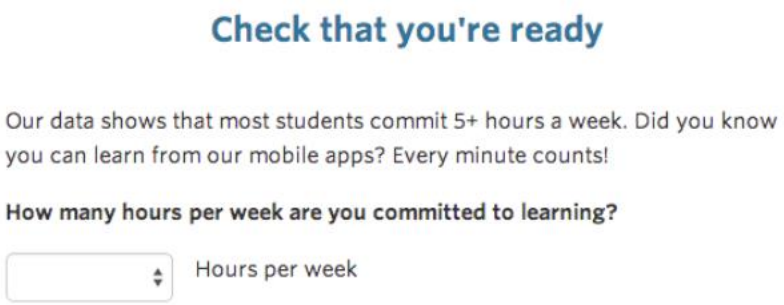
Udacity A/B Testing Final Project

Contents

1 Introduction of Experiment

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.



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.

2 Experiment Design

2.1 Metric Choice

2.1.1 Hypothesis

We should keep in mind of the hypothesis of the experiment:

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.

2.1.2 Invariant metrics

Since the change happened after clients click the 'start free trail'. Any activity before that will not be affected and can be regarded as invariant metrics.

Note:

Any place "unique cookies" are mentioned, the uniqueness is determined by day. (That is, the same cookie visiting on different days would be counted twice.) User-ids are automatically unique since the site does not allow the same user-id to enroll twice.

So, the invariant metrics are as follows:

1. **Number of Cookies** (# of unique cookies to view the course overview page)
   Number of Cookies is recorded when clients arrived, so it will not be affected by the activities later. Besides, this is the unit of diversion and even distribution amongst the control and experiment groups is expected.

2. **Number of Clicks** (# of unique cookies to click the 'start free trial' button which is before trigger the new feature.)
   Number of Clicks is recorded before the feature, so equal distribution amongst the experiment and control groups would be expected.

3. **CTP** (Click Through Probability) (# of clicks / # of Cookies)
   Before the "start free trial" button the user experience is same for all the users. Hence, we expect equal distribution in both the groups.
   Besides, it's obvious that both the numerator and the denominator will not be affected by the new feature.

Also, we should notice that we use cookie but not the user-id as the 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.

2.1.3 Evaluation Metrics

Evaluation metrics are chosen since there is a possibility of different distribution between experiment and control groups as a function of experiment. Each evaluation metric is associated with a minimum difference (d_{min}) that must be observed for consideration in the decision to launch the experiment.

Based on the hypothesis, the goal is to:

1. Reducing number of frustrated students who don't have enough time so that they leave free trail.
2. Keeping the number of students who past the free trail (become payed users) and eventually complete the course roughly the same.
3. Improving the overall student experience and improving coaches' capacity to support students who are likely to complete the course.

The third one is really a high-level goal and will happened if the first two is satisfied. So, we just need to focus on the metrics related to the first two sub-goals.

Analytically, it can be regarded as:

1. Increased retention, i.e., the ratio of users who remained enrolled past the 14-day boundary to the number of users to complete checkout should increase.

2. Decreased gross conversion and roughly the same (or increased) net conversion.

So, the evaluation metrics is as follows:

1. **Gross Conversion** (# of user_id complete check out and enroll in the free trail / # of unique Cookies to click 'start free trail')
   If the hypothesis is true, this metric should be lower after the change.
2. **Retention** (# of user_id to remain enrolled past the 14 days boundary and thus make payment / # user_id to check out and enroll in the free trail)
   If the hypothesis is true, this metric should be higher after the change.

3. **Net Conversion** (# of user_id to remain enrolled past the 14 days boundary and thus make payment / # of unique Cookies to click 'start free trail')
   If the hypothesis is true, this metric should be roughly the same after the change.

2.1.4 Unused Metrics

Number of user_id (The number of users who enroll in the free trial).

User-ids are tracked only after enrolling in the free trial and equal distribution between the control and experimental branches would not be expected. User-id count could be used to evaluate how many enrollments stayed beyond the 14-day free trial boundary. However, the number itself is meaningless, we need to look at the proportion.

2.1.5 Condition for launch the feature

Based on the hypothesis, if we want to launch the feature, there should be an increase in retention.

Meanwhile, the gross conversion will drop, so the net conversion will be affected since there are less clients enrolled in the free trail.

So, the increment of retention much cover the negative influence of gross conversion so that the net conversion will not change a lot or even increase.

2.2 Measuring Standard Deviation

For each of the metrics the standard deviation is calculated for a sample size of 5000 unique cookies visiting the course overview page. The standard deviation are calculated using 'Final Project Baseline Value' Excel

2.2.1 Result

The number of click and number of enrollments are binomial distributed. The standard deviation is sqrt(P*(1-P)/N). Note that the N should be the denominator of the metrics.

	STD
Gross Conversion	0.020230604
Retention	0.054949012
Net Conversion	0.015601545

2.2.2 Process for Calculation

	A	B	C	D	E
1					
2		given		now (pageview is given)	porpotion to given
3	Unique cookies to view course overview page per day:	40000		5000	0.125
4	Unique cookies to click "Start free trial" per day:	3200		400	
5	Enrollments per day:	660		82.5	
6	Click-through-probability on "Start free trial":	0.08		std of three metrics:	
7	Probability of enrolling, given click:	0.20625	Gross Conversion	0.020230604	SQRT((B7*(1-B7)/D4))
8	Probability of payment, given enroll:	0.53	Retention	0.054949012	SQRT((B8*(1-B8))/D5)
9	Probability of payment, given click	0.10931	Net Conversion	0.015601545	SQRT((B9*(1-B9))/D4)

2.2.3 Analytic Estimate vs. Empirical Estimate

When the unit of diversion is the same as the unit of analysis, the analytical standard deviation will match the empirical standard deviation, otherwise the variability will be different and tend to be overestimate.

The construction of these 3 metrics are as follows:

Gross Conversion (# of user_id complete check out and enroll in the free trail / # of unique Cookies to click ‘start free trail’)

Retention (# of user_id to remain enrolled past the 14 days boundary and thus make payment / # user_id to check out and enroll in the free trail)

Net Conversion (# of user_id to remain enrolled past the 14 days boundary and thus make payment / # of unique Cookies to click ‘start free trail’)

Since the unit of diversion is a cookie and the unit of analysis of retention is user_id, the two of them will be different for retention.
Besides, Since both the unit of diversion and unite of analysis of Cross/Net Conversion are cookies, the two of them will be very similar.

2.3 Sizing

2.3.1 Number of Samples vs. Power

There is no need to use the Bonferroni correction.
The multiple comparison should be used on condition that we want to do automatic detection of difference. In this kind of static analysis, we can reanalyze our data and make sure the same metric isn’t showing every time (or analysis one metric at a time) and see if the difference is repeatable.

Page views required for each evaluation metric is calculated separately using the online calculator. The alpha is 0.05 and beta is 0.2.
(Given data is from ‘Final Project Baseline Values’ Excel)

2.3.1.1 Gross Conversion

- Baseline Conversion: 20.625%
- Minimum Detectable Effect: 1%
- Alpha: 5%
- Beta: 20% -Sensitivity (1 - Beta): 80%
- Sample Size = 25,835 per variation (calculated)
- Number of groups = 2
- Total sample size = 51,670 enrollments
- CTP on ‘start free trail’ (Click/Pageview) = 0.08
- Pageviews Required = 51,670/0.08 = 645,875 enrollments (calculated)

2.3.1.2 Retention

- Baseline Conversion: 53%
- Minimum Detectable Effect: 1%
- Alpha: 5%
- Beta: 20%
- Sensitivity (1 - Beta): 80%
- Sample size = 39,155 per variation (calculated)
- Number of groups = 2
- Total sample size = 78,230 enrollments
- Enrollments/pageview: 660/40000 = 0.0165
- Pageviews = 78,230/0.0165 = 4,741,212 enrollments (calculated)

2.3.1.3 Net Conversion

- Baseline Conversion: 10.9313%
- Minimum Detectable Effect: 0.75%
- Alpha: 5%
- Beta: 20%
- Sensitivity (1 - Beta): 80%
- Sample size = 27,413 per variation (calculated)
- Number of groups = 2
- Total sample size = 54,826 enrollments
- Enrollments/pageview: 3200/40000 = 0.08 clicks/pageview
- Pageviews = 54,826/0.08 = 685,325 enrollments (calculated)

Pageviews required is maximum of pageviews required for Gross Conversion, Retention, Net Conversion. Therefore, the required pageviews is 4,741,212.

2.3.2 Duration vs. Exposure

There are 40,000 page-views per day for Udacity. Since we need 4,741,212 pageviews to test for retention, we still need 119 days even though we use 100% of our traffic, which is a crazy situation. It will let us to be in great risk and we will not be able to perform other tests since we use all the traffic. Besides, the time is too long for a traditional A/B – testing which is not related to the learning effect. So, my conclusion is that the retention is not testable. We should leave it aside and test two other metrics.

For net conversion and grass conversion, the pageview needed is 685,325. If we use all the traffic, they need 17 and 18 days. If we are aggressive and willing to take the risk, we can use all the traffic and run at the same time. In this way, we can decide whether to launch the feature in 18 days.

However, it's too risky for Udacity since we let half of our clients to see the potential harmful feature. Besides, we have no more traffic for other tests. I personally will use 60% of the traffic so that it's not that risky and we can make decisions in less than a month (29 days), which is still very efficient.

3 Experiment Analysis

3.1 Sanity Checks

3.1.1 Overview

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.** That is to check whether the difference within expectations.

Invariant metrics: number of Cookies, Number of Clicks, CTP (Click Through Probability)

The data used is 'Final Project Results' Excel, which contains the day of every single day for experiment and control group.

3.1.2 Cookies

Calculated by all the data in 'Pageviews' Column

The expected value for control is 0.5
Control: 345543, Experient: 344660
Then the observed value of control is 0.500639667

SE=sqrt((0.5*0.5)/(345543+344660))= 0.000601841
M=SE*1.96 = 0.001179608
So, the confidence interval is [0.498820392, 0.501179608]

The observed value is in the CI, so the sanity check of Cookies is pass.

3.1.2 Number of Clicks

Calculated by all the data in 'Clicks' Column

The expected value for control is 0.5

Control: 28378, Experiment: 28325
Then the observed value of control is 0.500467347

SE=sqrt((0.5*0.5)/(28378+28325))= 0.002099747
M=SE*1.96 = 0.004115504
So, the confidence interval is [0.495884496, 0.504115504]

The observed value is in the CI, so the sanity check of Number of Clicks is pass.

3.1.4 CTP (Click Through Probability)

The expected value for control is no longer 0.5. This metric should be same for both groups, so it should be the pooled CTP.

Total number of clicks: 56703; total number of cookies: 690203
The expected value for control is 0.082154091

The observed value of control is 0.082125814
The observed value of experiment is 0.082182441
The difference (D) is 5.66271E-05

SE_pool = sqrt(0.082154091*(1-0.082154091)*((1/344660)+(1/ 345543))) = 0.000661061
(The numerators are the total number of pageviews)
M = SE_pool * 1.96 = 0.001295679
CI = D+-M = [-0.001239052, 0.001352306]

Since the CI of CTP includes zero, so the sanity check of CTP is pass.

3.2 Result Analysis

3.2.1 Effect Size Tests

3.2.1.1 Overview

Based on previous analysis, there is just two evaluation metrics left；

Gross Conversion (# of user_id complete check out and enroll in the free trail / # of unique Cookies to click ‘start free trail’)
Net Conversion (# of user_id to remain enrolled past the 14 days boundary and thus make payment / # of unique Cookies to click ‘start free trail’)

Note:
Number of clicks and number of unique cookies should be calculated using the days that have enrollment and payment records.

3.2.1.2 Gross Conversion

	control	experiment	total
# enrollment	3785	3423	7208
# click	17293	17260	34553
rate	0.218874689	0.198319815	0.208607067

The pooled conversion is 0.208607067
SE = sqrt(0.208607067*(1-0.208607067)*((1/17293)+(1/17260))) = 0.004371675
M = SE*1.96 = 0.008568484
D (difference: experiment - control) = -0.020554875
CI = D +- M = [-0.029123358, -0.011986391]

Since the CI dose not includes 0, it's statistical significant.
The dmin of gross conversion is 0.01, which is out of CI, so it's also practical significance.

3.2.1.3 Net Conversion

	control	experiment	total
# payments	2033	1945	3978
# click	17293	17260	34553
rate	0.117562019	0.112688297	0.115127485

The pooled conversion is 0.115127485
SE = sqrt(0.115127485*(1-0.115127485)*((1/17293)+(1/17260))) = 0.003434134
M = SE*1.96 = 0.006730902
D (difference: experiment - control) = -0.004873723
CI = D +- M = -0.011604624, 0.001857179]

Since the CI includes 0, it's not statistical significant.
So, it's also not practical significance.

3.2.2 Sign Tests

The day-by-day conversion is calculated in the 'sign test' sheet of 'Final Project results table'

The sign tests are done using this online calculator:
https://www.graphpad.com/quickcalcs/binomial1.cfm

3.2.2.1 Gross Conversion

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.

The two tail p_value is 0.0026 which is less than 0.05, so it's statistically significant. We can reject null hypothesis and conclude that the difference is significant.

3.2.2.2 Net Conversion

Number of "successes": 10
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 0.3388
  This is the chance of observing 10 or fewer successes in 23 trials.
- The two-tail P value is 0.6776
  This is the chance of observing either 10 or fewer successes, or 13 or more successes, in 23 trials.

The two tail p_value is 0.6776 which is more than 0.05, so it's not statistically significant. We cannot reject null hypothesis and conclude that the difference is not significant.

3.2.3 Summary

The Bonferroni correction was not used. The Bonferroni correction is a method to limit / control the risk of Type 1 errors in multiple independent metrics comparisons. The multiple comparison should be used on condition that we want to do automatic detection of difference. In this analysis, I analysis one metric at a time and treat them separately. If a analysis both metrics at one time, I need to use Bonferroni correction.

The effect size test and sign test show the same results. The difference of gross conversion is statistically significant but the net conversion is not.

3.3 Recommendation

3.3.1 Overview

Keep in mind that our goal is to 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.

Besides, our evaluation metrics are gross conversion and net conversion. If the change is feature is as expected, the gross conversion should be lower and net conversion should be roughly the same or even higher.

3.3.2 Positive and negative parts

Before make recommendations, I'd like to summarize the positive and negative parts of the analysis results.

3.3.2.1 Positive

1. The results of both effective size test and sign test is as expected, the gross conversion is significantly lower and the net conversion is roughly the same. That is to say, the number of students enroll in free trail decrease and the number of students who pass the trail remain to be same.

2. The result also indicate that the proportion of free-trail students who are disappointed and thus don't pay decreased.

3.3.2.2Negative

1. For the net conversion, the difference is not significant and the point estimate is negative. So, we suffer from the risk of decreasing incomes.

2. For the goal of 'reducing the number of frustrated students who left the free trial', we don't have enough data to draw conclusions related to the retention.

3. We did not get any data about the ARPU for the experiment and control groups. I believe our ultimate goal is to make more total revenue. If so, we can launch the feature even though there is no significant change in net conversion.

3.3.3 Conclusion

I choose not to launch the experiment.

Now that it's impossible to do experiment on retention, we make naive from the change in gross conversion and net conversion that the retention is getting better. Even though the result looks good, I choose not to launch the experiment, since we are lack of critical information about the change of total income.

I am very concern about the change in revenue since there is no significant change in net conversion (the point estimate is even negative). To avoid potential risk of financial loss, we also need to run experiment on revenue.

4 Follow-Up Experiment

4.1 Analysis Result of Retention

4.1.1 Overview

Before proposing new experiment, I would like to analysis the retention based on the insufficient data. Even though the result is not solid, it can give us a intuition and can check if my previous naive conclusion is right.

The analysis result is in 'retention' sheet in 'Final Project Results' excel.

Retention (# of user_id to remain enrolled past the 14 days boundary and thus make payment / # user_id to check out and enroll in the free trail)

4.1.2 Effective size test for retention

	control	experiment	total
# payment	2033	1945	3978
# enrollment	3785	3423	7208
rate	0.53712	0.568215	0.551887

The pooled retention is 0.551887
SE = sqrt(0.551887*(1-0.551887)*((1/3785)+(1/3423))) = 0.01173
M = SE*1.96 = 0.02299
D (difference: experiment - control) = 0.031095
CI = D +- M = [0.008104, 0.0540852]

Since the confidence interval doesn't include 0, there is statically significant difference in retention. Also, the difference is also practically significant. The retention of experiment group is truly higher.

4.1.3 Sign test for retention

4.1.3.1 Overall

Number of "successes": 13
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 0.3388
  This is the chance of observing 13 or more successes in 23 trials.
- The two-tail P value is 0.6776
  This is the chance of observing either 13 or more successes, or 10 or fewer successes, in 23 trials.

The result shows that the difference is not statistically significant, which is different with parametric test. Then it was analyzed separately for weekdays and weekends.

4.1.3.2 Weekdays

Number of "successes": 7
Number of trials (or subjects) per experiment: 15
Sign test. If the probability of "success" in each trial or subject is 0.500, then:

- The one-tail P value is 0.5000
  This is the chance of observing 7 or fewer successes in 15 trials.
- The two-tail P value is 1.0000
  This is the chance of observing either 7 or fewer successes, or 8 or more successes, in 15 trials.

4.1.3.3 Weekends

Number of "successes": 6
Number of trials (or subjects) per experiment: 8
Sign test. If the probability of "success" in each trial or subject is 0.500, then:

- The one-tail P value is 0.1445
  This is the chance of observing 6 or more successes in 8 trials.
- The two-tail P value is 0.2891
  This is the chance of observing either 6 or more successes, or 2 or fewer successes, in 8 trials.

4.1.3.4 Conclusion

All the differences are not significant. However, in the weekdays, there is absolutely no difference between two groups. On the weekends, there is at least some difference, even though the difference is not statistically significant.

4.1.4 Overall Conclusion for Retention

Since the effective size test shows the result is statistically and practical significant, and it have more power than the size test (because of the hypothesis it has), I conclude that the retention seems to be better for experiment group with the insufficient data.

However, the sign test shows our clients behave differently on weekends and weekdays. It might be worth to dig deeper in the future.

4.2 Design for Follow-up Experiment

I suggest extend the duration and monitor different kind of payments for the payed students. As mentioned in the hypothesis, it will help improve coaches' capacity to support students who are likely to complete the course.

I assume the coaching is a payed service, we hope that the experiment group will spend more, at least as same the money comparing with the experiment group. (If it’s not payed, we can measure customer satisfaction or reduction in human recourses, which will not be discussed)

4.2.1 Setup

Use the same experiment previously but extend the time to make it enough longer to acquire enough data for payments after enrolled in the courses.

4.2.2 Null Hypothesis

The change of feature will not affect the payments of clients after enrolling the courses.

4.2.3 Unit of Diversion

User_id. The enrolled student already signed in and it's the unique identifier. In this experiment, we want to test the behavior of different clients which is user based but not event based.

4.2.4 Invariant Metrics

User_id. Because the experiment happened after our clients sign in.

4.2.5 Evaluation Metrics

Payment amount of clients during the experiment period.

4.2.6 Possible Outcome

The significant higher payment (higher ARPU) will lead to a higher revenue when there is no significant difference in net conversion (assume the acquired clients for website remains the same). Then, we will be very confident to launch the new feature.

However, if there is not significant difference, or even lower. Based on current situation of net conversion, we should not launch the experiment and dig some possible reasons for that. Just as a mentioned before, the difference in weekdays and weekends may be a good lead to follow.

5 Reference

[1] https://github.com/shubhamlal11/Udacity-AB-Testing-Final-Project

[2] https://olgabelitskaya.github.io/P7_Design_an_A_B_Test_Overview.html