Udacity A/B Test of Free Trial Screener

Vincent Zhang

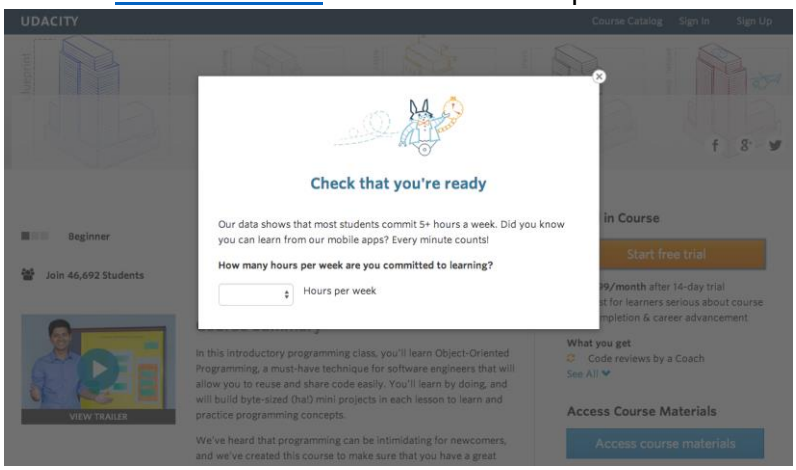
07/27/2018

- Experiment Overview
- Experiment Design
    - User Behavior Process
    - Metric Choice
        - Invariant Metrics
        - Variant Metrics
        - Explanation of Metrics Selection
        - Expected Results to Launch the Experiment
    - Measuring Standard Deviation
        - Analytical Estimation
        - Compare Analytics and Empirical Variability
    - Sizing
        - Number of Samples vs. Power
        - Duration vs. Exposure
            - Risk of Experiment
- Experiment Analysis
    - Sanity Check
    - Result Analysis
        - Effect Size Tests
        - Sign Tests
        - Recommendation

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.

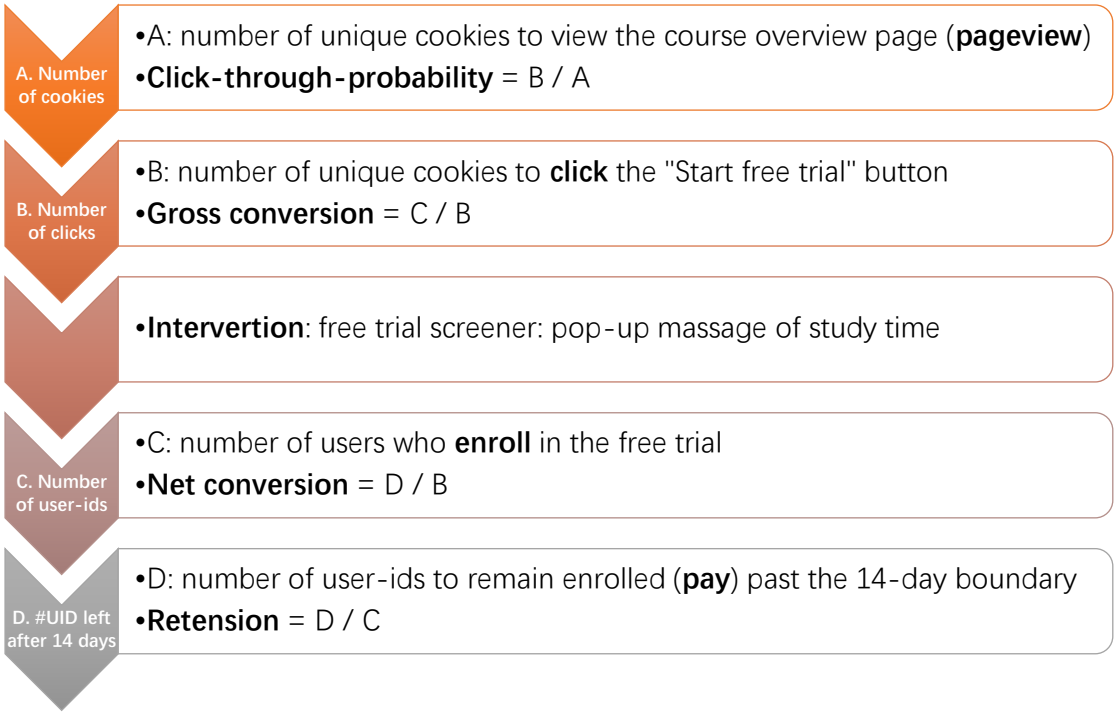
In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 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.

User Behavior Process



Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here.

Invariant Metrics

Invariant metrics are identical in both the experiment and control groups. In this case, invariant metrics are variables that calculated before starting the experiment—before clicking “Start free trial” button.

- **Number of cookies**
- **Number of clicks**
- **Click-through-probability**

Evaluation Metrics

Evaluation metrics measured the outcome of experiments with pre-set targets.
They might differently distribute in the experiment and control groups. There should be a minimum difference to consider whether the changed experiment should be launched or not.

- **Gross conversion**
- ~~**Retention**~~ *should be ignored after checking the sizing considering Retention*
- **Net conversion**

Explanation of Metrics Selection

Method of categorizing invariant metrics & evaluation metrics:
how the metric measured is related to the intervention. Metrics measured prior to an intervention are considered as invariant metrics. Metrics measured after an intervention can evaluate the result of experiments but might not be expected as appropriate evaluation metrics. We should also consider the way to gather relative data (how long to collect evaluation metric?) and advantages of numerous data types (raw count, ratio, probability, proportion, rate).

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

Invariant metric because it is consistent in the experiment. Cookie is a unit of diversion, which should not be changed all the time. The intervention of this experiment is the free trial screener and “number of cookies” has already been collected once users view the course overview page, which is prior to the intervention. So that “numbers of cookies” are identical for both the experiment and control group.

Number of user-ids: That is, number of users who enroll in the free trial. (d_{min}=50)

Not invariant because it is dependent on our experiment that is related to the Free trial screener. It measures the number for two groups after the intervention, but it is not a good evaluation metric. Although we check the sanity whether we have a proper same sized group, it is possible that we get some skewed data. So that if we have different numbers of cookies in two groups, the raw count can’t measure the result as we expect. Thus, we need a ratio type metric normalized by cookies. “Gross Conversion” (enrollments/cookies) is better than “number of user-ids”(enrollments). Therefore, “number of user-ids” is not an evaluation metric.

Number of clicks: That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is a trigger). (d_{min}=240)

Invariant metric because the reason is like that of “number of cookies”. Metric measured prior to the intervention.

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 metric because it is measured by the number of cookies and the number of clicks and both are invariant metrics. Explain in another way that CTP is for “Start free trial” button and behavior that clicking this button happens before the intervention.

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)

Evaluation metric because it is dependent in our experiment. It is a refined version of “number of user-ids” that normalized by “number of cookies”. A lower result can convince us that the tested intervention is effective in reducing the proportion of users to enroll in the free trial.

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)

~~Evaluation metric because~~ it is dependent in our experiment. *However, we should discard this evaluation metric because of the difficulty to calculate this metric with a long duration almost 119 days.*

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)

Evaluation metric because it is dependent in our experiment. We are interested in how many users left after the 14-days free trial.

Expected Results to Launch the Experiment

Udacity aims to 1) significantly reduce the number of non-profit frustrated users who left the 14-days free trial to save limited coaching resources and 2) not significantly reducing the number of students to continue past the free trial and eventually complete the course. The pop-up message of study time is designed for these two targets.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics.

Choices of users are independent whether choose or not so that it is Bernoulli experiment with population n and probability p, Mean(X) = p, Var(X) = p(1-p).
* *Bernoulli experiment is the Binomial trial that conducts once*

A standard error (SE) is really the standard deviation in a point estimate.
SE standard error of sample mean = σ / sqrt(n)
SE_Bernoulli = sqrt(p * (1-p) / n).

Given the following table:

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

Analytical Estimation of Standard Deviation given 5000 cookies per day

Gross conversion:
p = 0.20625
N = 5000 * 0.08 = 400
SE = sqrt (0.20625 * (1-0.20625) / 400) = 0.0202

Retention*: should be ignored after checking the sizing considering Retention*
~~*p = 0.53*~~
~~*N = 5000 * 0.08 * 0.20625 = 82.5*~~
~~*SE = sqrt (0.53 * (1-0.53) / 82.5) = 0.0549*~~

Net conversion:
p = 0.1093125
N = 5000 * 0.08 = 400
SE = sqrt (0.1093125* (1-0.1093125) / 400) = 0.0156

Compare Analytics and Empirical Variability of Evaluation Metrics

Above all, each experiment has only one unit of diversion, which is a subject when it comes to splitting users into experimental and control groups. However, our metrics may have different units of analysis, represented by the denominator of the ratio when computing the metric average among subjects. Thus, it is possible that some metrics' units of analysis are different from the unit of diversion. The difference means the analytic variance is not appropriate to describe our metrics and we should compute an empirical variance instead.

The similarity condition between analytics variability and empirical variability is **unit of analysis = unit of diversion**. For the "gross conversion" and the "net conversion", their denominators (unit of analysis) are "number of cookies", which is same as the unit

of diversion in the experiment. Thus, analytics estimate is likely to be the empirical variability.

Retention should be ignored after checking the sizing considering Retention
~~*However, the unit of analysis of the retention is the number of user-ids, which is different from the unit of diversion (number of cookies). Thus, analytics estimate is not close to empirical variability.*~~

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.

Can we apply Bonferroni correction?

Not apply Bonferroni correction.

Bonferroni correction is to address a problem that multiple tests lead to the increasing proportion of the false positive rate that rejecting the null hypothesis given H0 is true. In other words, having at least one type I error increases to $1-(1-\alpha)^m$ as the growth of the number of experiments m (in this case, we have two evaluation metrics, referring two experiments). To avoid a lot of false positives, the α needs to be reduced to account for the number of comparisons being performed, which is Bonferroni correction. $\alpha\_new = \alpha\_old / n$; n is number of tests/experiments/metrics.

The reduction in the false positive rate α results in the increase of false negative rate β and the reduction in the power $1-\beta$ (the power of any test of statistical significance is defined as the probability that it will reject the null hypothesis given H0 is false).

In term of the Free Trial Screener, “gross conversion” and “net conversion” are highly correlated since “gross conversion” is the base of “net conversion”. It is too conservative (means overly reduce the α beyond our expectation) if apply the correction. *Image an extreme case that two metrics are perfectly positive correlated (r=1), which indicates that they are represented as one metric. If we apply Bonferroni correction, the false positive rate will be reduced by 1/2 since we have two hypothesis tests. But the result is not what we want, the original false positive rate is the proper result.* In order to use Bonferroni correction, each experiment(metrics) should be independent, which can lower the conservative effect comparing with the correction on correlated metrics.

*Additional note: Bonferroni correction is only appropriate for “ANY” case, not “ALL” case. We can launch the experiment if ANY (at least one) metric satisfies our launch criteria instead of recommending the change only if ALL metric satisfies the pre-set criteria.

Number of pageview required to power the experiment appropriately

Pageviews are calculated by [online A/B test calculator: subjects required for A/B test] with **significant level** $\alpha = 0.05$ and statistical **power** $1 - \beta = 0.8$ ($\beta = 0.2$). The result of an online calculator is the sample size and we should convert it to the pageviews of the experiment group and time two to get total pageviews.

Baseline conversion & minimum detectable are given

	Gross conversion	Retention	Net conversion
Baseline conversion	20.625%	53%	10.93%
Minimum detectable	1%	1%	0.75%
Sample size	25,835	39,115	27,411
Experiment pageviews	25835/0.08 = 322,938	39115/0.20625/0.08 = 2,370,606	27413/0.08 = 342,663
Total pageviews	322938 * 2 = 645,875	2370606 * 2 = 4,741,212	342663 * 2 = 685,325

The total pageviews is the maximum pageviews of these three metrics 4,741,212. This value is too large compared with others. Given by the number of cookies that view the course overview page 40000, the estimated duration to run this experiment without considering the fraction of traffic is 4741212 / 40000 $\approx$119 days, which is too long to conduct an experiment. Thus, we discard *Retention* even if it can evaluate results but hard to collect.

Hence, the total pageviews required is 685,325 based on the maximum number of samples *Net conversion*.

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.

If we divert 100% of traffic to conduct experiments:
We need 685325 / 40000 $\approx$ 18 days

However, there might exist potential risks to divert all traffic for the experiment, it is a usual case that taking a part of traffic should be securer.
I choose 80% as the fraction of traffic.

How risky do you think this experiment would be for Udacity?

Participant risk: Udacity experiment doesn't exceed the "minimal risk", which is defined as the probability and magnitude of harm that a participant would encounter in normal daily life. Udacity only concerns the expected and actual study time and

whether users make payment or not, based on cookies. These data are not identifiable and can't do harm to users' daily life.

Application risk: Udacity experiment is about an influence of a pop-up message in a webpage. The backend database and key architectures are secure.

100% traffic in 18 days and 50% traffic in 685325/(40000*50%) = 35 days.
The duration should be as less as possible under cases that the current experiment should not be influenced by other experiments, the business loss is acceptable and the press from people blogging the new changes when Udacity is not sure about the test result. 100% is not reasonable with large risk. **50% with 35 days is my choice**.

Experiment Analysis

The experiment data is given by Final project results.csv with two groups of data

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.

Sanity is check is to check whether two groups of invariant metrics has significant difference or not.

For the count type metric (number of cookies, number of clicks), we should calculate a 95% confidence interval around the fraction of events we **expect** to be assigned to the control group, and the **observed value** should be the actual fraction that was assigned to the control group. Null hypothesis: the experimental and control groups are constructed equally.

Assume a binomial distribution P(ξ=K) = C(n,k) * p^k * (1-p)^(n-k), where C(n, k) = n!/(k!(n-k)!), n =1, k=1 (the splitting experiment happens once) since pageview/click is theoretical to be evenly splitting into two groups, p_theoretical = 0.5. Then calculate the std under a binomial distribution, std = sqrt(p*(1-p)). Next, calculated 95% CI = 1.96 * SE where SE = std/sqrt(n).

z = 1.96, in two-tailed experiment with 95% CI
N_c = Control group pageviews = 345543
N_e = Experiment group pageviews = 344660
X_c = Control group clicks = 28378
X_e = Experiment group clicks = 28325

Pageviews:

SE =sqrt(p_theoretical * (1- p_theoretical) / (N_c + N_e)) = 0.006018
Margin of error = z * SE = 1.96 * 0.00602 = 0.0011796
Lower bound = p_ theoretical – margin of error = 0.5 – 0.001180 = 0.4988
Upper bound = p_ theoretical + margin of error = 0.5 + 0.001180 = 0.5012
Observed = 345543 / (345543 + 344660) = 0.5006
Observed is in the 95% CI, thus pass sanity check

Clicks:
SE = sqrt(p_theoretical * (1- p_theoretical) / (X_c + X_e) = 0.0021
Margin of error = z * SE = 1.96 * 0.0021 = 0.0041
Lower bound = p_ theoretical – margin of error = 0.5 – 0.0041 = 0.4959
Upper bound = p_ theoretical + margin of error = 0.5 + 0.0041 = 0.5041
Observed = 28378 / (28378 + 28325) = 0.5005
Observed is in the 95% CI, thus pass sanity check

Click-through-probability:
Use the difference in proportions to check CTP sanity. The difference between two groups is theoretically to be 0, thus p_theoretical_diff = 0. "Observed" value should be centered around 0, which is to be expected under the null hypothesis that the experimental and control groups are constructed equally.

p_pool of clicks = (X_c + X_e) / (N_c + N_e) = 0.08215
SE_pool = sqrt(p_pool * (1-p_pool) * (1 / N_c + 1 / N_e)) = 0.00066
Margin of error = z * SE_pool = 1.96 * 0.00066= 0.0013
Lower bound = p_theoretical_diff – margin of error = 0 – 0.0013 = -0.0013
Upper bound = p_theoretical_diff + margin of error = 0 + 0.0013 = 0.0013
Observed(difference) = CTP_e – CTP_c = 28325 / 344660 - 28378 / 345543 = 0.00006
Observed is in the 95% CI, thus pass sanity check

Another way for the sanity check of CTP is to check whether the **observed** value of the control group is in the **expected** CI calculated by the experiment group. Here, we can use the same way as handling “pageviews” and “clicks”, and don’t require the pool-related concept.

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.

Significance definitions
A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the

confidence interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)

Note: before calculating probabilities, the total sample size of “number of clicks” should be concerned. According to the given table, there are no corresponding “enrollments” and “payments” records after 2th, Nov, but still has “pageviews” and “clicks” records. To present a reliable result, we don’t know “missing values” really miss due to problems of data gathering or are all zero. *Here, let me assume that those values are missed, otherwise zero should be put into the table. So, our total sample size of “clicks” is the sum of “clicks” that exists relative “enrollments” and “payments”.*

Note that a practically a significant difference dmin does not include the sign of the effect: it could be the experimental group being significantly higher than the control, or significantly lower than the control.

Gross conversion:
N_c = Control group clicks = 17293
N_e = Experiment group clicks = 17260
X_c = Control group enrolls = 3785
X_e = Experiment group enrolls = 3423
z = 1.96, two-tailed test with 95% CI
d_min = 0.01

p_pool = (X_c+X_e) / (N_c+N_e) = 0.2086
SE_pool = sqrt(p_pool * (1 - p_pool) * (1 / N_c + 1 / N_e)) = 0.0044
d_observed = X_e / N_e - X_c / N_c = -0.0206; expect in negative direction
lower bound = d_observed – SE_pool*z = -0.0291
upper bound = d_observed + SE_pool*z = -0.0120

Since 0 is not in CI, which indicates there exists clear differences between groups, so “gross conversion” is statistically significant;
Since d_min is beyond the CI, which indicates the difference matters to the business, so “net conversion” is practically significant.

Net conversion:
N_c = Control group clicks = 17293
N_e = Experiment group clicks = 17260
X_c = Control group pays = 2033
X_e = Experiment group pays = 1945
z = 1.96, two-tailed test with 95% CI
d_min = 0.0075

p_pool = (X_c+X_e) / (N_c+N_e) = 0.1151
SE_pool = sqrt(p_pool * (1 - p_pool) * (1 / N_c + 1 / N_e)) = 0.0034

d_observed = X_e / N_e - X_c / N_c = -0.0048; expect not much changes
lower bound = d_observed – SE_pool*z = -0.0116
upper bound = d_observed + SE_pool*z = 0.0019

Since 0 is in CI, which indicates there perhaps exist a no-difference between groups, so “gross conversion” is not statistically significant;
Since d_min is in CI, which indicates the practical change may lower than the practical significance, so “net conversion” is not practically significant.

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.

Online calculator
Distribution of number of positive changes (p=0.5) is likely binomial distribution with n experiment (our case is 23 days)

Gross conversion:
Number of days with positive changes: 4
Total number of days: 23
Probability of positive changes: 0.5
Two-tailed p value: 0.0026
It is statistically significant when α = 0.05

Net conversion:
Number of days with positive changes: 10
Total number of days: 23
Probability of positive changes: 0.5
Two-tailed p value: 0.6776
It is not statistically significant when α = 0.05

Summary

If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

There is no discrepancy between the effect size hypothesis tests and sign tests. Both tests show that “gross conversion” is statistically significant while “net conversion” is not.

Recommendation

I recommend not to launch the intervention.

The “Free trial screener” intervention performs results as we expected:

1. reduce the “gross conversion” with statistical and practical significance: significantly reduce the number of frustrated users who don’t have enough time to enroll the paid course after the free trial to save coaching resources.
2. Not significantly change the “net conversion”: not significantly reducing the number of students to continue past the free trial and eventually complete the course.

However, considering the business ROI, “net conversion” doesn’t show a statistical significance nor a practical significance. “Net conversion” is the key metric that matters to the business. The statistically significant reduction of “gross conversion” is not worthwhile enough for Udacity to pay the price of importing new feature. Moreover, the lower boundary of the 95% confidence interval of observed “net conversion” is negative, which means it is possible to reduce the “net conversion”. Thus, a not profitable enough choice (even a possibility of loss) is not appropriate to make changes.

References

Udacity Google A/B Test Course: https://classroom.udacity.com/courses/ud257

Optimizely A/B Testing: https://www.optimizely.com/optimization-glossary/ab-testing/

Udacity A/B Test Forum: https://discussions.udacity.com/c/standalone-courses/ab-testing

A Summary of Udacity A/B Testing Course: https://towardsdatascience.com/a-summary-of-udacity-a-b-testing-course-9ecc32dedbb1