Project 7 -- Design an A/B Test

By Jiemin Wang

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**: Number of cookies, Number of clicks, Click-through-probability
- **Evaluation Metrics**: Gross conversion, Retention, 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.

Number of cookies: Number of unique cookies to view the course overview page. It’s evenly distributed between control and experiment groups and it’s independent from the experiment since the visits happens before users see the change. Therefore, it is a good invariant metric.

Number of clicks: Number of unique cookies to click the “Start free trial” button. It happens before the free trial screener is triggered and therefore it is independent from the experiment. It is a good invariant metric.

Click-through-probability: Number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. It is independent from the experiment since the “click” and “pageview” happen before the users see the experiment. It is a good invariant metric.

Gross conversion: 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. The metric value is affected by the experiment since the number of enrollments will be affected. Therefore, it is a good evaluation metric. There should be a decrease in the enrollments.

Retention: 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. The metric value is affected by the experiment since the user payment will be affected. It is a good evaluation metric. We expect positive change in financial outcome.

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. The metric value is affected by the experiment since the number of enrollments will be affected. It is a good evaluation metric. There should not be a significant decrease in the number of user-ids to remain enrolled past the free trial days.

Number of user-ids: Number of users who enroll in the free trial. It is not a good invariant since the number of enrollments will be affected by the experiment. It is usable as an evaluation metric since it tracks the number of enrollments which is affected by the experiment. However, it is not well normalized as Gross conversion so that we don't use it as an evaluation metric in our analysis.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

Evaluation metrics	**Standard deviation**
Gross conversion	0.0202
Retention	0.0549
Net conversion	0.0156

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

For Gross conversion and Net conversion, the denominator of these two metrics is cookie which is the same as the unit of diversion. Therefore, the analytic estimate would be comparable to the empirical variability. However, for Retention, the unit of analysis is user-id which is different from the unit of diversion. It is likely that the analytic estimate will be different from the empirical estimate.

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.)

No, I will not use the Bonferroni correction during the analysis phase.

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

As we can see, the number of pageviews we need to power the experiment is the largest number: 4,741,212.

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. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

With 40,000 pageviews per day, I would divert 70% of the traffic to the experiment. The total days needed to run the experiment is approximately 4741212 / (40000 * 0.7) = 170 days which is more than half a year. It is too long to execute an experiment in our case. Therefore, Retention is not an appropriate evaluation metric and we decide not to use it. Instead, we choose to Gross conversion and Net conversion as our evaluation metrics. We need to revisit the previous question and the number of pageviews to power the experiment will then be 685,325. Given the new pageviews, the days needed to run the experiment will be 685325 / (40000 * 0.7) = 25 days which is reasonable and executable.

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

The reason to divert 70% of traffic to the experiment is that the experiment is of low risk and does not collect sensitive data of users. Specifically, the experiment does not brings harm that exceeds the “minimal risk” to users, for example, physical, emotional or social risks. The experiment just adds a screener to users to show the minimum time commitment required by the course, which is of very low risk. In addition, the data collected in the experiment is not sensitive in that it does not involve financial, health or other personally sensitive data. The experiment does not affect current enrolled students and payments, but it is possible that the new feature might affect new enrollments and their payments, we then choose to divert 70% of the traffic to be safe to run the experiment.

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.)

Invariant metric	**Lower bound**	**Upper bound**	**Observed**	**Pass**
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks on “Start free trial”	04959	0.5041	0.5005	Yes
Click-through-probability on “Start free trial”	0.0812	0.0830	0.0822	Yes

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.***

All sanity checks pass.

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.)

Evaluation metric	**Lower bound**	**Upper bound**	**Statistical significance**	**Practical significance**
Gross conversion	-0.0291	-0.0120	Yes	Yes
Net conversion	-0.0116	0.0019	No	No

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. (These should be the answers from the "Sign Tests" quiz.)

Evaluation metric	**p-value**	**Statistical significance**
Gross conversion	0.0026	Yes
Net conversion	0.6776	No

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.

The Bonferroni correction is not used in our analysis since it is useful when several dependent or independent statistical tests are being performed simultaneously and it avoids spurious positives. However, in our case, we only launch the experiment when both of the two metrics (Gross conversion and Net conversion) show significant change.

Recommendation

Make a recommendation and briefly describe your reasoning.

We would not launch the experiment.

The analysis based on Gross conversion shows that the experiment reduces the number of unprepared students and the change is practically significant. However, results from Net conversion indicates that the change is statistically and practically insignificant. Besides, the confidence interval of Net conversion includes the negative of the practical significance boundary, which means the change can reduce the number of continued enrollments by an amount that matters to the business. Therefore, we decide not to launch 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.

Students of Udacity get frustrated and cancel the course early mostly due to lack of help at the very beginning since they don't understand how to proceed with the course and complete it successfully. Although Udacity offers different ways of help like forums and 1-on-1 appointments, for those who have no experience in programming or related fields, it is still challenging to start the course and to be comfortable with the learning. Therefore, if we can assign a "buddy" to each student, e.g., we provide an email address that the student can contact directly whenever they have problems or concerns about the course and the "buddy" will check the progress of the student periodically, give advice on learning based on the background of the student, it is possible that students will be more comfortable to keep learning.

Specifically, when students sign up with Udacity, we provide them with an email address of their assigned buddy which can be a member from the Udacity team. The buddy will then be a mentor/supervisor of the student and offer suggestions, answer questions and keep track of the student’s progress to encourage them to keep learning.

The null hypothesis is that by assigning a buddy to the new signup students, the Retention will not increase by a practically significant amount.

The invariant metric is the number of user-ids. Only signup students will be assigned a buddy and the number of signup students will not be affected by the experiment.

The evaluation metric is Retention, i.e., 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. We expect more students remain enrolled and an increase in revenue with the new added feature.

The unit of diversion is user-id since only signup users will be affected by the experiment.

If Retention is positive and practically significant from the experiment, we can then launch the new feature.

References

http://mathworld.wolfram.com/BonferroniCorrection.html
http://www.evanmiller.org/ab-testing/sample-size.html