Udacity, A/B Testing

Alan Mosca

July 9, 2016

1 Metrics Choice

1.1 Invariant Metrics

- Number of Cookies This was chosen because it is an easy way of identifying unique visitors, which makes it a good choice for a unit of diversion. It has not been chosen as an evaluation metric because it shouldn't be affected by the change that is being evaluated.
- Number of Clicks This was chosen as an invariant (and not for evaluation) because it also isn't affected by the change that is being evaluated.

1.2 Evaluation Metrics

- Gross Conversion This measure is directly affected by the experiment, and it helps to determine if the number of people that click on the sign-up link for the trial and then leave after it ends can be reduced.
- Net Conversion This measure is directly affected by the experiment, and it helps to determine if the number of people that click on the sign-up link and pay is affected.

1.3 Discarded Metrics

- Number of user-ids This measure is affected by the experiment, as we can't expect the number of user-ids to be split evenly between control and experiment. It could be an evaluation metric because it measures the number of students that make it past the free trial, but it is not normalized by the number of clicks, so we prefer other measures.

- Retention This was initially selected as an evaluation metric because it directly measures the effect of the experiment, but using it would have made the length of the experiment 132 days, which was too high, so it has been subsequently discarded. It can't be used as an invariant because the experiment affects this value.
- Click-through-probability This metric is not affected by the experiment, given that the click happens before the diversion. Nevertheless, it does not tell us anything about the experiment and is not a unit of diversion, so it is not very useful.

1.4 Goal

The goal of the experiment is to determine whether adding the question "are you able to dedicate at least 5 hours per week to your studies" reduces the number of people that sign up for a trial and then leave before paying for the first time. Additionally, we want to know whether adding this question would prevent people who will eventually go through to payment from signing up. We can therefore state our null hypothesis h_0 to be that adding the question to the process does not affect the proportion of people signing up for the trial, or the proportion of people paying wrt the number of people clicking on the sign-up button. Broken down wrt the two measures, we define:

- $h_0(gc)$: the Gross Conversion is not affected by the experiment
- $h_0(nc)$: the Net Conversion is not affected by the experiment

In order to accept the change, we therefore want to ideally reject the null hypothesis on Gross Conversion $h_0(gc)$ with a *negative* value of $\hat{d}$ and accept the null hypothesis on Net Conversion $h_0(nc)$, or accept it with a *positive* value of $\hat{d}$ (which means that revenue would actually increase).

2 Variability

Given the following data about the experiment:

- Unique cookies to view page per day $N_{view} = 40000$
- Unique cookies to click "Start free trial" per day $N_{click} = 3200$

- Enrolments per day $E = 660$
- Click-through probability on "Start free trial" $P(click) = 0.08$
- $P(E|click) = 0.20625$
- $P(Pay|E) = 0.53$
- $P(Pay|Click) = 0.1093125$

If we assume a sample size $S = 5000$ unique cookies visiting the page, based on the original data, the number of clicks would be:

$$\hat{N}_{click} = S * P(click) = 400 \tag{1}$$

and the number of enrolments:

$$\hat{N}_{enroll} = S * P(click) * P(E|click) = 82.5 \tag{2}$$

Given that all the chosen metrics can be assumed to follow a Binomial distribution, the estimate of Standard Error follows:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \tag{3}$$

so for each metric, the SE is:

- Gross Conversion: 0.020231
- Net conversion: 0.015602

Given that Gross Conversion and Net Conversion are both using the unit of diversion as their denominator, I would expect the analytical variance to match the empirical one.

3 Sizing

3.1 Number of Samples

With a type I error rate of $\alpha = 0.05$ and a type II error of $\beta = 0.2$, the minimum detectable effects are $d_{gc,min} = 0.01$, $d_{r,min} = 0.01$, $d_{nc,min} = 0.0075$. I will not be using the Bonferroni correction, because the measures are covariant.

The ratios of clicks to pageviews is $\frac{N_{clicks}}{N_{views}} = 0.08$ and the ration of conversions to pageviews is $\frac{N_{enroll}}{N_{views}} = 0.0165$, which will be used for corrections.

The non-corrected experiment sizes are as follows:

- Gross Conversion: 25835
- Net Conversion: 27413

After the corrections they are as follows:

- Gross Conversion: 322938
- Net Conversion: 342662.5

Doubling the maximum because we need this number of pageviews for each hypothesis, we get a total of 685325 pageviews required.

3.2 Duration vs Exposure

The experiment does not affect any of the pages with content, so it does not affect existing users, and it is only adding a small prompt that wouldn't be considered annoying by most people. It cleary doesn't hurt anyone because the people that would be discouraged by the pop-up message are not likely to finish the course and receive a certification, so they would not be missing out. There is also no sensitive data being used for the experiment. Therefore, it makes sense to divert a large amount of traffic to it to reach a conclusion quickly. However, it also makes sense to keep a small amount of traffic out of the experiment, in case there was an error or bug in the experiment, so that it can be detected quickly. Therefore I would say that diverting 90% of Udacity traffic makes sense. Give a daily traffic of 40000 pageviews, this means there would be 36000 pageviews dedicated to the experiment. The total duration of the experiment would be of 19.03 days, which will be rounded to 20. This is an acceptable time.

4 Sanity Checks

We used a 95% confidence interval, with a Z-score of 1.96. For both measures we expect a probability of 0.5.

For the number of cookies, we had a SE of 0.0006, giving a margin of error of 0.0012 and a confidence interval of $[0.4988, 0.5012]$. The observed value was 0.5006, so it is considered a pass.

For the number of clicks, we had a SE of 0.0021, giving a margin of error of 0.0041 and a confidence interval of $[0.4958, 0.5041]$. The observed value was 0.5005, so it is considered a pass.

5 Effect Size Tests

The measured values from the experiment are as follows:

5.1 Gross Conversion

- $P(gc|control) = 0.2189$
- $P(gc|experiment) = 0.1983$
- $\hat{d} = -0.0206$
- $SE = 0.0044$
- $ME = 8.5652 \times 10^{-3}$
- $CI = [-0.0291, -0.0120]$
- $d_{min} = 0.01$

Gross Conversion is therefore both statistically and practically significant. We therefore conclude that we have observed a valid change in Gross Conversion.

5.2 Net Conversion

- $P(nc|control) = 0.1176$
- $P(nc|experiment) = 0.1127$
- $\hat{d} = -0.0049$
- $SE = 3.4340^{-3}$
- $ME = 6.7228 \times 10^{-3}$
- $CI = [-0.0116, 0.0018]$
- $d_{min} = 0.0075$

Net conversion is not statistically, or practically, significant. This effectively means that we do not see a change in Net Conversions. I did not use the Bonferroni correction, for the same reason as it was not used in the rest of the experiment (the metrics are covariate).

6 Sign Tests

6.1 Gross Conversion

- $d_{min} = 0.01$
- $N_{success} = 4$
- $N_{total} = 23$
- $p = 0.0026$
- $\alpha = 0.025$

Therefore the sign test also reports significance.

6.2 Net Conversion

- $d_{min} = 0.075$
- $N_{success} = 10$
- $N_{total} = 10$
- $p = 0.6776$
- $\alpha = 0.025$

This confirms the statistical insignificance of the difference.

7 Results Summary

The sign test confirms the results of the effect size test. I did not use the Bonferroni correction, because to accept or reject our null hypothesis we are considering both Gross and Net Conversion. Therefore the logic operation between the two sub-hypotheses is a conjunction, which means that it is not necessary to apply the Bonferroni correction.

8 Recommendation

The reduction in Gross Conversions is part of what was expected of the new change. The number of people signing up would be reduced, and there is no significant change in the Net Conversions. This means that there are less people signing up and then dropping out. Conversely, the lack of significance on the Net Conversions also means that there is no significant lost (or gained) revenue. This means that there is inherent savings in saved resources and a more focused trial. However, the confidence interval on Net Conversions exceeds the practical significance boundary on the negative side. As things stand, I do not recommend launching the change, because there is still a possibility that it would lead to a loss of revenue. Rather than completely abandoning the experiment, I would recommend collecting more data to reduce the boundary on this metric until its significance is clearer.

9 Follow-up Experiment

If the main concern is for people to be able to pass their initial trial, I would recommend a "mid-trial" quiz, where the user can answer a few questions about their expectations wrt their time commitment and if they think they will be able to dedicate enough time to the course, and then offer them the option to *freeze* the trial for a month. The hypothesis is that the user will be able to gauge the commitment needed for the course and reassess at a later date. The metrics used for this experiment would be number of users (invariant) and retention rate (evaluation), and the unit of diversion would be the user-id, because at this point the student is logged in and can be tracked with precision.