- Data Analyst Nanodegree
- Project 7 - Design an A/B Test
- Ricardo Yoshitomi

A/B Test: Free Trial Screener

Objective

The objective of this project is to perform an A/B test on a change in the Udacity website. The Udacity courses currently have two options on the homepage. The students can choose for **“start free trial”** where they will be enrolled in the paid version of the course. Choosing this option the students will have access to the videos, take the quizzes, receive coaching support, submit their final project for feedback and receive a verified certificate. After 14 days, they will be automatically charged unless they cancel first. In the other option **“access course material”**, the students will only have access to the videos and take the quizzes for free without any other support.
The experiment consist in a change in the “start free trial” option. When the students choose this option, a message will appear asking how much time they had available to devote to the course. It is expected to the students have at least 5 hours per week of commitment. If they indicated fewer than 5 hours per week, a message would appear indicating that the course require a greater time and suggesting that the student might like to access the course materials for free.

Hypothesis Test

The hypothesis was that the change would reduce the number of frustrated students who left the free trial because they didn’t have enough time. This change would maintain the number of students willing to complete the course. If the hypothesis was true, Udacity could improve the services offered to the students (ex: coaching support) until they finished the course.

Experiment Design

Metric Choice

Before a student clicks on the “start free trial” button, we only have the cookie, C. From the moment that the student clicks on the button and register, we have an user-id, U. The same user-id cannot enroll in the free trial twice. After 14 days, if the user don’t give up, he becomes a paid user, PU.
As the objective of the experiment is “reduce the number of frustrated students who left the free trial because they didn’t have enough time. This change would maintain the number of students willing to complete the course”, this means, in other words, reducing the proportion U/C without reducing the proportion PU/C.

Invariant Metric

Number of cookies (C): Since the unit of diversion is a cookie, the number of cookies is a good populating sizing metric. It is randomly assigned between the experiment and control groups and the number in each group is approximately the same.

Number of clicks (CK): The number of clicks is a good invariant metric. Since the students click the “start free trial” button before the free trial screener is trigger, the number of clicks will not be affected by the change of the experiment.

Click-through-probability (CK/C): Since the click-through-probability is a relation between two invariant metrics, defined by the number of unique cookies to click the “start free trial” button divided by the number of unique cookies to view the course overview page, it is also an invariant metric.

Evaluation Metric

Gross conversion (U/C): The gross conversion is a good evaluation metric because it is the relationship of two metrics, one after the free-trial screener (number of user-ids to complete checkout and enroll in the free trial) that is expected to change with the experiment and another before the free-trial screener (number of unique cookies to click the “start free trial” button) that is expected to hold constant. If the hypothesis holds true, it is expected a lower gross conversion rate in the experiment group than in the control group.

Net conversion (PU/C): The net conversion is also a good evaluation metric for the same reason of the gross conversion metric, it is the relationship between one metric that is measured after the free-trial screener (number of user-ids to remain enrolled past the 14-day boundary and thus make at least one payment) and another before the free-trial screener (number of unique cookies to click the “start free trial” button). The hypothesis states that the objective is not to reduce PU/C: “This change would maintain the number of students willing to complete the course”. In this case, it is expected that PU/C holds constant or increase.

Unused Metric

Number of user-ids (U): The number of user-ids is not an invariant metric because it is tracked after a student enrolls in the free-trial. The number of user-ids would be different between the experiment and control groups. It could be an evaluation metric, but it doesn’t have a relationship with a metric before the free-trial screener (it is not a measure of proportion). If a change occur in the number of unique cookies to click the “start free trial” button in the control and experiment groups for any other reason, the analysis using this metric could be misleading. So, it is less suitable than the other two evaluation metrics.

Retention (PU/U): With C invariant, the hypothesis is to reduce U without reducing PU. If this occurs, the proportion PU/U should increase. The retention will not be used as evaluation metric

because it takes a long time to evaluate, and the gross conversion and net conversion are enough to validate the hypothesis.

Measuring Standard Deviation

The gross conversion and net conversion metrics are probability metrics and they are assumed to have a binomial (normal) distribution. The analytic estimate of the standard deviation for this type of distribution is defined as:

$$SD = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{N}}$$

Where:

- $\hat{p}$ is the probability or the evaluation metric
- N is the sample size

From the table of baseline values:
The **gross conversion** is the probability of enrolling, given click which is 0.20625.
The **net conversion** is the probability of payment, given click which is 0.1093125.
In order to calculate the sample size of unique cookies to click the “start free trial” button per day, we need to multiply the sample size of cookies visiting the course overview page by the click-through-probability on “start free trial”. Given a sample size of 5000 cookies visiting the course overview page and click-through-probability of 0.08, we have:

$$N = 5000 \times 0.08 = 400$$

The calculated standard deviation for both evaluation metrics are:
Gross conversion:

$$SD_{GC} = \sqrt{\frac{0.20625\,(1-0.20625)}{400}} = 0.0202$$

Net conversion:

$$SD_{NC} = \sqrt{\frac{0.1093125\,(1-0.1093125)}{400}} = 0.0156$$

As the unit of diversion (number of cookies) and the unit of analysis (denominator of the evaluation metrics) of the evaluation metrics are all the same, the empirical variability tends to be closer to the analytic estimate. Therefore, it is not necessary to perform an empirical estimate of the variability in both cases.

Sizing

Bonferroni Correction

When we evaluate multiple metrics simultaneously, the more likely one of them will show a significant difference just by chance. As we increase the number of evaluation metrics, the probability of any false positive increases. To fix this we use a higher confidence level for each individual metric. The Bonferroni correction is the most common method used to estimate the confidence level for each metric, defined by:

$$\alpha_{individual} = \frac{\alpha_{overall}}{n}$$

Where:

- $\alpha_{overall}$ is the desired overall confidence level
- n is the number of metrics

In A/B test, the Bonferroni is applied in cases where several metrics are used to validate an objective. This avoids the probability of one of the tests being false positive, which increases with the number of used metrics. If we are using 10 different metrics to test the impact of a change, with 95% of confidence level, there is a chance of 40% that one of the tests will result in false positive and the experiment will be incorrectly defined as success. In this case, it is important to be conservative and apply the Bonferroni correction.
We won't use the Bonferroni correction for this analysis. Since we have a small number of evaluation metrics (gross conversion and net conversion) and they are correlated (they tend to move at the same time), the Bonferroni correction can be conservative and it is more feasible to analyse each of them separately.

Number of Samples vs. Power

In order to calculate the number of pageviews needed to adequately power the experiment (be statistically significant), first we need to calculate the sample size per variation of the unit of analysis for each evaluation metrics using the Evan's online calculator. Inputting the following values, we obtain:
Gross conversion:

- Baseline conversion rate (probability of enrolling, given click): 0.20625
- Minimum detectable effect (practical significance boundary, d_{min}): 0.01
- Statistical power (1 - beta): 0.8
- Significance level (alpha): 0.05

The calculated sample size per variation is 25835.

Net conversion:

- Baseline conversion rate (probability of enrolling, given click): 0.1093125
- Minimum detectable effect (practical significance boundary, d_{min}): 0.0075
- Statistical power (1 - beta): 0.8

- Significance level (alpha): 0.05

The calculated sample size per variation is 27413.

The number of required pageviews is obtained by dividing the sample size of the unit of analysis by the click-through-probability on “start free trial”. Calculating for both evaluation metrics, we have:
Gross conversion:

$$25835/0.08 = 322937.5$$

Net conversion:

$$27413/0.08 = 342662.5$$

Since we need samples of pageviews for both experiment and control groups, we need to multiply the last result by 2.
Gross conversion:

$$322937.5 \times 2 = 645875$$

Net conversion:

$$342662.5 \times 2 = 685325$$

The number of pageviews necessary to power the experiment is 685325, we chose the larger sample size because we can run the experiment adequately for both evaluation metrics.

Duration vs. Exposure

There is no risk for the participants during the experiment. The introduced feature is only a recommendation for the students to have at least 5 hours per week for studying, they won't have any sort of damage (physical, psychological, emotional, social or economical) with this experiment. Also we are not dealing with sensible data, the collected data are not confidential or would put the participant in risk if the data were exposed to non-trusted individuals. The clients can be informed that the new features will be tested on the website during a period of time and the website will work normally during this period.
For this experiment, we are considering divert 100% of the traffic because we are not putting in risk any participant or group of people and the website would work normally for the students who already use the services.
The number of days necessary to run the experiment is calculated by dividing the number of required pageviews by the number of pageviews per day:

$$685325/40000 = 17.13$$

It is necessary at least 18 days to run the experiment.

Experiment Analysis

Sanity Checks

After collecting the experimental data, the next step is to perform sanity checks in order to ensure that the experiment was run properly. There are two types of invariant metrics, population sizing metrics and any other metrics that is not expected to change. First we need to check the two populating sizing metrics, number of cookies and number of clicks, and see if their samples in the control and experiment groups are equivalent.
These two invariant metrics, number of cookies and number of clicks, are simple counts that can be randomly assigned to the control and experiment group with a probability of 0.5. As the two invariant metrics follow a binomial distribution, we can calculate the confidence interval around the expected probability of 0.5 for the control group (chosen group) and then check if the observed fraction (total number in the control group divided by the total number in both groups) fall within this confidence interval.
From the table Final Project Results, where we can find the results of the experiment, we need to count the total number of cookies and the total number of clicks for the control and experiments groups.
Number of cookies:

- Total number in the control group: 345543
- Total number in the experiment group: 344660

Number of clicks:

- Total number in the control group: 28378
- Total number in the experiment group: 28325

After that we need to calculate the standard deviation of the binomial distribution with probability of 0.5 for both metrics:
Number of cookies:

$$SD = \sqrt{\frac{0.5\ (1-0.5)}{(345543+344660)}} = 0.00060$$

Number of clicks:

$$SD = \sqrt{\frac{0.5\ (1-0.5)}{(28378+28325)}} = 0.00210$$

Given the 95% of confidence interval, the z-score is 1.96. The margin of error is obtained by:

$$m = z \times SD$$

Number of cookies:

$$m = 1.96 \times 0.00060 = 0.00118$$

Number of clicks:

$$m = 1.96 \times 0.00210 = 0.00412$$

Calculating the confidence interval around 0.5 for both metrics, we have:

$$C.I = 0.5 \pm m$$

Number of cookies:

$$lower\ bound = 0.5 - 0.00118 = 0.49882$$
$$upper\ bound = 0.5 + 0.00118 = 0.50118$$

Number of clicks:

$$lower\ bound = 0.5 - 0.00412 = 0.49588$$
$$upper\ bound = 0.5 + 0.00412 = 0.50412$$

Lastly, the formula of observed fraction of both metrics in the control group is:

$$\widehat{p} = total\ control/(total\ control\ +\ total\ experiment)$$

Number of cookies:

$$\widehat{p} = 345543/(345543 + 344660) = 0.50064$$

Number of clicks:

$$\widehat{p} = 28378/(28378 + 28325) = 0.50047$$

The next part of the sanity checks is to verify if the click-through-probability doesn't change. As the click-through-probability is not a simple count that can be randomly assigned between the control and experiment groups with a probability of 0.5 (comparable to a fair coin), but

actually it is a measure of probability or proportion of the students to click the “start free trial” button (comparable to a biased coin), we need to compare the probability estimated on the control side with the probability estimated on the experimental side.
Since the click-through-probability is an invariant metric, we expect that the probabilities in both experiment and control groups are the same, in other words, the difference between the probability in the experiment group and the probability in the control group is zero. To check this, we have to calculate the confidence interval around zero and see if the difference between the probabilities in both groups falls within the confidence level.
First we have to determine the probability across groups using the pooled probability expression:

$$\widehat{p}_{pool} = \frac{X_{cont} + X_{exp}}{N_{cont} + N_{exp}}$$

Where:

- X_{cont} is the total number of unique cookies who click in the control group
- X_{exp} is the total number of unique cookies who click in the experiment group
- N_{cont} is the total number of unique cookies in the control group
- N_{exp} is the total number of unique cookies in the experiment group

The calculated pooled probability is:

$$\widehat{p}_{pool} = \frac{28378 + 28325}{345543 + 344660} = 0.08215$$

Next we calculate the pooled standard error, which is given by the formula:

$$SE_{pool} = \sqrt{\widehat{p}_{pool} * (1 - \widehat{p}_{pool}) * \left(\frac{1}{N_{cont}} + \frac{1}{N_{exp}}\right)}$$

The pooled standard error is:

$$SE_{pool} = \sqrt{0.08215 * (1 - 0.08215) * \left(\frac{1}{345543} + \frac{1}{344660}\right)} = 0.00066$$

Given the 95% of confidence interval, the z-score is 1.96. The margin of error is:

$$m = 1.96 \times 0.00066 = 0.00129$$

The confidence interval around zero is:

$$lower\ bound = 0 - 0.00129 = -0.00129$$

$$upper\ bound = 0 + 0.00129 = 0.00129$$

The difference between the experimental probability and control probability is given by:

$$\hat{d} = \hat{p}_{exp} - \hat{p}_{cont} = \frac{X_{exp}}{N_{exp}} - \frac{X_{cont}}{N_{cont}}$$

Then:

$$\hat{d} = \frac{28325}{344660} - \frac{28378}{345543} = 0.0000566$$

The following table is a summary of the calculated values and the status indicating if the metric passes in the sanity check.

	Lower bound	Upper bound	Observed	Passes
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks	0.4959	0.5041	0.5005	Yes
Click-through-probability	-0.0013	0.013	0.0001	Yes

Since the observed fraction of the 3 metrics are within the confidence interval, all the metrics pass in the sanity check. Therefore, we can move forward and analyse the experiment results.

Result Analysis

Effect Size Tests

The result analysis consists in check if the experiment positively impacted in the evaluation metrics. In order to do that, we need to evaluate if the change was statistically significant using **hypothesis testing**. Hypothesis testing is a quantitative way to determine how likely it is that the results are due to chance. We need to examine two opposing hypotheses to calculate this probability: the null hypothesis and the alternative hypothesis. The null hypothesis or H_0 states that there is no difference or no effect between the experiment and control groups. The alternative hypothesis or H_A, is the statement that the experiment had an effect. For the evaluation metrics, we have to calculate the difference between the experiment and control groups, then calculate the confidence interval around this difference and check if the confidence interval doesn't include the zero. If that is the case, the metric is statistically significant.
From the table Final Project Results, the total number of clicks, the total number of enrollments and the total number of payments for the control and experiment groups (from October 11 to November 2) are:
Number of clicks:

- Total number in the control group: 17293

- Total number in the experiment group: 17260

Number of enrollments:
- Total number in the control group: 3785
- Total number in the experiment group: 3423

Number of payments:
- Total number in the control group: 2033
- Total number in the experiment group: 1945

First we need to calculate the pooled probability for both metrics:
Gross conversion:

$$\widehat{p}_{pool} = \frac{3785 + 3423}{17293 + 17260} = 0.20861$$

Net conversion:

$$\widehat{p}_{pool} = \frac{2033 + 1945}{17293 + 17260} = 0.11513$$

Next we calculate the pooled standard error:
Gross conversion:

$$SE_{pool} = \sqrt{0.20861 * (1 - 0.20861) * \left(\frac{1}{17293} + \frac{1}{17260}\right)} = 0.00437$$

Net conversion:

$$SE_{pool} = \sqrt{0.11513 * (1 - 0.11513) * \left(\frac{1}{17293} + \frac{1}{17260}\right)} = 0.00343$$

Given the 95% of confidence interval, the z-score is 1.96. The margin of error for both metrics is:
Gross conversion:

$$m = 1.96 \times 0.00437 = 0.00856$$

Net conversion:

$$m = 1.96 \times 0.00343 = 0.00672$$

The difference between the experimental probability and control probability is:
Gross conversion:

$$\widehat{d} = \frac{3423}{17260} - \frac{3785}{17293} = -0.02055$$

Net conversion:

$$\widehat{d} = \frac{1945}{17260} - \frac{2033}{17293} = -0.00487$$

The confidence interval around the difference is:
Gross conversion:

$$lower\ bound = -0.02055 - 0.00856 = -0.02911$$
$$upper\ bound = -0.02055 + 0.00856 = -0.01199$$

Net conversion:

$$lower\ bound = -0.00487 - 0.00672 = -0.01159$$
$$upper\ bound = -0.00487 + 0.00672 = 0.00185$$

The following table is a summary of the calculated confidence intervals for both evaluation metrics and the status indicating if the experiment has statistical and practical significances.

	Lower bound	**Upper bound**	**Statistical significance**	**Practical significance**
Gross conversion	-0.0291	-0.0120	Yes	Yes
Net conversion	-0.0116	0.0018	No	Yes

Since the confidence interval of gross conversion doesn't include zero, the experiment for this metric is statistically significant. However, the change in net conversion is not statistically significant because its confidence interval includes zero.
In addition, we need to check, from a business perspective, if the change in the gross conversion is practically significant. To check this, the confidence interval may not include the practical significance boundary (d_{min}). The practical significance boundary of gross conversion is 0.01, so this metric is also practically significant.
One part of the hypothesis is that the net conversion should not be reduced. However, the confidence interval of net conversion includes the practical significance boundary (d_{min}), which is -0.0075 (as we are testing whether the metric didn't reduce, the limit has negative sign). This implies that there is a risk that the net conversion was reduced by a percentage considered significant from the point of view of business.

Sign Tests

The p-value of the sign test is the probability of obtaining a result as extreme as, or more extreme than, the result actually obtained by chance (when the null hypothesis is true). In order to calculate the p-value we need to count the number of success of each day in the day-by-day data which is defined when the experiment group has a lower value of evaluation metric than the control group and then input the values in the Online Calculator.
Gross conversion:

- Number of trials or experiments (number of days): 23
- Number of success (number of days with positive change): 19
- Probability of success in each trial or subject: 0.5 (for sign tests)

The calculated two-tail p-value is 0.0026.

Net conversion:

- Number of trials or experiments (number of days): 23
- Number of success (number of days with positive change): 13
- Probability of success in each trial or subject: 0.5 (for sign tests)

The calculated two-tail p-value is 0.6776.

The following table is a summary of the calculated p-value for each evaluation metrics and if they have statistical significance. Since the calculated p-value of gross conversion is less than the significance level (alpha), which is 0.05, the sign test agrees with the hypothesis test indicating that this result is unlikely to come about by chance.

	p-value	**Statistical Significance**
Gross conversion	0.0026	Yes
Net conversion	0.6776	No

Summary

In this project, we have two different metrics, gross conversion and net conversion, which are used to evaluate two different objectives, the hypothesis is composed of two parts. The first evaluate whether the proportion between the clicks on the “free trial” button and the number of registered students on the free trial was reduced. The second evaluates whether the proportion between clicks on the “free trial” button and the number of students who became paying was not reduced. Each metric evaluates a different objective, and both objectives have to be reached for the experiment to be considered a success. Therefore, it would be conservative apply the Bonferroni correction for this experiment.
Both the effect size test and the sign test revealed the same results, the gross conversion metric is statistically significant and the net conversion is not. However, the reduction in net conversion

is higher than the practical significance boundary, this implies that the net conversion has a practical effect in terms of business.

Recommendation

I would not launch the experiment yet, I would dig deeper in the investigation. Although the net conversion is not statistically significant, which is desirable for the experiment, it is practically significant (the calculated confidence interval for net conversion includes the practical significance boundary), indicating that there is a risk, though small, that the introduction of the new feature causes a reduction in the net conversion and results in financial impact.
If we evaluate that the introduced feature is very important to the business, it is possible to rebuild the test with more statistical power by reducing the confidence interval of the effect size test. With the smaller confidence interval, we have more precise information about the size and direction of the variation in the evaluation metrics.

Follow-Up Experiment

In order to reduce the number of frustrated students who cancel early in the course, I would test an experiment where the students who decided to continue in the course (remain enrolled past the 14-day boundary) receive a message with a schedule and a text. This text would include tips on planning and organization and would emphasize the importance to dedicate on the course. In addition, I would include that the students don't need to strictly follow the schedule and if they had any problem with the time, the content of the subjects or any other reason to send an email to the support.
The hypothesis is that the message would help the students to find a free time in their schedule to dedicate to the course and the support would help them to solve any other problems. Since the students would be more aware of their commitments and would receive support if they needed, this would encourage them to complete the course.
The evaluation metric that I would choose for this experiment is the number of user-ids to complete the course divided by the number of user-ids to remain enrolled past the 14-day boundary (make at least one payment). This evaluation metric is a relationship between one metric that is measured after the message is sent (number of user-ids to complete the course) that is expected to change with the experiment and another before the message (number of user-ids to remain enrolled past the 14-day boundary) that is expect to be constant. If the hypothesis holds true, the rate of the evaluation metric in the experiment group should be higher than in the control group.
The number of user-ids to remain enrolled past the 14-day boundary is the invariant metric and also the unit of diversion of the experiment. It will be randomly assigned between the experiment and control groups and the number in each group will be approximately the same.