

# Final Project - A/B Testing

Ignacio Ferreras Astorqui

May 2017

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Experiment Design</b>	<b>2</b>
2.1	Metric Choice . . . . .	2
2.2	Measuring Standard Deviation . . . . .	4
2.3	Sizing . . . . .	4
2.3.1	Number of Samples vs Power . . . . .	4
2.3.2	Duration vs Exposure . . . . .	4
<b>3</b>	<b>Experiment Analysis</b>	<b>5</b>
3.1	Sanity Checks . . . . .	5
3.2	Result Analysis . . . . .	5
3.2.1	Effect Size Tests . . . . .	5
3.2.2	Sign Tests . . . . .	6
3.2.3	Summary . . . . .	6
3.3	Recommendation . . . . .	6
<b>4</b>	<b>Follow-Up Experiment</b>	<b>6</b>

# 1 Introduction

This project consist on the testing of a possible change to the way users enroll on the 15 day trial to any course in the Udacity website. The data I am going to use it is not the real one, but an adaptation with the same results. The technique I am going to use to demonstrate if this change really helps or not is A/B testing. For that we are going to have data from multiple users, differentiated into two groups. The control group is going to be people with the old interface, and the experiment group with the new changes applied. This way we are going to compare the results gathered and make a decision based on if these new changes actually help the Udacity users. The data is composed of the page views received, the number of click in the “Start Free Trial” button, the number of people to finalize the enrollment and finally the number of users to start paying for the course (first month at least). We only used the users who paid at least the first month because A/B test are generally fast, and finalizing a course is longer than the calculated time we will see the Duration vs Exposure section.

## 2 Experiment Design

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.

### 2.1 Metric Choice

For this project had these features available to choose to be either our invariants or evaluation metrics:

- Number of cookies
- Number of user-ids
- Number of clicks
- Click-through-probability: Number of user-ids to enroll divided by the number of unique cookies.
- Gross conversion: Number of user-ids to enroll divided by the number of unique cookies to click in the “Start Free Trial” button.

- Retention: Number of user-ids to make at least one payment divided by the number of users to enroll.
- Net conversion: Number of user-ids to make at least one payment divided by the number of unique cookies to click in the “Start Free Trial” button.

The Number of Cookies wasn’t chosen to be an evaluation metric given that it provides little information just by itself to be able to see if the experiment is taking any effect. However as an invariant it provides the possibility to make sanity check because the evaluation metrics rely on this metric and it is a population sizing metric. In order to launch the experiment I am going to be looking for

The Number of User-ids doesn’t provide any information alone and also we already have a population sizing metric that measures better the number of people to interact with our test. Also it isn’t a good evaluation metric because it doesn’t give us any information about the possible changes in the number of enrollments.

The Number of Clicks has been chosen to be an invariant metric because it directly affects the evaluation metrics so this provides a good way to measure the population sizing. I haven’t considered it to be a possible evaluation metric given that it doesn’t provide enough information by itself to decide if we have achieved our objective.

The Click-through Probability hasn’t been chosen for any of the to metrics for the following reasons. First, it wasn’t considered to be an evaluation metric given that it doesn’t provide any information about the number of people that enrolls past the 14 day trial. Second, it was chosen as a invariant given that it didn’t give any new information or possible control due to the fact that it is just a relationship between the already chosen invariants.

The Gross Conversion has been chosen as an evaluation metric because it gives you a good proportion of users that enroll given the number of clicks. It is a good way to measure “how many users go down the funnel”. It wasn’t chose to be an invariant given that it was a really good evaluation metric.

The Retention seemed to be a great evaluation metric. However, after some calculations I found out that in order to use this metric we would have needed more than 3 months of study, which for an A/B test is too much time. It wasn’t chosen to be a invariant metric given that it isn’t a population sizing metric and it is a metric that we expect to change.

Finally the Net Conversion was chosen to be an evaluation metric because it gives great insight about the engagement of the users to the course by measuring the proportion of users to keep working on the course past 14 days by the number of clicks. Also I didn’t considered to be a invariant metric given that we expect it to change and it isn’t a population sizing metric.

## 2.2 Measuring Standard Deviation

The standard deviations for the evaluation metrics are the following:

- Gross Conversion: 0.0202
- Net Conversion: 0.0156

The formula to calculate these results was the following:

$$\sqrt{\frac{P_1 * (1 - P_1)}{N}}$$

Given that the Gross conversion is a relative simple metric it is safe to assume that the distribution is more stable thus the estimate we have calculated should be comparable to the empirical variability. Same case for the Net Conversion given that these two metrics are closely related.

## 2.3 Sizing

### 2.3.1 Number of Samples vs Power

For this project I won't be using the Bonferroni correction. The reasons can be seen in the Summary section. Now we have to calculate the number of page views needed for the project to develop. For that we are going to use this on-line calculator. We have to provide the baseline conversion rate, in our case the Probability of enrolling given click and the Probability of payment given click, separately. The minimum detectable effect, which was provided to us. The  $\beta$  and the  $\alpha$  which were provided too. Due to the fact that we are not going to use Bonferroni we don't need to calculate individual alphas for each evaluation.

Once we got the sample sizes from the calculator we need to extrapolate these results because the inputted data was for 40000 cookies. Once I did that we found that the necessary sample size was: 685275 page views.

### 2.3.2 Duration vs Exposure

Now that we know the number of page views needed we can decide how many days we are going to invest in gathering this data. One thing I have to decide is how much of Udacity's traffic we will be using. For this case and given that this test doesn't pose any risk I decided to use 100% of Udacity's traffic. There is no risk for Udacity given that the test is not altering any key component of the website

So known the percentage of traffic we are going to be using and the number of page views needed we know that to gather the necessary data we are going to need 18 days.

## 3 Experiment Analysis

### 3.1 Sanity Checks

In order to see if the work we have been doing is correct we are going to exercise a sanity check in the invariant metrics. For that we are going to compute a 95% confidence interval for the value we expect to observe. So the sanity check will pass if the observed data falls into the confidence interval.

The results calculated are the following:

	Lower bound	Upper bound	Observed
Number of cookies	0.4988	0.5012	0.5006
Number of clicks	0.4959	0.5041	0.5005

As we can see both observed values fall inside the interval. This means that the sanity checks for both invariant metrics pass.

### 3.2 Result Analysis

#### 3.2.1 Effect Size Tests

For this part we are going to check if our evaluation metrics are statistically and/or practically significant. In order to do that I computed a 95% confidence interval around the difference between the experiment and the control groups.

The results obtained where the following:

	Lower bound	Upper bound
Gross Conversion	-0.0291	-0.012
Net Conversion	-0.0116	0.0019

In order to see if each metric is statistically and/or practically significant we need to follow these rules:

- A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change)
- A metric is practically significant if the confidence interval does not include the practical significance boundary

Knowing these rules I established that the Gross Conversion is both statistically and practically significant. The Net Conversion was found to be neither statistically nor practically significant.

### 3.2.2 Sign Tests

In order to check if the previous results are accurate we are going to work with the day-by-day data, by calculating the p-value. The results should show the evaluation metrics that are statistically significant, and the should coincide with the results from the Effect Size Tests.

	p-value
Gross Conversion	0.0026
Net Conversion	0.6776

The rule in this case to know if it is statistically significant is to check if it is lower than the probability. Which in this case is 0.5. Knowing these we can establish that the only statistically significant evaluation metric is the Gross Conversion.

### 3.2.3 Summary

For this project I didn't use the Bonferroni correction given that the two evaluation metrics are likely covariant. The Bonferroni correction works really well for multiple invariants that are not related given that in that case it tends to be too "conservative" and could have lead to future problems in the project.

## 3.3 Recommendation

Given the results obtained I would recommend launching the experiment. We had two different evaluation metrics and only one of them showed a statistically and practically significance. This means that it is likely to have a positive impact in the user experience given that the change we are examining is relatively small.

## 4 Follow-Up Experiment

The objective of the experiment is to reduce the number of early dropouts from desperate users. The experiment I would conduct is an alteration to the quiz solver, because it is the place where the user might get desperate. This alteration would be the appearance of a message once the user has tried to solve the quiz more than 5 times for example. The message would contain a text such as this:

"Don't desperate you are not alone! Go to the forum to look for answers!"

The objective with this message is to remind the users to look in the forum for answers to their questions.

The hypothesis for the experiment is to reduce the number of dropouts by redirecting the desperate users to the forum. This is a close adaptation of the previously stated objective that not only should reduce the number of dropouts but increase the participating members in the forum.

The metrics that I would use for the realization of the project are:

- Number of lessons passed: A count of all the lessons that have been passed
- Number of participation in the forums:
- User-ids: in order to keep track of all the users active
- Number of clicks in the forums link: in order to measure the number of people getting in to get help

Measuring the dropout would be a great way of measure but for a A/B test measure the dropout would require a much larger range of time.

Finally the unit of diversion should be the user-ids given that it is the unique key in the experiment. In addition it is a great way to track the users better than the cookie or something else because in order to access the course content one must be logged in.