DAND - A/B Testing

Author: Giacomo Sarchioni

This is my submission for the Final Project of the A/B Testing module of the Udacity Data Analyst Nanodegree. In the paragraph below I simply report an extract of the Project Instructions document provided by Udacity, in order to give some context to the experiment.

Experiment Overview

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.

Experiment Design

Metric Choice

ID	Metric	Description
1	Number of cookies	Number of unique cookies that visit the course overview page
2	Number of user IDs	Number of users who enroll in free trial
3	Number of clicks	Number of unique cookies who clicked the “Start free trial” button
4	Click-through probability	Number of unique cookies who clicked the “Start free trial” button divided by number of unique cookies to view the course overview page
5	Gross conversion	Number of user IDs to complete checkout and enroll in free-trial divided by the number of unique cookies to click the “Start free trial” button
6	Retention	Number of user IDs to remain enrolled past the 14-day boundary divided by number of user-ids to complete checkout
7	Net conversion	Number of user IDs to remain enrolled past the 14-day boundary divided by the number of unique cookies to click on the “Start free trial” button

Invariant metrics

I chose the following metrics as invariant for the experiment.

- *Number of cookies*. Visitors to the course overview page should not change as a result of this experiment. Since the “Start free trial” button only appears when somebody clicks on a course, this metric should not be affected.
- *Number of clicks*. For a very similar reason, I do not expect the number of clicks to change. The change introduced by the experiment only appears after the user has clicked on the “Start free trial” button, but it should not influence anything shown before,
- *Click-through probability*. Since this metrics is a combination of the previous two, and since both of them shall not be influenced by the experiment, I do not expect to see any change. Again, the experiment adds a change visible only after a user has clicked, so anything that happens before should not be affected.

I am not using any of these metrics as evaluation metrics because they are not measuring (at least directly) the objective described in the instructions.
In theory someone might argue that if the change introduced by the experiment is truly effective, then student experience improves. If this holds, happy students might suggest their friends/colleagues to have a look at Udacity course overview page, thus potentially affecting all

the three metrics above. This, however, is - if true - a secondary indirect effect which will be very difficult to measure with a short-term A/B test. It is much more plausible to assume that these metrics won't change throughout the duration of the experiment, thus making them invariant metrics to test for.

Evaluation metrics

I chose the following metrics as evaluation metrics for the experiment here considered.

- *Gross conversion*. The change introduced in the experiment should refrain visitors who are not ready to commit a certain amount of hours per week to the selected course from enrolling, thus affecting the proportion of enrolled users over the total number of visitors who clicked (i.e. *gross conversion*). If the change added through experiment works, I should see a reduced value for *gross conversion* in the experiment group.
- *Retention*. This is a very common sense metric to test for. If the change added through the experiment works, the proportion of students who remain enrolled in the course after the 14-day boundary should increase, exactly because these students should have been more willing to dedicate time to the course, thus making them more unlikely to withdraw.
- *Net conversion*. If the experiment works (i.e. better user experience for the students, after having filtered them), the number of students in the experiment group who keep on being enrolled after the 14-day boundary, as a proportion of the unique number of visitors, should increase **or, at least, not change**. This is the tricky part. Because I expect the gross conversion to go down, I expect that fewer users will enroll. Since I expect the number of cookies not to change (invariant metric), if the *net retention* for the experiment group does not change, given a smaller number of enrolled students in the experiment group, it means that a higher number of students remain enrolled and that the change was actually successful. Bear in mind that I will need to take into account the practical significance level of 0.75% for this.

Please notice that, in the end, I dropped the *retention* metric as an evaluation metric because it would have required an enormous sample size as well as a very long duration. More details are available in the *Sizing* chapter.

Other metrics

Initially, I considered the metric *Number of user IDs* to be an evaluation metric. Definitely, I wouldn't have considered it to be an invariant metric since the experiment might have an impact on the enrollment rate and, consequently, on the number of user IDs.
Having *Number of user IDs* as an evaluation metric would imply that Udacity is interested in seeing a different (increased or decreased) number of visitors who enroll in the course. However, this is not the purpose of the test, since Udacity is not necessarily interested in increasing/decreasing the number of enrolled students but rather on filtering off students who might not dedicate enough time to the course, without significantly reducing the number of students to continue past the free trial and eventually complete the course

Summary

In short, I would recommend to implement the change added through the experiment if:

1. Any change in the invariant metrics is not significantly different from zero;
2. The *gross conversion* for the experiment group is significantly lower than 1%;
3. The *net conversion* for the experiment group is not significantly different from zero, and the CI does not include the minimum detectable change.

Measuring Standard Deviation

In the table below I report the standard deviations for the two *conversion* metrics. For completeness, I also add the standard deviation for *retention*, although I have not used it in the end.

Metric	Unit of analysis	Baseline p	SD
Gross conversion	400 unique cookies who clicked	0.20625	0.0202
Net conversion	400 unique cookies who clicked	0.1093125	0.0156
Retention	82.5 enrolled students	0.53	0.0549

These calculations assume a sample size of 5,000 unique cookies and baseline values as per this file.

For the *conversion* metrics the unity of analysis (cookies) is the same as the unit of diversion (cookies) used in the experiment. Therefore, it is very likely that the analytical variabilities here calculated are going to be very similar to their corresponding empirical estimates.
In the case of *retention*, however, the unit of analysis (user IDs) is different from the unit of diversion. As a consequence the analytical variability could be much higher than reality, so, in this case, it would be better to also check the empirical estimate.

Sizing

Number of Samples vs. Power

I am not using the Bonferroni correction, so I keep the alpha level for all tests at 5%.
The numbers of pageviews required for each metric are reported below. The overall number of pageviews is the maximum number of pageviews among the required sizes for all the metrics.

Size is calculated assuming a type-II error probability of 20%, i.e. Beta for the test is 20%.
Please notice that the final number of pageviews is twice as much as the number reported by the online sample size calculator. This is due to the fact that the minimum sample size required apply to both the experiment and the control groups.

Metric	Baseline	Minimum detectable effect	Size	Baseline proportion (% pageviews)	Pageviews
Gross conversion	20.625%	1%	25,835 Users who clicked and enrolled (https://goo.gl/LIhnDd)	0.08 This is the baseline click-through probability on "Start free trial", i.e. the number of unique cookies who clicked the "Start free trial" button over the total number of unique cookies who visited the course overview page.	645,875
Net conversion	10.93125%	0.75%	27,413 Users who clicked and enrolled (https://goo.gl/FtPpLX)	0.08	685,325
Retention	53%	1%	39,087 Enrolled users who remain enrolled after the 14-day boundary (https://goo.gl/LWs5zb)	0.0165	4,737,819

As shown in the table above, in order to perform a test that meets the required levels of alpha and beta, Udacity will need to run the test on more than 4.7 million unique cookies (i.e. unique visitors) to the course overview page. This extremely high number is due to the fact that only a few number of visitors do actually enroll and pass the 14-day boundary window (the baseline value is 1.65%).

At the end of this paragraph I will provide a summary on which evaluation metrics I decided to choose, but for the time being I will consider this two possible scenarios.

Scenario	Evaluation metrics	Size (number of unique cookies)
1	• Gross conversion • Net conversion • Retention	4,737,819
2	• Gross conversion	685,325

	• Net retention	

Duration vs. Exposure

Starting from the baseline value of 40,000 unique cookies visiting the course overview page everyday, the duration of the experiment (by different diversion fractions) is reported in the table below.

Duration (in days)

	Fraction			
Scenario	*0.25*	*0.5*	*0.75*	*1*
1	474	237	158	119
2	69	35	23	18

If Udacity includes *retention* as one of the evaluation metrics (i.e. scenario 1), even if all traffic to the course overview page is diverted to the experiment (i.e. fraction = 1), the duration of the experiment would be ~ 119 days. This duration is way too long to test the effectiveness of the change introduced since:

- Users assigned to the control group (i.e. those who are not shown the prompt on weekly hours willing to commit to the course) might experience UX frustration;
- Udacity coaches will need to support non-filtered students (i.e. those who are not committing to the course for the minimum recommended amount of weekly hours) as before, thus potentially offering a lower coaching service that might impact students in the experiment group as well;
- For more than 5 months, all the traffic to the course overview page will be diverted to this experiment, thus having an impact on other tests that might be run in parallel. Reducing the fraction of traffic diverted to the test seems unreasonable since duration will be even higher.

As a consequence, I recommend dropping *retention* as an evaluation metric and only include *gross conversion* and *net conversion*. Since the change added through the experiment has a very minimal impact on the UX of visitors to the course overview page, I would be comfortable in choosing a diversion fraction equal to 1; this means that for 18 days all the traffic to the course overview page will be diverted to the experiment. If Udacity needs to run other tests on the same traffic, a 0.75 or a 0.5 diversion fraction would also be acceptable, since the experiment will not run for more than a month (approximately).

Experiment Analysis

Sanity Checks

Statistics for the invariance metrics are reported in the table below. All calculations are available here.

Metric	Expected value	Observed value	Lower CI	Upper CI	Invariant?
Number of cookies (proportion assigned to control group)	0.5	0.5006	0.4988	0.5012	Yes
Number of clicks (proportion assigned to control group)	0.5	0.5005	0.4959	0.5041	Yes
Click-through probability	0	-0.0001	-0.0013	0.0013	Yes

All sanity checks are ok, i.e. I can be confident at 95% significance level that the metrics I chose as invariant are actually the same in the control and the experiment group.

Result Analysis

Effect Size Tests

The table below shows relevant statistics for the two evaluation metrics I chose, i.e. *gross conversion* and *net conversion*.

Metric	d min	d observed	Lower CI	Upper CI	Statist. significant	Practic. significant
Gross conversion	0.01	-0.0206	-0.0291	-0.120	Yes	Yes
Net conversion	0.0075	-0.0049	-0.0116	0.0019	No	No

The *gross conversion* in the experiment group is significantly different than that in the control group. In particular, the *gross conversion* of the experiment is lower than that of the control group at the practical level (i.e. a change of at least 1%). In other words, Udacity should expect to see the *gross conversion* (i.e. the numbers who unique visitors who clicked) to go down by at least 1% in case it decides to implement the change tested in the experiment.

On the other side, there seems to be **no statistically significant difference** between the control and experiment group when it comes to *net conversion*. In theory this would be a good signal, since, given the lower number of enrolled students, if the proportion of students who remain enrolled after the 14-day boundary does not change, then it means that the experiment was successful. **However**, the lower bound of the CI is smaller than -0.75% (being -1.16%).

This means that the decrease in *net conversion* may be higher than the minimum detectable change accepted.

Sign Tests

The table below reports the number of successes for each evaluation metric and the p-value for the sign test.

Metric	Successes	Cases	P-value (two-tailed)	Significant
Gross conversion	19	23	0.0026	Yes
Net conversion	10	23	0.6776	No

The sign test confirms the finding above. The sign difference in *gross conversion* is statistically significant (two-tailed, alpha=5%), while that for *net conversion* is not.

Summary

In my analysis, I have **not used** the Bonferroni correction. Since I expect the two *conversion* metrics to be correlated, using the Bonferroni correction, i.e. reducing the alpha level for each test, would make me too much conservative when it comes to rejecting the null hypothesis. This may actually result in an increased probability of type-II error, i.e. failing to reject the null hypothesis.

Recommendation

Change in *gross conversion* is statistically and practically significant. This is a good sign, since I expected the change introduced through the experiment to filter the number of students enrolled in the course.

On the other hand, however, while the CI for the *net conversion* includes zero (i.e. I can assume that there is no statistically significant difference between the control and the experiment group), the same CI also includes values which are greater than the minimum detectable effect, thus making possible for the *net conversion* rate to be smaller than the minimum accepted level. In other words, for a lower *gross conversion* rate, Udacity may also see a comparably lower *net conversion* rate.

Through this test, Udacity wanted to reduce “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”.
It is true that the change added through the experiment helped to filter off some users (thus reducing the *gross conversion*), but it is also true that **such a change might actually reduce** the “number of students to continue past the free trial and eventually complete the course” - and

this is against the objective previously stated. Therefore, I **do not recommend** to implement the change.

Follow-Up Experiment

In the follow-up experiment, I am asked to identify and test a change that would help reduce the number of frustrated students who **cancel early** in the course.

The previous experiment used the "week-per-hours" dedicated to the course as a proxy of the likelihood of early cancellation, i.e. if a student does not dedicate enough time to the course, he is more likely to drop off.
However, students may cancel within the 2-week period for other reasons, e.g. the course did not meet their expectations, the course was too difficult/too easy, et cetera.
A possible experiment, then, would be that of adding additional information to the form that appears before the visitor clicks on the "Start free trial" and enroll in the course. Such information could include questions related to:

- Expected skills and capabilities to be acquired during the course, maybe in the form of a categorical response such as NA, Low, Medium, High;
- Minimum requirements, maybe in the form of a checklist.

The advantage of this experiment is that of maintaining the same setup of the previous one. Invariant and evaluation metrics would be exactly the same.
The major disadvantage, however, would be a longer and potentially more frustrating enrollment process, due to the longer form. This may discourage valid students from enrolling thus making Udacity lose numerous opportunities.

Another approach would be that of adding a scoring mechanism (similar to the badge systems in Khan Academy) where students who are consistent and dedicate enough hours to the course receive a badge/notification in their Udacity profile. In practice, every time a student completes a module or watches a lessons for a certain amount of time, he is awarded a progress badge. This should keep the student motivated and refrain him from losing momentum and commitment. Details for this experiment are reported below.

Unit of diversion	Since the experiment is performed exclusively on enrolled students, the unit of diversion will be user ID.
Invariant metrics	I expect students to be assigned to the control or the experiment group with equal probabilities. Therefore, the number of user IDs should be an invariant metrics.
Evaluation metric	In this case, *retention* would be the ideal candidate for evaluating any difference between the control and the experiment group.
Recommendation	If the *retention* rate is significantly higher than a specified practical significance level, I would recommend Udacity to implement the change.