A/B Testing – Final Project

Data Analyst Nano Degree
Udacity

by Diego Valdivieso

Contents

Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

Invariant metrics	Evaluation metrics
Number of cookies	Gross conversion
Number of clicks	Retention
Click-through-probability	Net conversion

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

Invariant metrics

Number of cookies: *is the number of cookies to view the course overview page*. Therefore it is prior to the change, and consequently it should not be affected by it.

Number of clicks: *is number of unique cookies to click the "Start free trial" button.* It is prior to the change and should not be affected by it.

Click-through-probability: *is number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.* As it is a combination of previous defined invariant metrics, it should not be affected by the change.

Evaluation metrics

Gross conversion: *is the number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button.*
I would expect the checkout and enroll to decrease after the screener is triggered, and consequently it should be lower in the experiment group.

Retention*: is the number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout.*
I expect a greater proportion of users who remain enrolled after 14 days, consequently, this metric should be greater in the experiment group than in the control group.

Net conversion*: number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button.*
I did not expect a substantial difference in this metric between the two groups.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics.

Unique cookies to view page per day = 40000
Unique cookies to click "Start free trial" per day = 3200
Enrollments per day =660
Unique cookies sample = 5000

Gross conversion
Probability of enrolling given click = 0.20625

Given that the sample size is 5000 cookies, we need to calculate the related number of clicks, taking into account that in the provided data there 40000 cookies and 3200 clicks.

Number of clicks = (3200/40000) * 5000 = 400

SE_{GC} = sqrt((0.20625 * (1- 0.20625)) / 400) = **0.0202**

Retention
Probability of payment, given enroll= 0.53

Given that the sample size is 5000 cookies, we need to calculate the number of clicks, taking into account that in the provided data there are 40000 cookies and 660 clicks.

Number of clicks = (660/40000) * 5000 = 82.5

SE_{R} = sqrt((0.53 * (1- 0.53)) / 82.5) = **0.0549**

Net Convertion
Probability of payment, given click = 0.1093125

Given that the sample size is 5000 cookies, we need to calculate the number of clicks, taking into account that in the provided data there are 40000 cookies and 3200 clicks.

Number of clicks = (3200/40000) * 5000 = 400

SE_{NC} = sqrt((0.1093125 * (1- 0.1093125)) / 400) = **0.0156**

Gross conversion	Retention	Net conversion
0.0202	0.0549	0.0156

Based on the fact that the unit of analysis and the unit of diversion should be the same in order to be able to compare the analytical estimate and the empirical variability, it can be concluded that only for Gross convertion and Net conversion the empirically calculated standard deviation

is close to the analytic, due to the fact that for both of them the unit of diversion and unit of analysis are the same (cookies). On the other hand, for Retention the unit of analysis and unit of diversion is not the same, therefore, the empirically calculated standard deviation will not match the analytic estimation.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

I did not use Bonferroni correction because the evaluation metrics are too correlated, and the alpha would be too conservative.

α = 0.05
β = 0.2
$Dmin_{GC}$= 0.01
$Dmin_{R}$ = 0.01
$Dmin_{NC}$ = 0.0075

Based on the above parameters and on the given probabilities I obtained the following numbers of pageviews (from online calculator):

$IniSize_{GC}$ =25,835
$IniSize_{R}$ = 39,115
$IniSize_{NC}$ = 27,413

The previous numbers of pageviews must be adapted by considering the units of analysis and the number of groups (which are two):

$Size_{GC}$ = $IniSize_{Gc}$ * (40,000 / 3,200) * 2 = 645,875
$Size_{R}$ = $IniSize_{R}$ * (40,000 / 660) * 2 = 4,741,333
$Size_{NC}$ = $IniSize_{NC}$ * (40,000 / 3,200) * 2 = 685,325
:

Gross conversion	Retention	Net conversion
645875	4741212	685325

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment.
Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

I do not consider it to be highly risky, because the change (a free trial screener) should not significantly affect any critical business process. Therefore, I diverted all the traffic.

The estimated days to run the experiment for each evaluation metric is the estimated number of pageviews divided the number of pageviews per day:

Gross conversion	Retention	Net conversion
17	119	18

My main idea was to pick the higher number of days as the answer, but taking into account that days required for Retention is longer than a few weeks, I decided to discard that evaluation metric and I finally took the next longer period, which corresponds to Net conversion metrics.

Experiment Analysis

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.)
For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. ***Do not proceed to the rest of the analysis unless all sanity checks pass.***

α = 0.5
Z = 1.96

For the first two cases:

P(success) = 0.5
P(failure) = 1 – Probability of success = 0.5

Number of cookies
Total events control = Total control pageviews = 345543
Total events experiment = Total experiment pageviews = 344660

SE = sqrt((0.5 * 0.5) / (345543+ 344660)) = 0.006018
m (margin of error) = Z * SE = 0.01176

Confidence interval = [0.4988, 0.5011]

Observed value = Total experiment pageviews / Total page views = 344660 / 690203 = 0.5006

Number of clicks on "Start free trial"
Total events control = Total control clicks = 28378
Total events experiment = Total experiment clicks = 28325

SE = sqrt((0.5 * 0.5) / (28378+ 28325)) = 0.002099
m (margin of error) = Z * SE = 0.004115

Confidence interval: [0.4958, 0.5041]
Observed value: 0.5005

Click-through-probability on "Start free trial"
Unlike the previous two cases, the standard deviation and confidence interval are firstly calculated based only in the control group data, and them we check if the probability of the experimental group lies within that interval.

Click-through-probability Control = Number of Clicks / Page Views = 28378 / 345543 = 0.08212

Click-through-probability Experiment = 28325 / 344660 = 0.08218

SE = sqrt((0.082 * (1 - 0.082)) / 345543) = 0.000467
m (margin of error) = Z * SE = 0.00915

Confidence interval = [0.0812, 0.0831]
Observed value = 0.0822

Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)

Gross conversion
Control:
Total clicks: 17293
Total enrollments: 3785
Control Probability = Total enrollments / Total clicks = 0.21887

Experiment:
Total clicks: 17260
Total enrollments: 3423
Experiment Probability = Total enrollments / Total clicks = 0.19832

Difference = Experiment Probability – Control Probability = -0.0205
Pool Probability = (Tot Enroll Cont + Tot Enroll Exp) / (Tot Click Cont + Tot Click Exp) = 0.2086071

SE pool = sqrt((Ppool * (1-Ppool))/ (Tot Enroll Cont + Tot Enroll Exp)) = 0.0043717

m = SE * Z = 0.0085685

Confidence interval = [Difference – m, Difference + m] = [-0.0291, -0.0120]

Since the zero is not included in the confidence interval, it has statistical significance.
Since the min difference (+/- 0.01) is greater than the confidence interval, it has practical significance.

Net conversion

Control
Total clicks: 17293
Total payments: 2033
Control Probability = Total payments / Total clicks = 0.117562019

Experiment
Total clicks: 17260
Total payments: 1945
Experiment Probability = Total enrollments / Total clicks = 0.112688297

Difference = Experiment Probability – Control Probability = -0.004873

Pool Probability = (Tot Paymen Cont + Tot Payment Exp) / (Tot Click Cont + Tot Click Exp) = 0.1151

SE pool = sqrt(Ppool * (1-Ppool) * ((1/Tot Payment Cont)+(Tot Payment Exp))) = 0.0034

m = SE * Z = 0.0067

Confidence interval: [-0.0116, 0.0019]

Since the zero is included in the confidence interval, it doesn't have statistical significance.
Since the min difference (+/- 0.0075) is included in the confidence interval, it doesn't have practical significance.

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant.

Gross conversion

Given that the number of success is 19 (considering a success that the gross conversion in the experiment group is lower than in the control group) and the total number of cases is 23, the probability density of the binomial distribution returns a p-value of 0.0026, which is statistically significant.

Net conversion

Given that the number of success is 10 and the total number of cases is 23, the probability density of the binomial distribution returns a p-value of 0.6776, which is not statistically significant.

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

I did not use Bonferroni correction because evaluation metrics are high correlated, and the alpha would be too conservative.
I do not see a discrepancy between Effect Size and Sign tests, due to the fact that the p-value obtained using the Sign test do not contradict what was obtained in the Effect Size test.

Recommendation

Make a recommendation and briefly describe your reasoning.

My recommendation is not to apply the change, because even though Gross conversion results are statistically and practical significant, which means that many students who were not able to study more than five hours per week decided not to enroll, the Net conversion results are not statistical nor practical significant, and as the negative min difference is between the confidence interval, it is possible that the numbers of students who stay enrolled more than 14 days get reduced, which is not an option based on the stated objectives for the experiment.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

Since asking students if they were able to study at least five hours per week have not seemed to be an appropriate approach to keep the rate of students who remained enrolled past the 14-days boundary, in a further experiment I will try a different approach which will consist in encourage them by using positive stimulus.

The student activity (progress in lessons, number of times he visit the page per day, etc) will be tracked, and if it is detected that the students is not doing it as well as he expected (it could be used some machine learning algorithm which will be trained based on the behavior of previous students who did not pass the 14-days boundary), he will be motivated through emails and popups. He will receive support messages, and he will be offered the possibility to be put in contact with voluntary students who are performing well, in order to get tips and advises.

Hypothesis
The number of students who left the free trial will be lower, and therefore will increase the number of students who remained enrolled past the 14-days boundary.

Invariant metrics
Number of cookies, Number of clicks, Click-through-probability and Gross conversion.
The metrics should not be affected because the change that will be introduced in the experiment group occurs after enroll in the free trial.

Evaluation metrics
Net conversion: more students are expected to pass the 14-days boundary, consequently the result of this metric in the experimental group is expected to be higher than in the control group.

Unit of diversion
Cookies. Because it is the most suitable option for the chosen metric.