Experiment Design

Experiment Overview: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the home 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.

Metric Choice

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

- Number of cookies: That is, number of unique cookies to view the course overview page. (dmin=3000) (Invariant metric)

This is an invariant metric as the number of unique cookies to view the course overview page would not change because students have seen the time commitment question. For same reason this metric cannot be considered as evaluation metric as it cannot show any effect on the result before experiencing the changes.

- Number of user-ids: That is, number of users who enroll in the free trial. (dmin=50) (Evaluation metric)

  Number of user-ids will be changed depends on the number of students who enroll in the free trial, so they have seen the trial screener and if they were not able to commit to the minimum number of hours they would not have enrolled. So it cannot be considered as invariant.

  Although it can be considered as an evaluation metric, I would not choose it as Evaluation metric for my investigation since it is a raw number rather than a normalized value like rate or probability. So it would not be a useful metric to measure the changes through the experiment.

- 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). (dmin=240) (Invariant metric)

  This is an invariant metric as it happens before the free trial screener is trigger. For same reason this metric cannot be considered as evaluation metric as it cannot show any effect on the result before experiencing the changes.

- 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. (dmin=0.01) (Evaluation metric)

  Click-through-probability is not an invariant because clicking the "Start free trial" button happens after seeing the trial screener which will affect the number of students that enroll.

  It cannot be an evaluation metric because it cannot measure the number of students that will stay enrolled after 14 days.

- 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. (dmin= 0.01) (Evaluation metric)

  Since Gross conversion is based on a variant, user-ids, it cannot be considered as invariant.

  Instead it can be an evaluation metric as it is an indication of effect of the trial screener on the number of enrollments, because it uses the number of user-ids in the nominator and the number of unique cookies to click the "Start free trial" button in the denominator.

  Since students became aware of required time commitment for the course before enrollment, there will be less number of students who decide to continue the course due to the ability of time allocation for the course. So we should expect less number of students enrolling for the course, hence we should see decrease on the Gross conversion metric.

  Any higher value than 0.01 for its practical significance boundary between control and experiment values, makes it a proper decision criteria in order to launch the change.

- 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. (dmin=0.01) (Evaluation metric)

  Since Retention is also based on a variant, user-ids, it cannot be considered as invariant.

  I would not use it for my examination the launch because 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 cannot be a good evaluation metric. Later on when I calculated the number of Samples given Power for this metric, it was showing over 4 million required pageviews that is too large to be used for our evaluation.

- 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. (dmin= 0.0075) (Evaluation metric)

  Since Net conversion is also based on a variant, user-ids, it cannot be considered as invariant. Instead it can be an evaluation metric as it can show how the trial screener can change the number of final enrollments. **Since screener informs student about required time commitment for the course, it will make it possible for the students to decide not to continue the course if they are not able to allocate enough time for the course. So we should expect less number of students enrolling for the course, hence we should see decrease on the Net conversion metric but hopefully not that significant that impacts the overall enrollment.**

If the Net conversion shows less than 0.0075 reducing the number of students for experiment values, then we can launch the change as it would not show significant reduction to the number of students to continue past the free trial and eventually complete the course, but we can rely on the change to plan for coach assignment with higher confidence.

Measuring Standard Deviation

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

Gross conversion: # user-ids to enroll / # unique cookies to click
Unit of diversion: cookie
Unit of analysis: cookie
Since the unit of diversion is similar to unit of analysis, then the analytic estimate would be comparable to the empirical variability and since Gross Conversion has a binomial distribution, the analytic estimates will be good.

sqrt((0.20625*(1-0.20625))/(5000*(3200/40000))) = 0.020230604

Retention: # user-ids to enroll / # user-ids to complete checkout
Unit of diversion: cookie
Unit of analysis: user-ids
Since the unit of diversion is not similar to unit of analysis, then the analytic estimate would not be comparable to the empirical variability because 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. In this case it might be worth doing an empirical estimate if there is time.

sqrt((0.53*(1-0.53))/(5000*(660/40000))) = 0.054949012

Net conversion: # user-ids to remain enrolled past the 14-day boundary / # unique cookies to click
Unit of diversion: cookie
Unit of analysis: cookie
Since the unit of diversion is similar to unit of analysis, then the analytic estimate would be comparable to the empirical variability and since Net Conversion has a binomial distribution, the analytic estimates will be good..

sqrt((0.1093125*(1-0.1093125)/(5000*(3200/40000))) = 0.015601545

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. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

Use Bonferroni? No, because as we know although the Bonferroni method is the most conservative method which is free of dependence and distributional assumptions, $\alpha_{\{\text{per comparison}\}} = \bar{\alpha}/k$, its accuracy increased by number of independent comparisons k or a lower $\bar{\alpha}$. I didn't use Bonferroni method because for the proposed hypothesis for this experiment we want to see decrease in Gross Conversation AND insignificant decrease on Net Conversion in order to be able to recommend the change. Since Bonferroni correction will recommend the change if ANY metric matches the expectations, it would not be useful for us in this evaluation.

Which metrics? Gross conversion, Retention, Net conversion

How many pageviews will you need?
I calculate the number of pageviews for each metric individually by using the R script and choose the largest value as the amount need.

```
> # sample size for Gross conversion:
> required_size(s=sqrt(0.20625*(1-0.20625)*2),d_min=0.01,
  Ns=1:200000) * (1/.08) * 2
[1] 642475

> # sample size for net Retention:
> required_size(s=sqrt(0.53*(1-0.53)*2),d_min=0.01, Ns=1:200000) * (1/0.0165) * 2
[1] 4739879 → this sample size make unacceptable duration of test, for that reason
I removed it from my evaluation metric list.

> # sample size for net conversion:
> required_size(s=sqrt(0.1093125*(1-0.1093125)*2), d_min=0.0075, Ns=1:200000) * (1/.08) * 2
[1] 679300
```

Duration vs. Exposure

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

First my risk assessment of this experiment on the following categories:

1- **Is there a chance that anyone gets hurt because of the duration of our experiment?**
   By adding a pop up window to ask student about the time that they could devote for the course, no one can get hurt because of seeing it or because of the duration of our experiment.
2- **Are we dealing with sensitive data? (Political attitudes, personal disease history, sexual preferences)**
   There will be no sensitive data involved during this experiment.

At the same time, there is always a chance that data collected over a short period of time is influenced by specific events. This is why collecting data over a longer period of time helps get a sense of the differences between week days/weekends, different weeks, or even months.

It is this combination of disadvantages of collecting 100% of data at any time, and the benefits of collecting data over a longer period of time, that makes it more beneficial to collect a smaller percentage over a longer period.

Because there is a very low to no risk for this experiment, we potentially can divert whole traffic for this experiment, but we need to consider other considerations before diverting the whole traffic.

By removing Retention from my Evaluation metrics list, I chose the sample size for "Net Conversion", 679300, as it was higher than the sample size for "Gross Conversion" and with this we can have higher confidence that our conclusion will fulfill the significance of both metrics.

Even if the sample size is much larger than the daily pageviews, 40000, I would not put the Udacity webpage by diverting the whole traffic to this experiment. Because the collected data might have been influenced by specific seasonal events or time related issues. So repeating the experiment in different time intervals for shorter number of users might be more beneficial than whole users in a short time.

At the other hand I would not take too long time for evaluating this experiment as it might enter other unknown factors and make too long and complicated to get a reliable result. In order to accurately measure the number of final enrollments with payment, I would not take the test duration anywhere less than 14 days trial time when the first payment is being made and we suppose the student has made his decision to continue the course.

By considering above mentioned notes, I would divert .5 of Udacity daily traffic to this experiment which will give me the experiment length of **34 days**, (679300 / 20000 = 34).

Experiment Analysis

Sanity Checks

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.**

Number of Cookies		SD=sqrt((.5*.5)/(tot-exper + tot-control))	m=SD*1.96	Interval= 0.5+/-m	p^=tot-control/(tot-exper + tot-control)	**Passes**
Total Control	345543	0.0006	0.0012	0.4988	0.5006	**Yes**
Total Experiment	344660			0.5012		
probability	0.5					
Number of Clicks						
Total Control	28378	0.0021	0.0041	0.4959	0.5005	**Yes**
Total Experiment	28325			0.5041		
probability	0.5					

Since both invariant metrics are within 95% confidence interval, then the sanity check for them passes.

Result Analysis

Effect Size Tests

	For Gross Conversion:	**for Net Conversion:**
Ncont (sum of all clicks in control)	17293	17293
XCont (sum of all enrollments in control)	3785	2033
NExp (sum of all clicks in experiment)	17260	17260
XExp (sum of all enrollments in experiment)	3423	1945

ppool (XCont + XExp / NCont + NExp)	0.208607067	0.115127485
SEpool (sqrt(ppol * (1-ppool) * (1/ncont + 1/nexp)))	0.004371675	0.003434134
pCont (XCont / NCont)	0.218874689	0.117562019
pExp (XExp / NExp)	0.198319815	0.112688297
dhat (pExp - pCont)	-0.020554875	-0.004873723
ME (SEpool * 1.96)	0.008568484	0.006730902
Upper (dhat + ME)	-0.011986391	0.001857179
Lower (dhat - ME)	-0.029123358	-0.011604624
dmin	0.01	0.0075
Statistical Significant	Yes	No
Practical Significant	Yes	No

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.

Sign and binomial test:

	For Gross Conversion:	**For Net Conversion:**
Number of "successes" you observed	4	13
Number of trials or experiments	23	23
Probability	.5	.5
Sign Test P-value	.0026	0.6776

For 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.

Since the P-value for Gross Conversion metric, 0.0026, is much smaller than .5 confidence interval, then we have statistical significance that the change will decrease the Gross Conversion metric.

For Net Conversion:

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.

Since the P-value for Net Conversion metric, 0.6776, is bigger than .5 confidence interval, then we don't have statistical significance that the experiment makes significant change on decreasing the Net Conversion metric.

The results of above two sign tests proves the hypothesis that this might set clearer expectations for students upfront, thus reducing the number of frustrated students (affecting Gross Conversion) 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 (Net Conversion not to be affected).

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

I didn't use Bonferroni method because for the proposed hypothesis for this experiment we want to see decrease in Gross Conversation AND insignificant decrease on Net Conversion in order to be able to recommend the change. Since Bonferroni correction will recommend the change if ANY metric matches the expectations, it would not be useful for us in this evaluation.

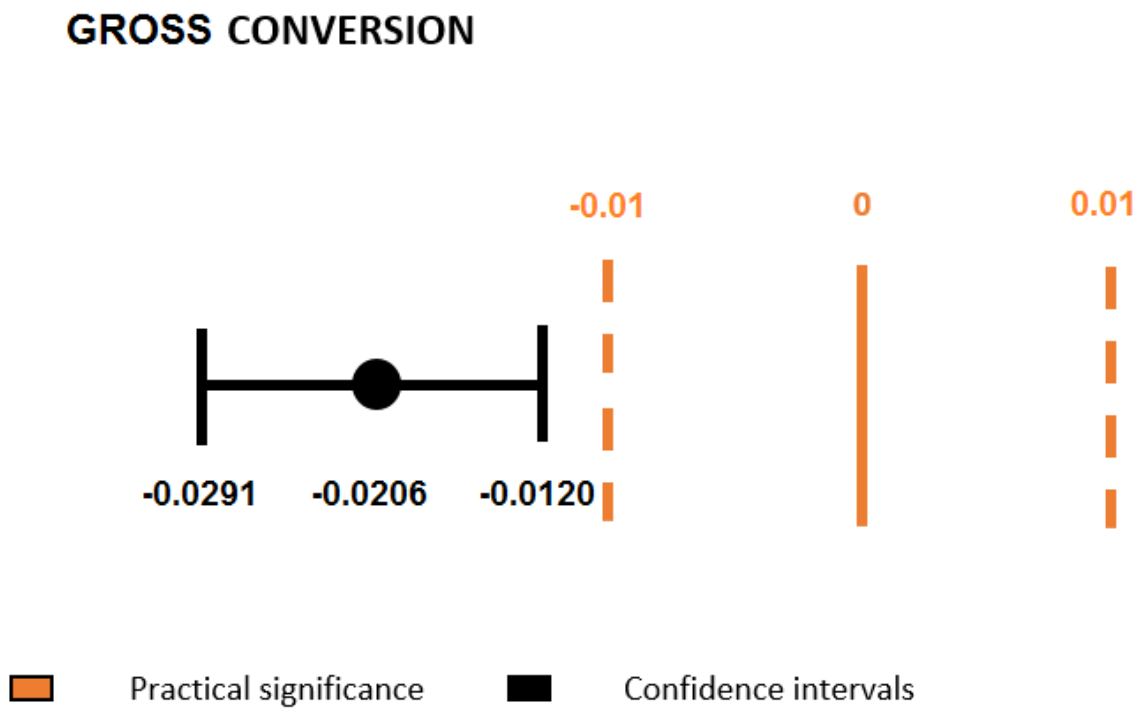
When we are considering multiple metrics at the same time, and we need all of them to inform our decision (in our case, we are looking at both gross and net conversion to decide whether to launch or not), the risk is of a type II error. That is not what the Bonferroni correction is designed for. The Bonferroni correction is designed to reduce the risk that one metric is deemed significant by mistake. If we were in the situation where we need just one metric to meet expectations in order to launch an experiment, then we would need Bonferroni. In our case we would need multiple metrics to match our expectations to launch the experiment therefore Bonferroni is neither necessary nor helpful.

I don't see any discrepancy between the effect size hypothesis tests and the sign tests for the selected evaluation metrics as follows:
For 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) we expected to see decrease on the number of enrolled users after seeing the trial screener because more students might decide to drop the enrollment if they see they cannot make time commitment. It was verified that its effect size hypothesis was Significant statistically and practically. The sign tests for this metric showed decrease on number of enrollment for most of the days and its P-value of 0.0026 was much smaller than .5 means that we have a statistical significant for which we can reject the null hypothesis for Gross Conversion in favor of having meaningful decrease in the number of students who are continuing for enrollment.

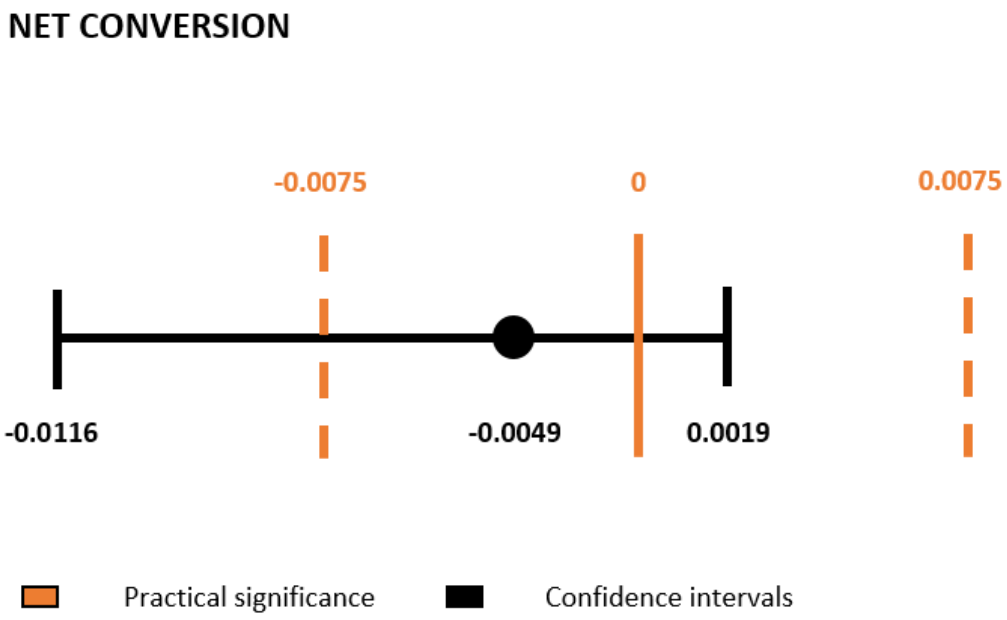
For Net conversion, (That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button.) we wanted to see if there is any significant difference (decrease or increase) on the number of enrolled users after seeing the trial screener because more serious student with clearer idea about the courses will be among enrolled users. It was verified that its effect size hypothesis was not Significant statistically and practically. The sign tests for this metric showed increase on number of enrollment for almost half of the days and its P-value of 0.6776 was close to the probability of "success" value of .05 which means the chance of change in number of paid students is almost equal on control and experiment group hence there is no significant change on the number of those students as desired for this experiment.

Recommendation

Based on the above analysis, it seems that we could see a strong statistical and practical significant difference on Gross Conversion, Not only the d-hat value(-0.0206) is below the negative value of dmin for this metric, but also its upper and lower values are outside the practical significance boundary which makes the result of the experiment for this metric statistically significant and behaves exactly as expected. The number of enrolled students will be decreased as expected and according to the hypothesis.



For the Net Conversion metric, the expected result is that we should not see significant difference between control and experiment. Although the statistical boundary contains zero and the d-hat is above negative value of dmin, the lower boundary of d-hat is smaller than dmin which indicates the risk of significant difference for this metric and the risk of loosing student enrollment.
For this reason, we cannot make same conclusion for Net Conversion as Gross conversion because it is more important for decision makers as if we should proceed with launching this change on the Udacity webpage or not.
For this reason, I recommend to run more tests with longer time period in different seasons of the year and compare the results.



Follow-Up Experiment

For the follow-up Experiment I would suggest to use user-id as diversion which is much more stable than a cookie and we can measure how long a student remains enrolled before cancellation in order to see if our follow-up experiment is working. I would study the average completion time for each course for previous students and sort the courses from easy to difficult in order to reduce the chance of cancellation for the frustrated students. For example I would definitely put A?B testing as the last course in data analysis Nanodegree program as if it was one of the early courses we would have a lot of cancellations;-)
The hypothesis would be that the sorting the courses from easy to complicated will increase the length of the student enrollments. I would consider something like Retention: That is, number of user-ids to cancel enrollment each month after paying their first payment divided by number of user-ids to make the first payment. (dmin=0.01)
Following metrics will be invariant for this follow-up:

- Number of cookies: That is, number of unique cookies to view the course overview page. (dmin=3000) (Invariant metric)

- Number of user-ids: That is, number of users who enroll in the free trial. (dmin=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). (dmin=240) (Invariant metric)

- 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. (dmin=0.01) (Invariant metric)
- 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. (dmin= 0.01) (Evaluation metric)
- 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. (dmin= 0.0075) (Evaluation metric)