# Udacity's data analysis Nanodegree program
# Project 7 - A/B test experiment
# By: khaled salah
# March, 2017

## Table of contents

## 1. Abstract

This experiment is the final project for the A/B testing course offered by Udacity for students enrolled in data analysis Nanodegree program.

All statistical calculation shown here are done using google sheets.

I'll be analyzing the data from an actual experiment made by udacity (will be explained below), the specific numbers have been changed, but the patterns have not.

## 2. 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.

# 3. Experiment Design

## 3.1. Metric choice

### Invariant metrics :

Invariant metrics are the metrics which aren't expected to change during the experiment, and they are useful in two ways, first they will help us to choose the sample size needed for the experiment by distributing them between the two control and experiment groups. Second by measuring them for the two groups, if they are equal ( not exact equal but fairly equal, more details in sanity check part) then we can trust our experiment, otherwise we would know that something went wrong in our experiment.
Invariant metrics here are divided between number of clicks on the "start free trial" button and the number of views for the course overview page, which both happen before the message is showed to the student, and both aren't expected to change.

- Number of clicks ($d_{min}$=240)*: That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger).
- Number of cookies ($d_{min}$=3000): That is, number of unique cookies to view the course overview page.
- Click-through-probability ($d_{min}$=0.01): 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.

### Evaluation metrics :

On the other hand Evaluation metrics are expected to change and we will use them to measure the result of our experiment, each of them is explained below :

- Number of user-ids ($d_{min}$=50): That is, number of users who enroll in the free trial. I'm expecting this count to decrease during the experiment, but it would be a better idea for using probability instead of count. So I'm not going to use this metric and instead I'll use Gross conversion.
- Gross conversion ($d_{min}$= 0.01): 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. Which is ( #enrollments / #clicks ) and we want this metric to decrease, as this means that some of the students who saw the message, didn't enrolled in the free trial, thus reduction the number of  frustrate students.
- Retention ($d_{min}$=0.01): 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. Which is ( #payments / #enrollments ), and we want this metric to increase as the percentage of students who exit the course after enrollment should be decreased.
- Net conversion ($d_{min}$= 0.0075): 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. Which is ( #payments / #clicks ),

As the experiment only decrease the number of frustrated students who cancel early in the course, which will not affect the number of students who start paying, so I'm not expecting this metric to decrease as if it decreased thus it won't be a good idea to launch the experiment, so it's very important to track this metric.

* $d_{min}$ = The practical significance boundary for each metric, that is, the difference that would have to be observed before that was a meaningful change for the business. All practical significance boundaries are given as absolute changes.

## 3.2. Measuring Variability

This spreadsheet contains rough estimates of the baseline values for the metrics per day, and for each of them evaluation metrics I choose, I'm going to calculate its analytical estimate of standard deviation given a sample size of 5,000 cookies visiting the course overview page.

Assuming the binomial distribution for all of them, which gives the SD by the equation (P(1-P)) / N, where P is the probability given in the spreadsheet and N is the number of the unit of analysis ( the denominator of the metric ) equivalent to sample size of 5000 cookies.

- Gross conversion : ( #enrollments / #clicks ) the denominator here is number of clicks, and as given the number of clicks per day is 3,200 clicks, views per day is 40,000 and we are calculation for 5,000 views, so N will be 400 clicks. P = Probability of enrolling given click = 0.20625, and then the standard deviation will equal 0.0202 .
- Retention : ( #payments / #enrollments ) the denominator here is the number of enrollments, and as given the number of enrollments per day is 660, views per day is 40,000 and we are calculation for 5,000 views, so N will be 82.5 enrollments. P = Probability of payment given enroll = 0.53, and then the standard deviation will equal 0.0549 .
- Net conversion : ( #payments / #clicks ) the denominator here is number of clicks, and as calculated above N will be 400 clicks. P = Probability of payment given click = 0.1093125, and then the standard deviation will equal 0.0156 .

In case of the unit of analysis ( the denominator ) is the same as the unit of diversion ( number of clicks ), the empirical and analytical standard deviation tends to be the same, otherwise it's different. So for retention we may want to compute its empirical standard deviation later.

## 3.3. Sizing

### Number of Samples :

Using the analytical estimate of standard deviation we just calculated to compute the total page views ( for the two groups control and experiment ) needed for the test.

Considering an alpha of 0.05 and beta of 0.2, I will calculate the sample size per group for each of the metrics using this online sample size calculator, compute the needed page views, and then multiply each of them by two to get the total sample size needed for each metric, and we will use the highest value for our test.

| metric | Conversion rate | $d_{min}$ | Sample size Per group | Page views per group | Total page views |
|---|---|---|---|---|---|
| gross conversion | 20.625% | 1% | 25,835 clicks | 322,937.5 | 645,875 |
| retention | 53% | 1% | 39,087 enrolls | 2,370,606.5 | 4,741,213 |
| Net conversion | 10.93125% | 0.57% | 27,413 clicks | 342,662.5 | 685,325 |

The highest page views size is for retention, and actually +4 millions is a huge sample size we can't give up on it. So for that reason we will remove retention from our evaluation metrics, and use gross conversion and net conversion, which yields to sample size of 685,325 page views.

## Duration vs exposure :

Applying the experiment to 100% of the traffic seems to be ok, as there is no chance that anyone got hurt by the screener and also there is no sensitive data to worry about.
With daily 40,000 page views, and our experiment needs 685,325 page views to complete, then it will take 18 days to finish.

# 4. Experiment analysis

## 4.1. Sanity check

The experiment data can be found on this spreadsheet.
As I mentioned before, the invariant metrics will be useful to indicate whether the experiment includes an error or not, by measuring the invariant metrics for each of the control and experiment groups. If they are equal then we are safe, the problem is that they aren't exactly equal but are close to each other. To check if that close is ok or not, we are going to compute 95% confident interval around each metric. Then compute the ratio for each of the invariant metric per control group ( observed )
( example: #cookies-ratio-for-control-group = #cookies-for-control-group / total-#cookies ),
And if that ratio falls between the confident interval then we are good.
The following table shows the values for each invariant metric.

| metric | expected* | observed | CI lower boundary | CI upper boundary | result |
|---|---|---|---|---|---|
| Number of clicks | 0.5 | 0.5005 | 0.4959 | 0.5041 | pass |
| Number of cookies | 0.5 | 0.5006 | 0.4988 | 0.5012 | pass |
| Click through probability | 0.0 | 0.0001 | -0.0013 | 0.0013 | pass |

- For the click through probability the CI is built around 0, and by calculating pooled standard error for a binomial distribution.
- For clicks and cookies the value is .5 as they are a count and the probability of distributing the values between the two groups is .5.

## 4.2. Result analysis

### Effect size test :

Now after we passed the sanity check, it's time to check the evaluation metrics to see the outcome of the experiment. For each of the two evaluation metrics we choose, we are going to calculate a 95% confident interval and then check whether each metric is statistically and/or practically significance. 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).
Also I’m not using the Bonferroni correction ( the reason can be found in the summary section). Our alpha will equal 0.05.
As I’m assuming a binomial distribution, I’m going to use the pooled probability/SE for the calculations. The below table shows results in steps.

| | Gross conversion | Net conversion |
|---|---|---|
| Xcontrol | 3785 | 2033 |
| Xexperiment | 3423 | 1945 |
| Ncontrol | 17293 | 17293 |
| Nexperiment | 17260 | 17260 |
| Pcontrol | 0.2189 | 0.1176 |
| Pexperiment | 0.1983 | 0.1127 |
| Ppooled | 0.2086 | 0.1151 |
| SEpooled | 0.0044 | 0.0034 |
| Margin of error | 0.0086 | 0.0067 |
| CI lower bound | -0.0291 | -0.0116 |
| CI upper bound | -0.012 | 0.0019 |
| $d_{min}$ | 0.01 | 0.0075 |
| Statistically significant | Yes (CI doesn’t include 0) | No (CI does include 0) |
| Practically significant | Yes (CI doesn’t include $d_{min}$) | No ( CI does include $d_{min}$) |

## Sign test :

For further test we do the sign test using the day-by-day breakdown. For each metric I’ll compare the control and experiment values, and count the number of values where control is bigger than experiment ( or vice versa, it would give us the same result) and call it successes, then using this online sign test calculator compute the two-tailed p-value.

| metric | successes | Trails (total) | p-value | significant? |
|---|---|---|---|---|
| Gross conversion | 19 | 23 | 0.0026 | Yes |
| Net conversion | 10 | 23 | 0.6776 | No |

Summary :

- Regarding Gross conversion, it’s statistically significant which means that there is a change in the number of enrollments given clicks and as the confident interval is negative so it dropped down - as needed. Also it’s practically significant which means that the change is a big change which matters to us, and supports our hypothesis that the number of students who enrolled in the free trial will decrease.
- Regarding Net conversion, it’s not statistically significant nor practically significant. Actually the Net conversion ( aka number of students who started paying for courses ) is alittle bit decreased, which is not a good thing.
- The sign test supports the other tests.
- Launching the experiment is depending on the significance of both metrics, so I’m not going to use the Bonferonni correction.

## 4.3. Recommendation

The main goal of our experiment is to decrease the number of students who enrolls in the free trial and then exit before completing the 14 days. Which will decrease the number of enrollments given clicks which is approved by the experiment, but also will increase the number of students who started paying given the number of enrollments, which we couldn’t compute because it needed a very large sample size (Retention).
Regarding the net conversion, I wasn’t expecting it to decrease. More than the fact it decreased, it also included the negative of the practical significance boundary which indicate that the number went down by an amount which would matter to the business.
So to recall the null hypothesis for gross conversion is that there is no change, and for net conversion that it will decrease, and to launch the experiment we have to reject both null hypothesis. For gross conversion we rejected the null, but for net conversion we didn’t and worse it decreased because of the experiment, so my recommendation is not to launch the experiment.

## 5. Follow-up experiments

As our goal is to reduce the number of frustrated students who cancel early in the course, lets search for a way that motivate them to complete the course.

Students would feel motivated if they knew that they are not alone, and if they knew how much that course is valuable for their career.

So my experiment is to add students of some course to a group ( Google+ groups for example) which would serve the two points above, first they will motivate each other and meet other students who shares the same course, second mentors could remind them for while and awhile how important that course for their career and who they will benefit from it.

The unit of diversion will be the number of user-ids who enrolled in some course.

The invariant metric will be  the number of enrollments.

The evaluation metric will be “Retention” the number of users started payments given enrollments, our goal is to increase this metric.

The null hypothesis is that Retention won’t change.

The alternative hypothesis is that Retention will increase.

As the unit of diversion is the same of unit of analysis, so it will require less number of sample size, unlike the previous experiment.