P7: A/B testing

Elias Laura Wiegandt

Conducting A/B test of Udacity website update

April 7, 2017

1 Experiment Design

The objective of this experiment is to decrease the number of students who enroll in the free trial without decreasing the number of students who move from the free trial to making their first payment for the course.
Throughout the report the students/possible students will generally be referred to as "cookies" as unique cookies are the unit of diversion.

1.1 Metric Design

The goal of the experiment is to check if students who will leave the course during the free trial can be sorted out earlier, by issuing a warning if they state they will only work '0-5 hours a week on the course.
The chosen metrics can be seen in table 1.

Table 1: Metrics used for A/B test

Invariant metrics	**Variant metrics**
Number of cookies	Gross conversion rate
Number of clicks	Net conversion rate

1.1.1 Invariant metrics

The invariant metrics are expected to not change between the experiment and the control group, as the difference in the website between the two groups appears *after* assignment of cookie and clicking on the "start free trial" button.
As unique cookie is the unit of diversion, and the control group is assumed to be of the same size as the experiment group, the number of cookies in each group have to be close to each other. If they are not, there is an issue in the mechanism that assigns cookies to the experiment and control group. The number of clicks on the "start free trial" button should be equal between the two groups, because it is assumed that people are assigned randomly to

the control and experiment groups. This implies the two groups should "be alike", including equally likely to press the button. If this turns out not to be the case, the assignment to test and control group might not be completely random.
Other invariant metrics were considered. Most notably, this include the "click-through-probability". This probability measures the number of unique cookies who click the "start free trial" button, divided by the total number of unique cookies to view the course overview page. But this rate is already indirectly caught by the two invariant metrics already used, as it is just one divided by the other.

1.1.2 Variant metrics

The variant metrics are the metrics that measures whether the experiment alters behavior, and are often also refereed to as the "evaluation metrics". The gross conversion rate is defined as the percent of cookies who clicks the "start free trial" button and enrolls on the same day. The net conversion rate denotes the number of cookies who completes the free trial and make their first payment.
The gross conversion rate can be affected because the warning causes either fewer or more cookies to enroll in the free trial. Intuitively, if there is an effect from the warning, it seems most likely that it cause the gross conversion rate to fall. But there is no compelling reason to ignore the possibility of an increase. This could happen because cookies view the course as more "serious" if it requires more work per week, causing them to believe it is of higher quality and therefore make them more likely to enroll in the free trial.
The net conversion rate could both increase or decrease. If fewer cookies enroll in the free trial, but the cookies' probability of paying given they enrolled (the "retention" rate) does not increase, the net conversion rate will decrease. On the other hand, if gross conversion rate falls, but the cookies who actually enroll are more likely to reach the payment, the retention rate will increase. Depending on the sizes of the fall in gross conversion and increase in retention rate, this can cause net conversion rate to either increase, decrease or stay unchanged.

1.1.3 Launch of experiment

The purpose of the experiment is to find out, if the experiment group have a decreased gross conversion rate without having a decreased net conversion rate. This is consistent with the goal of the experiment, as described at the very beginning of this report.
If we are confident that the experiment group's gross conversion rate falls more than our practical significance boundary, while we are also confident that the experiment group's net conversion rate *does not* fall more than its practical significance boundary, we should launch the experiment.

1.1.4 Discarded metrics

From the above description, the retention rate appear important, and clearly could have been a valid variant metric. But, as will be shown later, this metric requires a much larger number of unique cookies to be with satisfactory significance and power, relative to gross and net conversion rates, and is therefore not feasible.

Another possible variant metric was the total number of users who enroll in the free trial, denoted "number of user-ids" in the project material. If the test and control groups are of equal size and assigned randomly, this would be comparable between the two groups. But using this variable would require making an assumption about its distribution. With the net and gross conversion rates, it is assumed they follow a binomial distribution. This should not be a problematic assumption, since whether you enroll or pay can easily be translated to a "succes/failure" methodology. But this is not the case for the number of user-ids, which would be a daily number that varied from day to day. It *could* be normally distributed, but could also follow e.g. a power law or another distribution which it would be difficult to work with. As there would also be significantly fewer observation than for the gross and net conversion rates (same as with retention, see the description of this below), this metric is not a good metric for evaluating the experiment.

1.2 Measuring standard deviation

The standard deviation of a variable with a binomial distribution is given by:

$$SE = \sqrt{\frac{\beta\left(1-\beta\right)}{N}} \tag{1}$$

where β is the probability of success and N is the number of observations. β is estimated with $\hat{p}$. The task states there is 5000 cookies in the sample. As the gross and net conversion rates are calculated given a person clicks the "start free trial", the N to be used in equation 1 is 5000 times the "click-through-probability". For the net conversion rate the calculation is thus:

$$N = 5000 \cdot 0.08 = 400 \tag{2}$$

$$SE = \sqrt{\frac{0.109313\left(1-0.109313\right)}{400}} = 0.0156 \tag{3}$$

The result for both evaluation metrics an be seen in table 2.

Table 2: Standard deviation of evaluation metrics

Metric	**Standard deviation**
Gross conversion rate	0.0202
Net conversion rate	0.0156

The analytical variation is likely to be close to the empirical variation. How different they are largely depends on whether or not we have independent sampling. Since the unit of diversion, unique cookies, matches the unit of analysis, also unique cookies, we can assume that our samples are independent, and thus be confident that the analytical variation is close to the empirical variation.

1.3 Sizing

1.3.1 Bonferoni correction

The Bonferoni correction is not used. This would normally be the case when testing several metrics, since the chance of a finding a significant variable by sheer chance increase the more variables are tested. The correction divides the desired α with the number of evaluation metrics. But this is not reasonable in this experiment, because we are looking for a changed gross conversion rate, but an unchanged net conversion rate. Increasing the required significance level by using the Bonferoni correction would simultaneously make it more likely that a change in either where not deemed significant. In the worst-case scenario, if gross conversion rate is more significantly changed than net conversion rate, this could cause us to accept the gross conversion rate is changed but net conversion was not, while net conversion was actually also changed.
Although the Bonferoni correction is not used, confidence intervals with the test effect sizes using the Bonferoni correction have been included later, for comparison.

1.3.2 Needed page views

To calculate the needed page views, the current rates of conversion ($\hat{p}$), a minimum required effect size (d_{min}), a required α and a required power (β), are needed. These number are all given in the project descriptions. Using this webpage, the corresponding "unique cookies who click the button" are calculated. This is then transformed into required page views using the probabilities from the project description. Finally, it is assumed that the test and control group are of equal size, and therefore the result has to be timed by two.

Table 3: Input to sizing calculation

Metrics	α	1-β	$\hat{p}$	d_{min}
Gross conversion rate	0.05	0.80	0.20625	0.0100
Net conversion rate	0.05	0.80	0.109313	0.0075

The constant used for the calculation can be seen in table 3. The transformation from "cookies who click button" to the necessary "Total page views" can be seen in table 4. From this table, it can be seen that the net conversion rate requires 685,325 page views, slightly more than the gross conversion rate, due to the different $\hat{p}$ and d_{min} between the two evaluation metrics. As it is the maximum of the required page views that define the required page views for the entire experiment, these 685,325 are the total required page views to run the experiment, with the chosen significance and power.

1.3.3 Duration

The warning is only shown to cookies who state they only expect to work between 0 and 5 hours a week on the course, so it will only affect a subset of the experiment group. Further, the experiment does not collect sensitive information. Cookies who sign up can possibly be identified through the information they submit during enrollment, but this was the case all

Table 4: Sizing calculation

Metric	Cookies who click button	Test page views	Total page views
Gross conversion rate	25,835	322,938	645,875
Net conversion rate	27,413	342,663	685,325

along. No new information is collected. Finally, the experiment is not likely to hurt anyone. In theory, it could happen that someone do not sign up for the free trial due to the warning, but would have gone on to have a very successful career in data science if they had signed up. But this is far fetched, and the risk to cookies must be considered minimal. The risk to Udacity is that cookies get a negative view of Udacity due to the warning, but this risk should also be minimal (if one warning can kill the brand, it's a pretty bad brand).
It is therefore assumed that this experiment only entails a low risk for both Udacity and the unique cookies.
Thus, the fraction of traffic diverted to the test and control group is set to 1, e.g. all traffic. With a daily traffic of 40,000 unique cookies and 685,326 unique cookies, this results in a duration of 17.3 days. Rounded up, this gives a total of 18 days.
As mentioned earlier, including the retention rate was considered, but would required 4,741,212 page views. The experiment would then have to run for 119 days. This is unacceptably long, hence the retention rate was not chosen as an evaluation metric.

2 Experiment analysis

2.1 Sanity checks

The invariant metrics are checked. The expectation is that they are not significantly different between the test and the control group. The results can be seen in table 5. All tests passes. This increase the confidence in the test results.

Table 5: Sanity checks

Metric	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

2.2 Result analysis

2.2.1 Effect size test

The effect size of the evaluation metrics are calculated, and the confidence interval of the effect size is used to conclude if the effect is either statistically of practically significant. The result can be seen in table 6.

Table 6: Effect size test

Metric	Lower bound	Upper bound	Statistical significance	Practical
Gross conversion rate -0.0291	-0.0120	Yes	Yes	
Net conversion rate	-0.0116	0.0019	No	No

The 95 percent confidence interval for the gross conversion rate neither includes 0 (the limit for statistical significance) nor -0.01 (the limit for practical significance), and the variable is thus both statistically and practically significant.
But the 95 percent confidence interval for the net conversion rate includes both 0 (the limit for statistical significance) and -0.0075 (the limit for practical significance). But we do not want this variables to fall below -0.0075, and with the current confidence interval, 22 percent of the probability mass appear *beneath* that threshold. The variable is thus not significant, but we cannot, with 95 percent confidence, rule out that the experiment has an effect on the net conversion rate that is greater than $d_{min} = -0.0075$.
For comparison, the confidence interval with the Bonferoni correction is calculated, and can be seen in table 7. The results does not change, but as stated earlier, implementing the Bonferoni correction is not deemed necessary.

Table 7: Effect size test, with Bonferoni correction

Metric	Lower bound	Upper bound	Statistical significance	Practical
Gross conversion rate -0.0304	-0.0108	Yes	Yes	
Net conversion rate	-0.0126	0.0028	No	No

2.2.2 Sign test

A sign test is calculated on the evaluation metrics. This is done because it does not require any parametric assumptions, and therefore can serve as a check on the credibility of the effect sizes found in the section above. To calculate the sign test, this webpage is used. The field "proability" is set to 0.5. This denotes the expected probability of the experiment observations being higher than the control observations under the null hypothesis. Setting it to 0.5 implies that they, under the null, would be expected to be very close to each other. The results can be seen in table 8. The results are compliant with the results from the effect size test. The test indicates that the control and experiment observations for gross conversion rate are systematically different, while no difference between the control and experiment observations for net conversion rate can be identified.

2.2.3 Summary

The sanity checks are passed, the results of the effects size test and the sign test agrees. The Bonferoni correction was not used, but using it does not alter the results. Gross conversion

Table 8: Sign test

Metric	p-value	Statistical significance
Gross conversion rate	0.0026	Yes
Net conversion rate	0.6776	No

rate decreases significantly, but we cannot rule out that the net conversion rate decrease below the practical significance boundary of -0.0075.

2.3 Recommendation

There is statistically significant evidence that the experiment decrease the gross conversion rate below the practical significance boundary. But there is not statistically significant evidence that the net conversion rate will **not** decrease below its practical significance boundary. From here on, it depends on the cost of coaching students who do not stay enrolled into the payment period, compared to the payment. If management are willing to run the risk of lower net conversion rate, the experiment should be launched.
Looking into other options appear to be a better call. This could be either increasing the test size or trying another experiment. The test size could be increased, to try and achieve lower standard errors and smaller confidence intervals. Given the current lower bound for the net conversion rate of -0.0116, this would demand large amount of new observation. Trying another experiment seems like a better idea.
In conclusion, I would recommend not launching. There is a large network effect in Udacity's business model; the more who complete a course or nano degree the better. Running the risk of a lower net conversion rate is a bad idea, and in the worst case scenario it could lead to a lower graduation rate. Trying another experiment is a better idea.

3 Follow-up experiment

An experiment could be to put links to relevant forum posts beneath quizzes.
I believe many students do not learn to use the forums until after they have made their first payment. But finding information on the forums, instead of e.g. Google, is often quicker.
This could be implemented by having links to 3 relevant forum posts (ranked by e.g. number of posts, key words or how recent they were) below each quiz. Frustrated students looking for hints below the quiz would then discover these posts, and hopefully either find information in them or use them as a gateway to other posts on the forum. In the best case scenario, this would reduce the rate of frustrated students.
The hypothesis would be, that students shown links to relevant forum posts below quizzes would be less frustrated and therefore more likely to make their first payment.
Since the experiment takes place *after enrollment*, the gross or net conversion rates cannot be used. The best evaluation metric would be the retention rate, as this is what the experiment seeks to increase. The unit of diversion would be user-id. It was shown earlier that the retention rate required a much larger amount of page views, which indicates the experiment

might not be feasible.
An invariant metric that could be used to test the experiment setup could be the number of new user-ids per day. If the experiment was correctly set up, this metric should not be significantly different between the control and experiment groups. Further, as user-id is also the unit of diversion in this experiment, we could actually test if the rate of assignment to the control and experiment groups follows a binomial distribution with 0.5 probability for each outcome (assuming that the control and experiment groups were supposed to be of equal size).