Problem Set 7: A/B Testing - Free Trial Screener

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.

Invariant Metrics

- Number of cookies (viewing the course overview page)
- Number of clicks (to "Start Free Trial")
- Click-through probability

The above metrics were chosen as they should be comparable between the experiment and control group. Viewing the course overview page and clicking on "Start Free Trial" would take place before the screener is shown. As such, these metrics are unlikely to vary between the experiment group that sees the screener and the control group that doesn't. The click through probability, being a ratio of the Number of clicks to the Number of cookies, will also likely be invariant between experiment and control groups.

Evaluation Metrics

- Gross Conversion
- Retention
- Net conversion

Gross conversion is expected to differ in the experiment group compared to the control group because seeing the screener would likely affect the decision of the experiment group in continuing or terminating the checkout.

Retention is expected to differ between the groups as the user-ids in the experiment group that have seen the screener and decided to complete checkout are more likely to have made an informed decision about completing checkout. One would expect then that the proportion of user-ids remaining enrolled past the 14-day boundary to be higher for the experiment group.

Likewise, net conversion is expected to differ in the experiment group. Users in the experiment group who have seen the screener could be more likely to have clearer expectations regarding the course, hence influencing their decision to continue beyond the 14-day boundary.

Unused Metrics

- Number of User IDs

The number of user ids enrolling in a trial *could* be used as an evaluation metric. However, as the unit of diversion is a cookie, we would not know how many unique users are in the experiment group and control group and whether the numbers are comparable. Therefore, comparing the absolute numbers of user-ids in each group enrolling in a trial would not be a very robust metric as to whether or not the experiment group was more or less likely to enrol.

To launch the experiment, I would look for comparability between the experiment and control groups for the invariant metrics. There would also need to be statistical and practical significance between the 2 groups for all evaluation metrics. I would want to see a decrease in the Gross Conversion, an increase in Retention and an increase in Net Conversion.

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 Metric	**Standard Deviation (Analytic Estimate)**	**Expected to be comparable to empirical estimate?**
Gross Conversion	0.0202	Yes
Retention	0.0549	No
Net Conversion	0.0156	Yes

**see excel spreadsheet "Calculations.xlsx" – SD_Eval_Metric for calculations*

The unit of analysis for Gross Conversion and Net Conversion is cookie (clicking on "Start Free Trial"), which is the same as the unit of diversion. As such, the analytic estimates and empirical estimates are expected to be comparable.

On the other hand, the unit of analysis for Retention is the user-id (i.e. users who enrol in free trial). The empirical variability could be expected to be different, therefore, it might be worth doing an empirical estimate if there is time.

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 your experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

Bonferroni correction will not be used during the analysis phase.

Total Number of pageviews needed with 3 evaluation metrics = 4,741,212

Total Number of pageviews needed with 2 evaluation metrics
(Gross Conversion & Net Conversion) = 685, 275

The number of pageviews required with all 3 evaluation metrics is too large and would result in a long experiment duration, which is not feasible.

The experiment design will be revised to include only 2 of the original 3 evaluation metrics i.e. Gross Conversion and Net Conversion will be used.

Evaluation Metric	Baseline	dmin	Min. Pageviews Per Group
Gross Conversion	0.20625	0.01	322938
Retention	0.53	0.01	2370606
Net Conversion	0.1093125	0.075	342638

**see excel spreadsheet "Calculations.xlsx" – Power for calculations. Alpha = 0.05; Beta = 0.2*

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?

100% of the traffic would be diverted to this experiment. With daily traffic at 40,000, the experiment would be run for 18 days.

The experiment has minimal risk as the change is only the inclusion of a screener for the experiment group which serves as a prompt to clarify expectations. There are no major changes to the site or course.

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.**

Evaluation Metric	**95% Confidence Interval**		**Actual Observed Value**	**Sanity Check Passed?**
	Lower Bound	Upper Bound		
Cookies (Pageviews)	0.4988	0.5012	0.5006	Yes
Clicks	0.4959	0.5041	0.5005	Yes
Click-through probability	-0.0013	0.0013	0.0001	Yes

**see excel spreadsheet "Calculations.xlsx" – Sanity Checks for calculations.*

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	**95% Confidence Interval**		**dmin**	**Statistically Significant?**	**Practically Significant?**
	Lower Bound	Upper Bound			
Gross Conversion	-0.0291	-0.0120	0.001	Yes	Yes
Net Conversion	-0.0116	0.0019	0.0075	No	No

**see excel spreadsheet "Calculations.xlsx" – Result Analysis for calculations.*

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	**Count of positive observations**	**p-value**	**Statistically Significant?**
Gross Conversion	4 (of 23)	0.0026	Yes
Net Conversion	10 (of 23)	0.6776	No

**see excel spreadsheet "Calculations.xlsx" – Sign Test for calculations.*

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 was not used. To launch the experiment, we already need changes measured by both metrics to be statistically and practically significant. Applying the Bonferroni correction in this case would therefore be overly conservative and could lead to a higher likelihood of false negatives.

The effect size hypothesis tests and the sign tests gave consistent results.

Recommendation

Make a recommendation and briefly describe your reasoning.

To launch the experiment, both the Gross Conversion metric and the Net Conversion metric need to be statistically and practically significant.

Gross Conversion appears to have been reduced in the experiment group, which is a desired outcome of the experiment i.e. a reduction in the number of enrolments of students in the experiment group (who might otherwise have enrolled and become unsatisfied if they did not see the screener). The change was statistically and practically significant, where the magnitude of values reflected in the 95% confidence interval was greater than dmin(=0.001).

On the other hand, the change in Net Conversion was not statistically significant. In terms of practical significance, dmin(=0.0075) lies above the (positive) upper bound of the 95% confidence interval, so there is no practical significance in terms of *improvement* in Net Conversion. In addition, the confidence interval contains negative values, with the lower bound at -0.116, which is of a magnitude greater than dmin. As such, the *deterioration* in Net Conversion (an undesired outcome) could be significant and the change could be detrimental to the company.

Based on the considerations above, the recommendation is not to launch the experiment.

Follow-Up Experiment: How to reduce early cancellations

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.

Experiment:

To test a change where after a student completes check-out, a reminder is shown that coaching support is available for courses.

Hypothesis:

The reminder would make it clear to students that customized help is available if they were stuck on courses, thus reducing the number of frustrated students who cancelled early during the 14-day trial period.

Unit of Diversion:

Since the reminder is shown only after check-out (which is tied to unique IDs), the unit of diversion chosen is the User ID.

Invariant Metric:

- Number of User IDs (completing checkout)

The number of users completing check-out in both the experiment and control groups should be comparable as the reminder would only be shown after checkout.

Evaluation Metrics:

- Retention (Number of User IDs remaining enrolled past the 14-day boundary divided by number of User IDs to complete checkout)
- Number of User IDs remaining enrolled past the 14-day boundary
- Number of pageviews by enrolled students

Retention can be used as an evaluation metric as we want to see the impact the change (coaching support reminder) has on the experiment group. The change is expected to encourage students in the experiment group to persevere in the course. One would expect then that the proportion of User IDs remaining enrolled past the 14-day boundary to be higher for the experiment group. The absolute number of User IDs remaining enrolled is likewise expected to be different for the experiment group. *(Note: as in the case of the free trial screener experiment, a longer duration may be required for the experiment in order to obtain sufficient power for this metric)*

The number of pageviews is also expected to differ for the experiment and control groups. The experiment group is hypothesized to be less likely to be frustrated due to the change, and therefore more engaged in the course during the 14-day period. This is expected to lead to more pageviews for the experiment group.