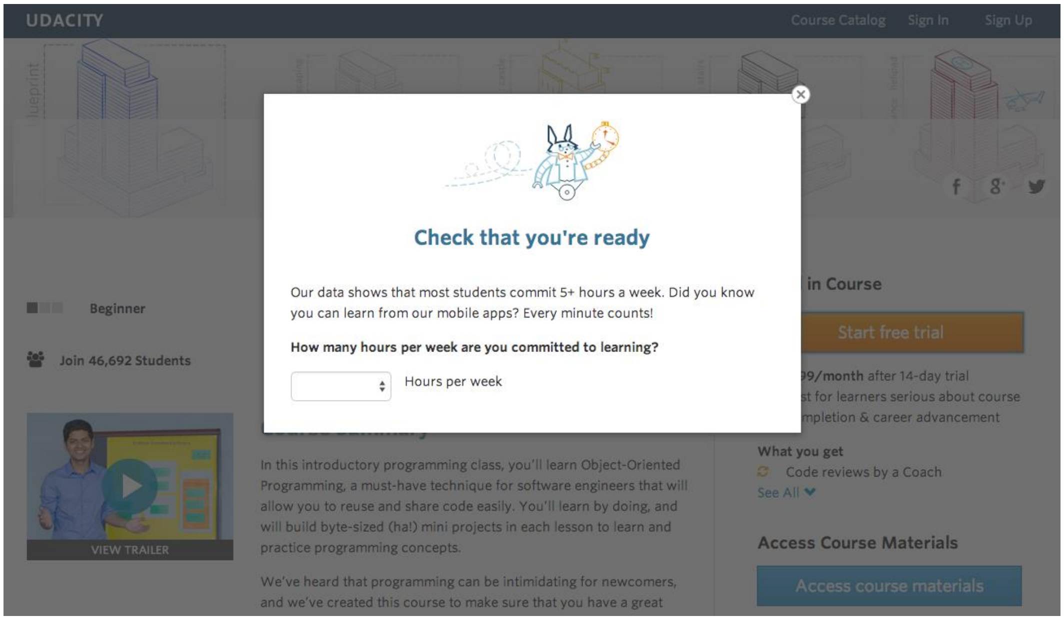
At the time of this experiment, Udacity courses currently have two options on the course overview page: "start free trial" and "access course materials".

- 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.

Experiment Description

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.



Check that you're ready

Our data shows that most students commit 5+ hours a week. Did you know you can learn from our mobile apps? Every minute counts!

How many hours per week are you committed to learning?

Hours per week

99/month after 14-day trial
st for learners serious about course
mpletion & career advancement

What you get
Code reviews by a Coach
See All

In this introductory programming class, you'll learn Object-Oriented Programming, a must-have technique for software engineers that will allow you to reuse and share code easily. You'll learn by doing, and will build byte-sized (ha!) mini projects in each lesson to learn and practice programming concepts.

We've heard that programming can be intimidating for newcomers, and we've created this course to make sure that you have a great

Experiment Setup

Primary Aim: improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

Null Hypothesis: not make a significant change and might not be effective in reducing the early Udacity course cancellation.

Alternative Hypothesis: reduce the number of frustrated students who left the free trial because they did not have enough time, without significantly reducing the number of students to continue past the free trial and eventually complete the course.

Experiment Design

Unit of Diversion

Cookies. 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: used for sanity checks and will remain invariant throughout the experiment.

- Number of cookies: number of unique cookies to view course overview page (d_{min}=3000).
- Number of clicks: number of unique cookies to click the "Start free trial" button which happens before the free trial screener is trigger (dmin=240).
- 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 (d_{min}=0.01).

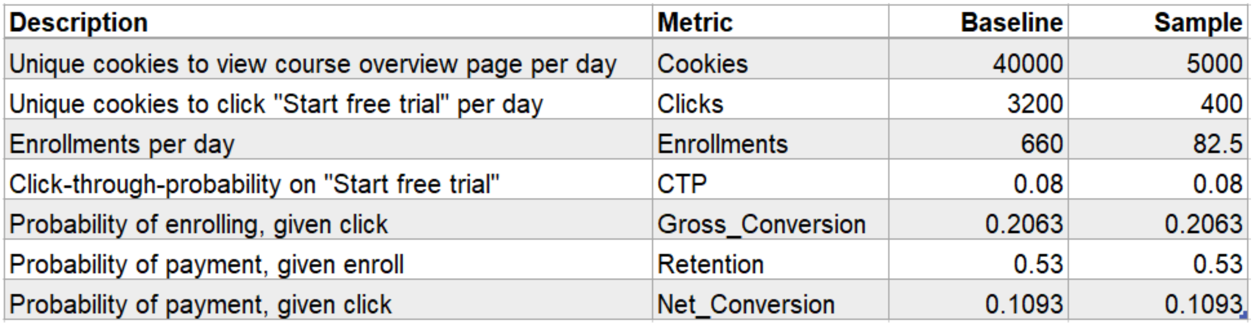
Evaluation metrics: must be observed for consideration in the decision to launch the experiment.

- 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 (dmin=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).

Unused metrics:

- Number of user-ids: who enroll in the free trial (d_{min}=50). The number of user-ids would be lower in the experiment group, since 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. This metric is not normalized.

Measuring Standard Deviation



Description	Metric	Baseline	Sample
Unique cookies to view course overview page per day	Cookies	40000	5000
Unique cookies to click "Start free trial" per day	Clicks	3200	400
Enrollments per day	Enrollments	660	82.5
Click-through-probability on "Start free trial"	CTP	0.08	0.08
Probability of enrolling, given click	Gross_Conversion	0.2063	0.2063
Probability of payment, given enroll	Retention	0.53	0.53
Probability of payment, given click	Net_Conversion	0.1093	0.1093

Udacity provided the baseline values for each metric and given a sample size of 5000 cookies visiting the course overview page, for each metric selected as an evaluation metric, I made an analytic estimated of its standard deviation (SE):

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- **Gross Conversion:** n = 5000*0.08 = 400, p = 0.20625, SE = 0.0202
- **Retention:** n = 5000*(660/40000) = 82.5, p = 0.53, SE = 0.0549
- **Net Conversion:** n = 5000*0.08 = 400, p = 0.109313, SE = 0.0156

Sizing

Number of Samples vs. Power

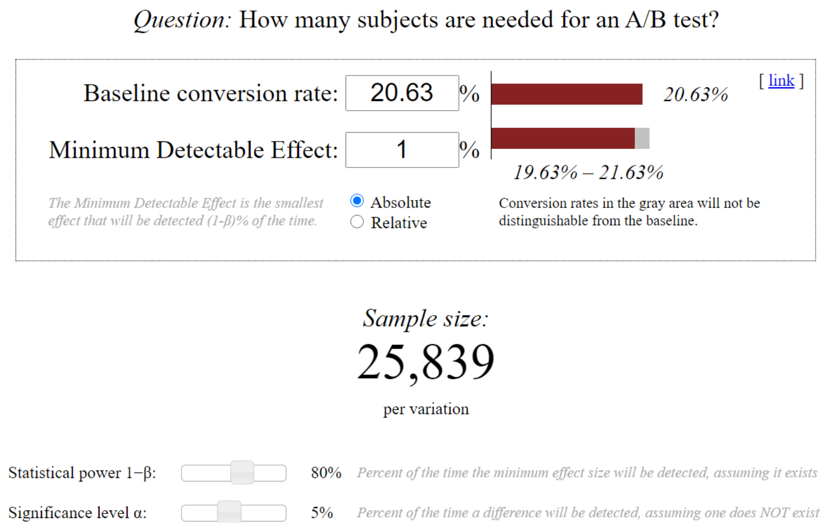
Given alpha = 0.05 and beta = 0.2 for selected evaluation metric, the sample size is based on the distribution of metrics, alpha (significance level), beta (false negative probability or type II error), baseline conversion rate and minimum detectable effect (practical significance or dmin).

Here I do not use Bonferroni Correction in the analysis phase, since the evaluation metrics in this experiment are closely related to each other, Bonferroni Correction would be too conservative.

According to the following conditions, every evaluation metric (gross conversion, retention, net conversion) is binomial distribution.

- Outcome either success or failure
- Independent events
- Constant probability
- Fixed number of events

By using the online calculator, I got the sample size 25,839 for each group (e.g. gross conversion):



Question: How many subjects are needed for an A/B test?

Baseline conversion rate: 20.63 %

Minimum Detectable Effect: 1 %

The Minimum Detectable Effect is the smallest effect that will be detected (1-β)% of the time.

● Absolute
○ Relative

Conversion rates in the gray area will not be distinguishable from the baseline.

Sample size:

25,839

per variation

Statistical power 1−β: 80% *Percent of the time the minimum effect size will be detected, assuming it exists*

Significance level α: 5% *Percent of the time a difference will be detected, assuming one does NOT exist*

- **Gross Conversion**

  Baseline Conversion: 20.63%
  Minimum Detectable Effect: 1%
  Alpha: 5%
  Beta: 20%
  Statistical Power or Sensitivity (1-beta): 80%
  Sample Size = 25839 enrollments/group
  Number of groups = 2 (experiment and control)
  Total sample size = 25839*2 = 51,678 enrollments
  Clicks/Pageview: 3200/40000 = 0.08 clicks/pageview
  Pageviews = 51678/0.08 = 645,975

- **Retention**

  Baseline Conversion: 53%
  Minimum Detectable Effect: 1%
  Alpha: 5%
  Beta: 20%
  Statistical Power or Sensitivity (1-beta): 80%
  Sample size = 39155 enrollments/group
  Number of groups = 2 (experiment and control)
  Total sample size = 39155*2 = 78230 enrollments
  Enrollments/pageview: 660/40000 = 0.0165 enrollments/pageview
  Pageviews = 78230/0.0165 = 4,741,212

- **Net Conversion**

  Baseline Conversion: 10.93%
  Minimum Detectable Effect: 0.75%
  Alpha: 5%
  Beta: 20%
  Statistical Power or Sensitivity (1-beta): 80%
  Sample size = 27413 enrollments/group

Number of groups = 2 (experiment and control)
Total sample size = 27413*2 = 54826 enrollments
Clicks/pageview: 3200/40000 = 0.08 clicks/pageview
Pageviews = 54826/0.08 = 685,325

Pageviews require the maximum of pageviews, in order to have enough power for each evaluation metric (gross conversion, retention and net conversion). Therefore, the required pageviews is 4,741,212.

Duration vs. Exposure

Q: What percentage of Udacity's traffic would you divert to this experiment (assuming there were no other experiments you wanted to run simultaneously)? Is the change risky enough that you would not want to run on all traffic?

> Based on the required pageviews 4,741,212, if divert 100% of traffic, given 40,000 pageviews per day, the experiment would take 119 days, which is too long and presents both a business risk, potential for frustrated students, lower conversion and retention, inefficient use of coaching resources and an opportunity risk (performing other experiments).

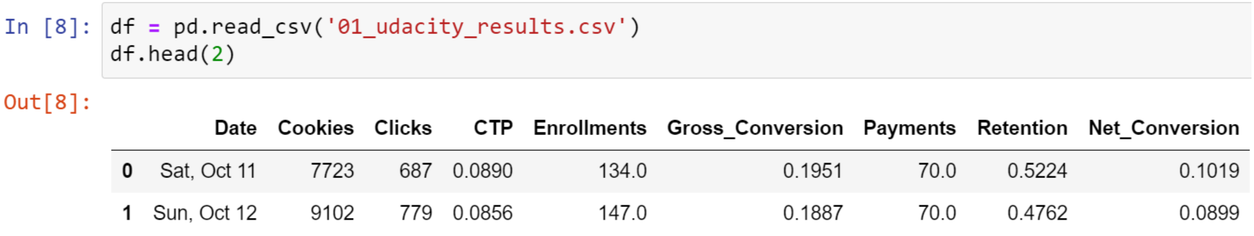
Q: Given the percentage you chose, how long would the experiment take to run, using the analytic estimates of variance? If the answer is longer than a few weeks, then this is unreasonably long, and you should reconsider an earlier decision.

> Since the experiment indicates that payments are made 14 days after enrollment, I expect it to run for at least 14 days. Therefore, I choose to reduce the required pageviews to 685,325 and an 18-day experiment with 100% diversion or 35-day given 50% diversion. For Gross Conversion and Net Conversion, if use 100% of the traffic, the experiment will take 18 days and the duration is short. If divert 80% of the traffic, the experiment will last 21 days. In general, an 18-day experiment is more reasonable, but percentage diversion may be scaled down depending on other experiments of interest to be performed concurrently.

Experiment Analysis

Sanity Checks

Sanity check is primarily for invariant metrics, including number of pageviews (cookies), number of clicks (on "Start free trial"), click-through-probability (clicks/pageviews). It should expect equal diversion into control and experiment groups, the data for analysis can be found here.

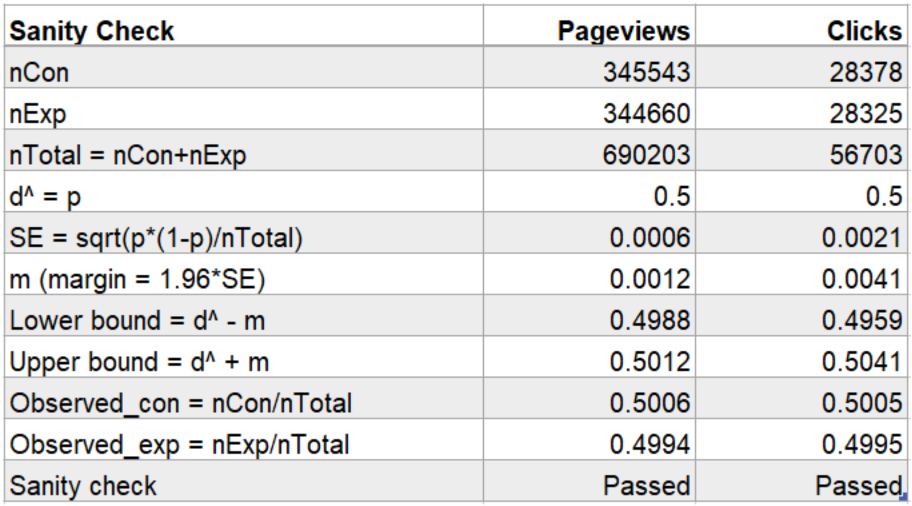


```
In [8]: df = pd.read_csv('01_udacity_results.csv')
        df.head(2)
```

Out[8]:

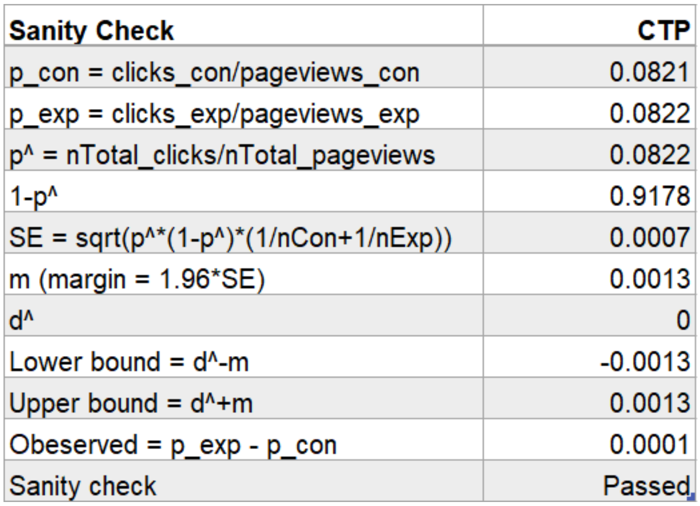
	Date	Cookies	Clicks	CTP	Enrollments	Gross_Conversion	Payments	Retention	Net_Conversion
0	Sat, Oct 11	7723	687	0.0890	134.0	0.1951	70.0	0.5224	0.1019
1	Sun, Oct 12	9102	779	0.0856	147.0	0.1887	70.0	0.4762	0.0899

After evaluated the data, I suppose all three invariant metrics are binomial distribution. According to the following information, we can see observed probability of both pageviews and clicks are within their intervals, given 95% of confidence interval. Therefore, these two metrics passed the sanity check.



Sanity Check	Pageviews	Clicks
nCon	345543	28378
nExp	344660	28325
nTotal = nCon+nExp	690203	56703
d^ = p	0.5	0.5
SE = sqrt(p*(1-p)/nTotal)	0.0006	0.0021
m (margin = 1.96*SE)	0.0012	0.0041
Lower bound = d^ - m	0.4988	0.4959
Upper bound = d^ + m	0.5012	0.5041
Observed_con = nCon/nTotal	0.5006	0.5005
Observed_exp = nExp/nTotal	0.4994	0.4995
Sanity check	Passed	Passed

Similarly, I calculated the lower and upper bounds of click-through-probability, the interval includes 0, which means there is no significant difference between control and experiment groups. CTP passed the sanity check.

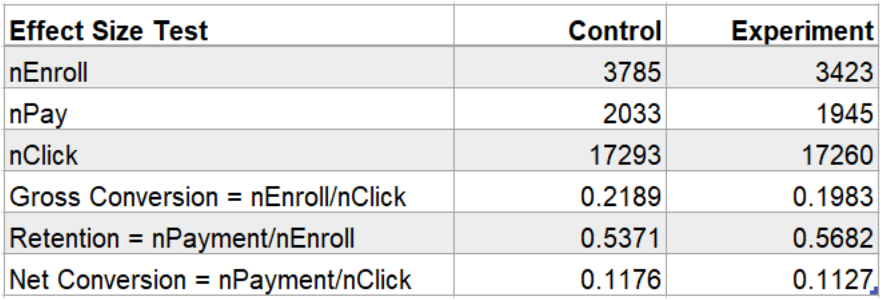


Sanity Check	CTP
p_con = clicks_con/pageviews_con	0.0821
p_exp = clicks_exp/pageviews_exp	0.0822
p^ = nTotal_clicks/nTotal_pageviews	0.0822
1-p^	0.9178
SE = sqrt(p^*(1-p^)*(1/nCon+1/nExp))	0.0007
m (margin = 1.96*SE)	0.0013
d^	0
Lower bound = d^-m	-0.0013
Upper bound = d^+m	0.0013
Obeserved = p_exp - p_con	0.0001
Sanity check	Passed

Result Analysis

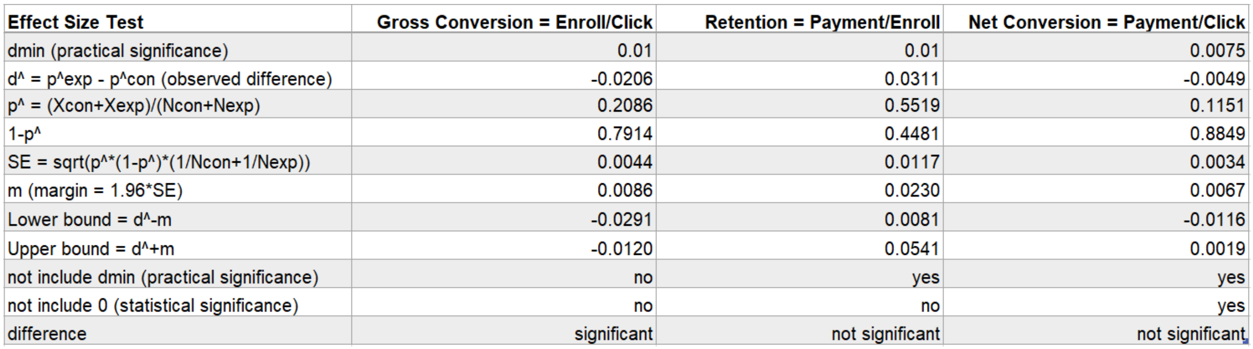
Effect Size Tests

Since all sanity checks were passed, it is time to decide whether Udacity should launch the experiment or not. Before that, we need to test is there a statistical or practical significant difference in each of evaluation metrics (Gross Conversion, Retention, Net Conversion) between control and experiment groups.



Effect Size Test	Control	Experiment
nEnroll	3785	3423
nPay	2033	1945
nClick	17293	17260
Gross Conversion = nEnroll/nClick	0.2189	0.1983
Retention = nPayment/nEnroll	0.5371	0.5682
Net Conversion = nPayment/nClick	0.1176	0.1127

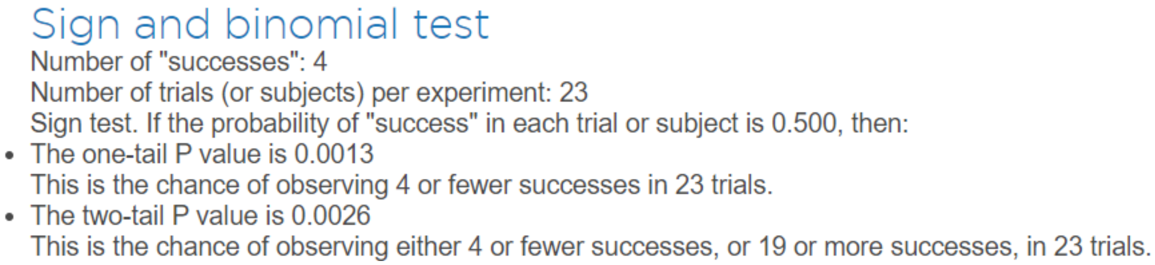
Based on the above summarized data, I got the confidence intervals for all metrics. A metric is statistically significant if the confidence interval does not include 0 (can be confident there was a change) and it is practically significant if the confidence interval does not include the practical significance boundary (can be confident there is a change that matters to the business). From the following table, we can see only gross conversion is significant different, there is no big difference in either retention or net conversion.



Effect Size Test	Gross Conversion = Enroll/Click	Retention = Payment/Enroll	Net Conversion = Payment/Click
dmin (practical significance)	0.01	0.01	0.0075
d^ = p^exp - p^con (observed difference)	-0.0206	0.0311	-0.0049
p^ = (Xcon+Xexp)/(Ncon+Nexp)	0.2086	0.5519	0.1151
1-p^	0.7914	0.4481	0.8849
SE = sqrt(p^*(1-p^)*(1/Ncon+1/Nexp))	0.0044	0.0117	0.0034
m (margin = 1.96*SE)	0.0086	0.0230	0.0067
Lower bound = d^-m	-0.0291	0.0081	-0.0116
Upper bound = d^+m	-0.0120	0.0541	0.0019
not include dmin (practical significance)	no	yes	yes
not include 0 (statistical significance)	no	no	yes
difference	significant	not significant	not significant

Sign Tests

Sign test is another method to validate the results, here I use the online tool to run the tests. Suppose the hypothetical probability of success in each trial is 0.5, because if there is no difference, it should be 50% chance of positive change in each trail.



Sign and binomial test

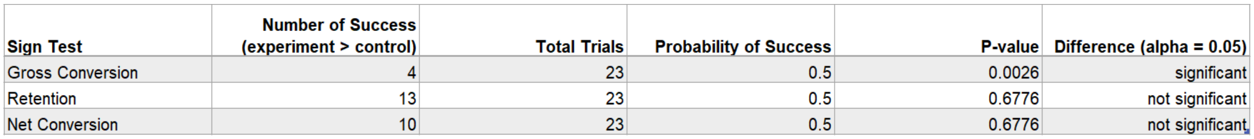
Number of "successes": 4

Number of trials (or subjects) per experiment: 23

Sign test. If the probability of "success" in each trial or subject is 0.500, then:

- The one-tail P value is 0.0013
  This is the chance of observing 4 or fewer successes in 23 trials.
- The two-tail P value is 0.0026
  This is the chance of observing either 4 or fewer successes, or 19 or more successes, in 23 trials.

According to the Udacity data, there are total 23 trials in each group and I observed rates from experiment group are higher than control group for 4 times in gross conversion, 13 times in retention and 10 times in net conversion. The p-value is as below:



Sign Test	Number of Success (experiment > control)	Total Trials	Probability of Success	P-value	Difference (alpha = 0.05)
Gross Conversion	4	23	0.5	0.0026	significant
Retention	13	23	0.5	0.6776	not significant
Net Conversion	10	23	0.5	0.6776	not significant

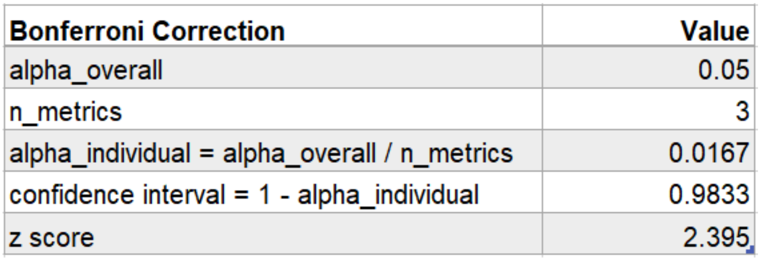
$$p\text{-value} < \alpha \Rightarrow \text{reject } H_0 \Rightarrow \text{accept } H_\text{a}$$

$$p\text{-value} \geq \alpha \Rightarrow \text{fail to reject } H_0$$

Given alpha = 0.05, only gross conversion p-value = 0.0026 < 0.05, we can say there is a significant difference in this metric, but not for the other two.

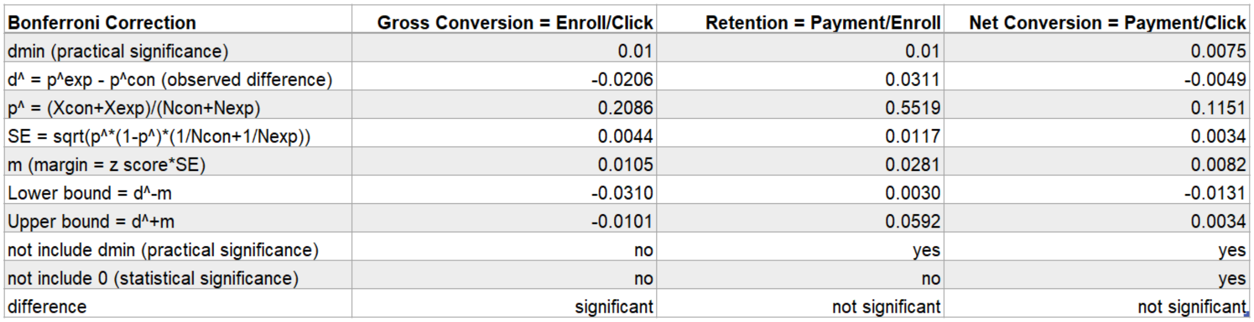
Bonferroni Correction

The Bonferroni Correction is a method for controlling for type I errors (false positives) when using multiple metrics in which relevance of ANY of the metrics matches the hypothesis, here I use it to test three evaluation metrics.



Bonferroni Correction	Value
alpha_overall	0.05
n_metrics	3
alpha_individual = alpha_overall / n_metrics	0.0167
confidence interval = 1 - alpha_individual	0.9833
z score	2.395

Based on the new z score = 2.395, the analysis results are same as effect size test and sign test. Compare with control and experiment groups, there is a significant difference only in gross conversion.



Bonferroni Correction	Gross Conversion = Enroll/Click	Retention = Payment/Enroll	Net Conversion = Payment/Click
dmin (practical significance)	0.01	0.01	0.0075
d^ = p^exp - p^con (observed difference)	-0.0206	0.0311	-0.0049
p^ = (Xcon+Xexp)/(Ncon+Nexp)	0.2086	0.5519	0.1151
SE = sqrt(p^*(1-p^)*(1/Ncon+1/Nexp))	0.0044	0.0117	0.0034
m (margin = z score*SE)	0.0105	0.0281	0.0082
Lower bound = d^-m	-0.0310	0.0030	-0.0131
Upper bound = d^+m	-0.0101	0.0592	0.0034
not include dmin (practical significance)	no	yes	yes
not include 0 (statistical significance)	no	no	yes
difference	significant	not significant	not significant

Recommendation

This experiment is designed to understand whether the free trail screener will help to filter out students who would not commit to the study time, without reducing the number of students who will make the payments after 14 days free trial.

The results show that Gross Conversion will be reduced significantly, which means when screener requires student to input enough study time before enrolling, it does help filter out significant number of students who are not ready for studying.

However, there are no significant changes in Retention and Net Conversion, which means the screener only help reduce the enrollments, but not enough evidence to show that more students will stay till the end of the free trial and trigger the final payments.

Therefore, I would not recommend to launch this screener, but rather to pursue other experiments.

Follow-Up Experiment

In order to increase the number of final payments, we want students to be clear about that the course is worth the money and is going to be very helpful to get a job.

- Is the tuition competitive?
- Any available financial aid?
- If students regret and cancel, how much money can they get refund?
- What accredited certificates can students get after finishing the course?
- Any discount on bundle sales, like coach service or professional badges?
- Is this a high-quality course and are the instructors professional?
- How is the course review and what are the comments from other students?
- What pre-requisite requirements can help students achieve the goal?
- What job do students want to apply, is this course going to be the right help?

We can add these information after the enrollment and setup a new experiment:

- Setup: randomly assign enrollment students into control and experiment groups.
- Null Hypothesis: no significant increase in retention and payments after 14 days free trail.
- Unit of diversion: user-ids, since the changes take place after the student creates an account and enrolls in a course.
- Invariant Metrics: number of user-ids, number of clicks, click-through-probability.
- Evaluation Metrics: retention, net conversion.

If there are statistically and practically significant positive changes in retention and net conversion, then we would consider to launch the experiment.