AB Testing Final Project

Gabor Sar

March 2016

1 Experiment Design

1.1 Metric Choice

1.1.1 Number of Cookies

The number of cookies is the number of unique cookies to view the course overview page. It is a population sizing metric, and it should be split evenly between the control and experiment groups. Therefore, I will use it as an invariant metric. Since this metrics is not expected to differ between the control and experiment groups, I will not use it as an evaluation metric.

1.1.2 Number of Clicks

The number of clicks is the number of unique cookies to click on the "start free trial" button. Since students click on that button before the free trial screening appears, the number of clicks should not be affected by the experiment, and should not differ between the control and experiment groups. Therefore, I will use it as an invariant metric. Since this metrics is not expected to differ between the control and experiment groups, I will not use it as an evaluation metric.

1.1.3 Number of User-IDs

The number of user-IDs is the number of users who have enrolled in the free trial. Since this metric can be affected by the experiment, I will not use it as an invariant metric. It could be used to detect an absolute difference in the number of enrollments between the control and experiment groups. Therefore, it would be a good evaluation metric. However, since gross conversion can indicate a relative difference between the number of enrollments, I will rather not use this metric as an evaluation metric either.

1.1.4 Click-through-probability

Click-through-probability is the number of unique cookies to click on the "start free trial" button divided by the number of unique cookies to view the course overview page. Since neither the number of clicks nor the number of cookies should change across the control and the experiment groups, click-through-probability should not either. Therefore, I will use it as an invariant metric. Since this metrics is not expected to differ between the control and experiment groups, I will not use it as an evaluation metric.

1.1.5 Gross Conversion

Gross conversion is the number of users who have enrolled in the free trial divided by the number of unique cookies to click on the "start free trial" button. Since the number of enrollments can be affected by the experiment, gross conversion can be as well. Therefore, I will use it as an evaluation metric. Since this metric is expected to differ between the control and experiment groups, I will not use it as an invariant metric.

To launch the experiment I expect it to be significantly less in the experiment group, as that would mean that the experiment successfully reduced the number of students who did enroll in the free trial, without having enough time to finish the course successfully.

1.1.6 Retention

Retention is the number of users to remain enrolled past the 14-days boundary (and thus make at least one payment), divided by the number of users who have enrolled in the free trial. Since both the number of payments and the number of enrollments can be affected by the experiment, retention can be as well. Therefore, I will use it as an evaluation metric. Since this metric is expected to differ between the control and experiment groups, I will not use it as an invariant metric.

To launch the experiment, I expect it to be similar across the experiment and control groups. That would mean that the experiment did not affect the number of students who did enroll in the free trial, with having enough time to finish the course successfully.

1.1.7 Net Conversion

Net conversion is the number of users to remain enrolled past the 14-days boundary (and thus make at least one payment), divided by the number of unique cookies to click on the "start free trial" button. Since the number of payments can be affected by the experiment, net conversion can be as well. Therefore, I will use it as an evaluation metric. Since this metric is expected to differ between the control and experiment groups, I will not use it as an invariant metric.

To launch the experiment, I expect it to be similar across the experiment and control groups. That would mean that the experiment did not affect the number of students who did enroll in the free trial,

with having enough time to finish the course successfully.

1.2 Measuring Standard Deviation

1.2.1 Gross Conversion

The analytical standard deviation of gross conversion is 0.0202. As the unit of analysis (number of cookies to click) and the unit of diversion (number of cookies) are very close to each other, I do not expect the analytical and empirical variabilities to be different.

1.2.2 Retention

The analytical standard deviation of retention is 0.0549. As the unit of analysis (number of users who have enrolled int the free trial) and the unit of diversion (number of cookies) are far from each other, I expect the analytical and empirical variabilities to be different. Therefore, I would measure the empirical variability if I would have the time.

1.2.3 Net Conversion

The analytical standard deviation of net conversion is 0.0156. As the unit of analysis (number of cookies to click) and the unit of diversion (number of cookies) are very close to each other, I do not expect the analytical and empirical variabilities to be different.

1.3 Sizing

1.3.1 Number of Samples vs. Power

I will not use the Bonferroni correction during the analysis phase. To adequately power the experiment, I would need 4,741,212 pageviews total, across both groups.

Metric	Pageviews
Gross Conversion	646,450
Retention	4,741,212
Net Conversion	685,325

1.3.2 Exposure vs. Duration

During the experiment, no student going to be exposed anything beyond minimal risk. Neither personal, financial nor otherwise sensitive data will be collected, and no student can be hurt either. Assuming that there are no other experiments that we would like to run simultaneously, I would divert 100% of the traffic to the experiment.

Based on the number of required pageviews and the proportion of the diverted traffic, I would need 119 days to run the experiment. That is too much time, especially as during it, we could not run any other experiments simultaneously. Revising my previous decision, I decided not to use retention as an evaluation metric. Net conversion can provide the same sort of information (significant change in the number of payments).

Based on the decision to not use retention as an evaluation metric, the number of required pageviews is 685,325 and I would need 18 days to run the experiment.

Metric	Duration
Gross Conversion	17
Retention	119
Net Conversion	18

2 Experiment Analysis

2.1 Sanity Checks

2.1.1 Number of Cookies

A Z-test indicated that there was no statistically significant difference between the expected and observed probabilities of a cookie is being assigned to the control group, $p = 0.5006, 95\%CI[0.4988, 0.5012]$.

2.1.2 Number of Clicks

A Z-test indicated that there was no statistically significant difference between the expected and observed probabilities of a click is being assigned to the control group, $p = 0.5005, 95\%CI[0.4959, 0.5041]$.

2.1.3 Click-through-probability

A Z-test indicated that there was no statistically significant difference between the expected and observed click-through-probabilities of the experiment group, $p = 0.0822, 95\%CI[0.0812, 0.0830]$. The expected click-through-probability of the experiment group was calculated based on the observed click-through-probability of the control group.

2.2 Result Analysis

2.2.1 Effect Size Test

2.2.2 Gross Conversion

A Z-test indicated that the gross conversion was statistically and practically significantly lower in the experiment group than in the control group, $d_{min} = 0.01, 95\%CI[-0.0291, -0.0120]$.

2.2.3 Net Conversion

A Z-test indicated that the net conversion was neither statistically nor practically significantly different between the experiment and control groups, $d_{min} = 0.0075, 95\%CI[-0.0116, 0.0019]$.

2.2.4 Sign Test

2.2.5 Gross Conversion

A sign test indicated that gross conversion was statistically significantly lower in the experiment group than in the control group, $p = 0.0026, \alpha = 0.05$.

2.2.6 Net Conversion

A sign test indicated that net conversion was not statistically significantly different between the experiment and control groups, $p = 0.6776, \alpha = 0.05$.

2.2.7 Summary

The Bonferroni correction is designed to limit the risk of Type I errors in multiple comparisons. It can be used in cases when multiple independent tests performed simultaneously, and to make a decision, we expect at least on of them to be statistically significant. To launch the experiment, I expect both gross conversion and net conversion to be significant. Therefore, the Bonferroni correction would have been too conservative, and thus I did not use it.

2.3 Recommendation

The analysis of gross conversion indicates that the experiment successfully reduced the number of students who did enroll in the free trial, without having enough time to finish the course. On the other hand, the analysis of net conversion does not provide statistically significant evidence to conclude that the experiment did not reduce the number of students to continue past the free trial and eventually complete the course. Therefore, I would not launch the experiment, and I would try to improve net conversion.

3 Follow-Up Experiment

Most courses on Udacity consists of a set of lessons, each followed by a set of problems. In some cases, the number of videos per lesson, and the number problems per lesson can be large. Hence, the time between a student watches a lesson video, and being exposed to a problem related to it can be long. In a scenario like that, if a student cannot solve a problem it can be challenging and frustrating to find the exact lesson video, which contains the missing bit of information.

To address this issue, I would show a hint box to students who fail to solve a problem. The box would contain the list of lesson videos that are related to the problem, or could assist in solving it.

My hypothesis is that the change would help students to get over their difficulties easier, and therefore, it would reduce the number of students who cancel early in the course.

I would use user-ID as the unit of diversion, as the change would only be visible to users whose are enrolled in a course.

I would use user-ID as an invariant metric, and the number of payments divided by the number of user-IDs as an evaluation metric.

References

[1] Hervé Abdi: The Bonferonni and Šidák Corrections for Multiple Comparisons
`http://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf`