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.)

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.

Answer

An invariant metric is a metric that should be comparable between test and control groups. These metrics can be used to check to see if the sample and control populations are the same for the purpose of sanity testing. The following could be used as invariant metrics for this experiment:

- Number of cookies: number of unique cookies to view the course overview page.
    - The user will view the course overview page before the free trial screener is triggered. Therefore the number of cookies should be split evenly between the experiment and control groups. If there were differences in this metric (e.g. because users in the test group clear their cookies more often), it would be important to be aware of this.
- Number of clicks: number of unique cookies to click the "Start free trial" button
    - Given that users would be randomly assigned to the experiment and control group, the number of unique cookies to select the “start free trial” button should be the comparable in both the test and control groups because this event occurs before the free trial screener is triggered.
- Click-through-probability: the number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.
    - Probability measures are useful when measuring the total impact of an event. In this experiment, both the sample and control groups are expected to have a similar proportion of unique clicks on the “Start free trial” button as a proportion of unique cookies to view the course overview page. The reason for this is that the action of clicking on the “Start free trial” button occurs before the free trial screener is triggered. This metric would be particularly useful if the size of the test and control groups were not equal.

An evaluation metric refers to the metric that enables the researcher to compare differences between the experiment and control group. The following could be used as evaluation metrics for this experiment:

- Gross conversion: number of user-ids to complete checkout and enrol in the free trial divided by number of unique cookies to click the "Start free trial" button.
    - It is anticipated that a lower number of user-ids in the test group will enrol in the free trial, compared to the control group, as a proportion of the number of unique

cookies to click the “Start free trial” button. If the hypothesis is correct, users who cannot commit the required time to the course will opt to access the course materials for free, rather than enrolling, and would therefore reduce the rate of enrolment in the course.

- 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 aim of this intervention is to set clearer expectations for students upfront, thus reducing the number of frustrated students who leave the free trial (without significantly reducing the number of students to continue past the free trial and eventually complete the course). The people who remain enrolled past the 14-day boundary, as a proportion of people who click the “Start free trial” button, should be comparable between the test and control groups. Even though no difference between the groups is anticipated, this is an evaluation metric because any difference between the trial and control groups (as a result of the experiment) is of interest. This metric should not be held constant at the start of the experiment, as would be the case for an invariant metric.

In order for this experiment to be launched, gross conversion and net conversion metrics need to be valid. If the gross conversion metric is valid, the rate of enrolment in the class should decline because students choose not to enrol if they are unable to commit sufficient time to the course. If the net conversion is valid, this would indicate that the intervention did not significantly reduce the number of students to complete the course. If the expected results are observed, this would indicate that by setting clearer expectations upfront, Udacity would reduce the number of frustrated students who leave the course without significantly reducing the number of students who continue beyond the 14-day boundary.

Metrics not in use:

- Number of user-ids: number of users who enrol in the free trial.
    - This metric would not be an appropriate invariant metric because users are only tracked by the user-id after they enrol in the free trial. The number of users who enrol in the free trial could be affected by whether the user sighted the free trial screener, therefore this would make a better evaluation metric. However, this metric focuses on the number of users (rather than a proportion), and results may be affected if there are different numbers of participants in the experiment and control groups. As a result, other evaluation metrics were selected to measure differences between the trial and control groups in this experiment.
- 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.
    - This metric was initially included as an evaluation metric; however too many page views were required for the experiment when this metric was in use. This metric was therefore excluded from the experiment. It would also have been possible to reduce the number of page views required by increasing dmin, alpha or beta; or

targeting a specific segment of the population. These changes were not made because other evaluation metrics were available to test the hypothesis.

Measuring Standard Deviation

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

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to 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.

Answer

Metric	Standard deviation
Gross conversion	0.0202
Net conversion	0.0156

The unit of diversion is chosen in the design phase of the experiment to provide a consistent monitoring of the user experience throughout the whole experiment: In this experiment, non-logged in traffic needs to be monitored to compute some metrics of interest. As a result, cookies were chosen as the unit of diversion.

An analytical estimate is likely to be accurate when the unit of analysis is the same as the unit of diversion. The unit of analysis for gross conversion and net conversion metrics is a cookie. Therefore, an analytical estimate can be used for these metrics.

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.)

Answer

The more metrics that are included in an experiment, the higher chance of a false positive result. The Bonferroni correction aims to address this issue by estimating alpha for an individual metric by dividing the desired alpha (overall) by the number of metrics in use. This method is very conservative and assumes independence of metrics.

In this experiment, the evaluation metrics are correlated and it is therefore likely that application of the Bonferroni correction would result in a false negative. As a result, the Bonferroni correction will not be used during the analysis phase, however there is a higher chance that a statistically significant result will occur by chance.

Page views total: 685,325

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.)

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

Answer

In order to run this experiment, a minimum of 685,325 page views are required. On average, 40,000 cookies view the page per day. If 100% of all traffic was included in the experiment, the experiment would need to run for approximately 18 days.

Given that this experiment is not very risky (just a single, small pop-up on electing to start the free trial) it would be appropriate to divert 100% of traffic to the experiment. This would also minimise the duration of the experiment, allowing Udacity to obtain results in a timely manner.

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.

Answer

The following sanity checks assume that the experiment was run for 18 days.

Number of cookies
CI: (0.4988, 0.5012)
P(allocated to control): 0.5006

0.5006 falls within the bounds of the confidence interval, indicating that the result is not statistically significant. This metric passes the sanity check.

Number of clicks
CI: (0.4959, 0.5041)
P(allocated to control): 0.5005

0.5005 falls within the bounds of the confidence interval, indicating that the result is not statistically significant. This metric passes the sanity check.

Click-through-probability
CI(control): (0.0812, 0.0830)
P(experiment): 0.0821

0.0821 falls within the bounds of the confidence interval, indicating that the result is not statistically significant. This metric passes the sanity check.

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.)

Answer

Gross conversion
CI: (-0.0291, -0.0120)
The confidence interval does not contain zero, indicating that the result is statistically significant. The practical significance value for this metric is 0.01. Given that the confidence interval range falls below dmin (-0.01), this result is also practically significant.

Net conversion
CI: (-0.0116, 0.0019)
The confidence interval contains zero and the practical significance value (dmin=0.0075), indicating that the result is not statistically or practically significant.

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.)

Answer

Gross conversion
Number of days with 'positive' outcomes: 4
Number of days in experiment: 23
p-value: 0.0026

The p-value of 0.0026 is less than the 0.05 level of significance. Therefore this result is statistically significant. This confirms the result observed in the hypothesis test.

Net conversion
Number of days with 'positive' outcomes: 10
Number of days in experiment: 23
p-value: 0.6776

The p-value of 0.6776 is greater than the 0.05 level of significance. Therefore this result is not statistically significant. This confirms the result observed in the hypothesis test.

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.

Answer

Answer to Bonferroni correction – see section 'Sizing'.
Explanation of sign test results – see section 'Sign Tests'.

Recommendation

Make a recommendation and briefly describe your reasoning.

Answer

Gross Conversion reduced by at least the practical significance boundary, indicating that the intervention significantly reduced the number of students who chose to enrol in the free trial. This is a positive result for Udacity because the cost of enrolment (both monetary and the cost of unsatisfied students) will decrease if this change is introduced.

There was no statistically significant change for net conversion, but the confidence interval included dmin. The result indicates that it is possible that the number of enrolments decreased by an amount that is significant to Udacity.

Given these results, it is recommended that Udacity does not launch the change. Ideally, Udacity should undertake another experiment with greater power to determine if a practically significant

result is obtained for net conversion. However, Udacity may not have the resources (time or money) to run another experiment investigating this change. If this is the case, it is recommended that Udacity abandon the experiment as it should not be launched unless both metrics are practically significant.

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.

Answer

The first step towards reducing the number of frustrated students who cancel early in the course would be to conduct an experiment to better understand why students are cancelling early. For example, this experiment could be conducted using a survey which appears when a student requests to cancel their enrolment. Based on the results of this qualitative feedback, Udacity may be in a better position to make changes to reduce the number of students who cancel enrolment prior to the 14-day boundary.

For example, the survey could indicate that a common reason for leaving the course prior to the 14-day boundary was that course videos take too long to load, making it difficult to view the content and complete the course. Udacity could undertake an experiment to see if lower resolution videos with faster load times would reduce the number of students who leave the course prior to the 14-day boundary.

The hypothesis for this experiment is that reducing load time for course videos will reduce the number of students who leave the free trial prior to the 14-day boundary. This hypothesis could be tested using an evaluation metric such as ‘number of user-ids to remain enrolled past the 14-day boundary divided by the number of user-ids to complete checkout’. This would enable a comparison of the proportion of enrolled students who remain enrolled after 14 days for the control and experiment group. Useful invariant metrics would include number of user-ids and number of cookies.

An appropriate unit of diversion would be the user-id, to ensure a consistent user experience across all platforms.