Final Project A/B Testing

This is the final project about A/B Testing. We will design and analyze an experiment that was run by Udacity. For specific details about the experiment, please refer to the document *Experiment.pdf*. The figures were calculated in the Excel file *Final_Project_Calculations.xlsx*.

Experiment Design

Metric Choice

Invariant Metrics:

- Number of cookies (number of unique cookies to view the course overview page)
- Number of clicks (number of unique cookies to click "Start free trial")
- Click-through-probability (number of unique cookies to click "Start free trial" / number of unique cookies to view page)

Evaluation Metrics:

- Gross Conversion (UserIDs that enroll in free trial / cookies that click on start free trial)
- Net Conversion (UserIDs that pay / cookies that click on start free trial)

Number of cookies - Is an invariant metric because the change of the sign in process after clicking on the “Start free trial” button should not impact this metric. It is randomized between experiment and control group. It is a bad evaluation metric as it is not different between experiment and control.

Number of user-ids – Is not an invariant metric because it is very likely to be different between experiment and control group. Could be chosen as an evaluation metric but is not my preferred metric because I prefer normalized indicators such as Gross Conversion. Since the numerator of Gross Conversion is the same as the Number of user-ids they would move in a similar direction. The Gross Conversion is more precise (normalized) and sufficient for this A/B Test.

Number of clicks – Is an invariant metric because the change of the sign in process after clicking on the “Start free trial” button should not impact the number of cookies to click the “Start free trial” button. It is randomized between experiment and control group. It is a bad evaluation metric as it is not different between experiment and control.

Click-through-probability – Is an invariant metric because we expect it to be equal between experiment and control as the number of cookies and number of clicks are invariant metrics.

Gross Conversion – Is an evaluation metric because we can expect this metric to change due to the alert that the workload of a nanodegree is considerable. The students are “warned” before they actually enroll in the course. We want to reduce the number of frustrated students who left the free trial because they did not have enough time. Is not an invariant metric as there is a high chance that it will change significantly.

Retention – Can be considered as evaluation metric as it might change due to the experiment. It has not been chosen because this metric requires a large sample size since the unit of diversion (cookies) and unit of analysis (user-ids) are different from each other.

Net Conversion – Is an evaluation metric because we would like to determine if the modified sign in process causes a decrease in number of user-ids who pay. This metric is very good to prove the hypothesis of the A/B Test that the number of students who pay does not change significantly.

Measuring Standard Deviation

Standard Deviations:

- Gross Conversion: 0.0202
- Net Conversion: 0.0156

We can expect to see significant differences between the analytically and empirically computed variability of our metric if the unit of analysis is different from the unit of diversion. In our case the unit of analysis is cookies. Similarly, the unit of diversion is a cookie up to the point when the students enroll in the free trial. Since the unit of analysis and unit of diversion are not different from each other, we can expect that the analytic estimate will be comparable to the empirical estimate.

Sizing

Number of Samples vs. Power

At this point, I decided to de-select Retention as an evaluation metric since we would need 4.741.212 cookies to view the course overview page. If we diverted 50% of the traffic to this experiment, we would need 238 days to run the experiment. This is far too long. Additionally, Retention is very similar to Net Conversion which is why we decide to not use Retention as evaluation metric.

With the remaining two evaluation metrics Gross Conversion and Retention we “only” require 685.325 pageviews.

I did not use the Bonferroni correction since this correction is very conservative and neglects the statistical significance very often.

Duration vs. Exposure

I would divert 60% of the traffic to this experiment. The experiments will last 29 days.
I would not divert more traffic to the experiment because we do not know exactly how the users react to our change. Maybe our hypothesis is not true and the Net Conversion reduces drastically, then we would not launch the change and do not want 100% of our traffic to be impacted by that experiment. However, this experiment does not imply a high risk for the users. Nobody can be harmed and we do not create any sensitive data. Therefore, we could still increase the fraction of traffic for this experiment if the duration of the experiment was too long.

Nevertheless, this is a reasonable tradeoff between proportion of traffic and duration of the experiment.

Experiment Analysis

Sanity Checks

95% Confidence Intervals:

- Number of cookies: [0.4988, 0.5012], observation: 0.5006, passed
- Number of clicks on “Start free trial”: [0.4959, 0.5041], observation: 0.5005, passed
- Click-through-probability on “Start free trial”: [-0.0013, 0.0013], observation: 0.0001, passed

Every sanity check passed.

Result Analysis

Effect Size Tests

95% Confidence Intervals:

- Gross Conversion: [-0.0291, -0.012], dmin = |0.01|, statistically and practically significant
- Net Conversion: [-0.0116, 0.0019], dmin = |0.075|, statistically and practically not significant

Sign Tests

P-Values:

- Gross Conversion: p-value = 0.0026, 0.0026 < α, statistically significant
- Net Conversion: p-value = 0.6776, 0.6776 > α, statistically not significant

Summary

I did not use the Bonferroni correction to compute the confidence intervals of the evaluation metrics. This method should not be used routinely as it is very conservative which makes it very difficult to reach statistical significance. It should be considered in case of multiple metrics which might result in the fact that we see a significant difference just by chance. We take into account only two different metrics which is why the Bonferroni correction is not needed.

The effect size test and sign test results support each other and confirm a statistically significant impact on the Gross Conversion. However, the Net Conversion is not impacted significantly.

Recommendation

I would recommend launching the new feature and modify the enrollment process. The reason for that recommendation can be explained by looking at the original idea of the experiment.

The idea was to set clear expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time. As the Gross Conversion decreased significantly both from a statistical and a business point of view, we can assume that less students will leave the free trial due to workload issues.

However, we do not intend to significantly reduce the number of students to continue past the free trial and eventually complete the course. The Net Conversion lies in a confidence interval between -0.0116 and 0.0019. Since the confidence interval contains 0, it is statistically not significant. It may lead to small decreases in Net Conversion and the lower bound exceeds the practical significance boundary of |0.075|. Therefore, in a 95% confidence interval it is possible that the Net Conversion decreases significantly.

Based on these results, we cannot recommend launching the new feature. We would need to run more tests to gather additional data.

Follow-Up Experiment

I would like to run the following experiment: we could change the course overview page and display already at this stage, without asking the user to enter the time they can devote to the course, that Udacity Nanodegrees require a substantial time investment of at least 10h per week.

The hypothesis remains similar: the idea is to go one step further compared to the original A/B Test by setting very clear expectations.
We would use cookies as unit of diversion. This unit best fits to this experiment because it is very precise (one cookie per browser and device).
My invariant metric would be number of cookies to view the course overview page as I do not expect this value to change.
In terms of evaluation metrics we could use Click-through-probability (number of unique cookies to click "Start free trial" / number of unique cookies to view page), Gross Conversion (UserIDs that enroll in free trial / cookies that click on "Start free trial") and Net Conversion (UserIDs that pay / cookies that click on "Start free trial").

I expect the Click-through-probability to decrease significantly as we indicate the time investment already on the course overview page. As in the original A/B Test, the Gross Conversion is also very likely to decrease. However, I do not expect the Net Conversion to decrease significantly because for people who are really interested in doing the Nanodegree (those who actually pay for it), it is not a big surprise that it will require a lot of time to learn new skills.

Bibliography – Literature and Sources

- Richard A. Armstrong: When to use the Bonferroni correction, source: https://onlinelibrary.wiley.com/doi/pdf/10.1111/opo.12131