P7: Design and A/B test - 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.)

Invariant metrics

Invariant metrics are metrics that shouldn't change between the control group and experiment group. As we ask students "how much time do you have available to devote to the course?" after they click on the "Start free trial" button, the metrics **number of cookies, number of clicks and Click-through-probability** shouldn't change between the control and experiment groups because they are measured before clicking on "Start free trial" button. However the other metrics are measured after the question "how much time do you have available to devote to the course?" is shown, so we can expect Number of user-ids, Gross conversion, Retention and Net conversion to vary between control and experiment groups.
Later during our sanity check we will look only at the invariant metrics **number of cookies and number of clicks and click-through-probability**.

Evaluation metrics

In the experiment we want to analyze if the question "how much time do you have available to devote to the course?" has an impact on the number of people who decide to checkout and enroll in the free trial.

We can perform the following hypothesis testing on the **gross conversion** metric:

The hypothesis would be that the gross conversion for the experiment group is lower than the gross conversion for the control group because the question sets clearer expectations for students upfront in terms of weekly investment, thus reducing the number of student enrolling in the free trial by more than 1%.

We also want to use the **net conversion metric** as an evaluation metric to see if the number of students who pass the free trial is not reduced significantly as a result of asking this question.

The hypothesis would be that the net conversion metric remains the same between the control and experiment groups because the number of people who pass the free trial is the same between the control and experiment groups. Those who would have left during the free trial don't enroll in the free trial in the first place, because the expectations are set clearly. So for this

reason the lower bound of the confidence interval for the difference for the net conversion shouldn't be lower than 0.0075.

We could look at the retention metric; however, it's simply the difference between the net conversion and the gross conversion. For this reason we will only look at **the net conversion and the gross conversion metrics**.

Measuring Standard Deviation

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

Let's make an analytical estimate of the standard deviation for the gross conversion and net conversion metrics, given a sample size of 5000 cookies visiting the course overview page.

We want to measure if the practical significance boundary is realistic with the variability of these metrics.

For the metrics gross conversion and net conversion, the unit of analysis (denominator of the metric) and unit of diversion are the same: number of cookies. So we can use the analytical estimate instead of empirical estimate. If we would have used the retention as an evaluation metric, we would have to compute the empirical estimate. When the unit of diversion and unit of analysis are not the same, such as in the case of the retention, the empirical variability tends to be much higher than the analytical variability.

To compute the standard deviation of the gross conversion and net conversion metrics, we use the following table including the baseline values.

Here is the Standard Deviation for each metric:

Metric	Standard Deviation
Gross conversion	0.0202
Net conversion	0.0156

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.)

Here we won't use the Bonferroni correction.

To get the number of pageviews we need for the experiment, we will use the online calculator.

Metric	Number of pageviews
Gross conversion	685 325
Net conversion	645 875

In order to power the experiment appropriately for these two metrics we take the largest number of page views calculated for each metric. So we need to collect 685,325 pageviews.

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.)

We will run the experiment on all traffic. It shouldn't be a risky test as it's a small popup on electing the free trial.

If we run the experiment on all traffic, it will take 18 days to collect 685,325 pageviews. 18 days is already a long experiment and the experiment is run on a mix of weekends and weekdays. That confirms that we should run the experiment on all traffic.

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.)

Here are the results:

Metric	Lower bound	Upper bound	Observed
Number of cookies	0.4988	0.5012	0.5006
Number of clicks on "Start free trial"	0.4959	0.5041	0.5005
Click through probability	0.0812	0.0831	0.0822

For the **number of cookies** and **number of clicks on "Start free trial"**, **click through probability**, the observed fraction is included in the the confidence interval, so the invariant

metric, the number of cookies and Number of clicks on “Start free trial and click through probability, pass 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.)

Here are the confidence interval for each metric:

Metric	Lower bound	Upper bound
Gross conversation	-0.02912	-0.01198
Net conversation	-0.0116	0.00186

For the gross conversion metric, the confidence interval doesn’t include 0, so the test is statistically significant. The confidence interval [-0.02912,-0.01198] is inferior to the practical boundary (-0.01), so the test is practically significant.

For the net conversion metric, the confidence interval [-0.0116 , 0.00186] includes 0 so the test is not statistically significant. It also includes the practical boundary (-0.0075), so the test is not 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.)

Here is the p-value for each metric:

Metric	p-value
Gross conversation	0.0026
Net conversation	0.6776

For the gross conversion metric, the two-tail P-value returned by the calculator is equal to 0.0026, so the statistical test is significant. It means that these results are unlikely to happen by chance. The sign test agrees with the effect size test.

For the net conversion metric, the two-tail P-value returned by the calculator is equal to 0.6776, so the statistical test is not significant. It means that these results are likely to happen by chance. The sign test agrees with the effect size 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.

For this test, we won't use the Bonferroni correction because to launch the experiment, we need to make sure that both metrics, gross conversion and net conversion, match the hypothesis. It's different than the case where any of the metrics need to be significant in order to launch the experiment. For this case, we would have used the Bonferroni correction.

Recommendation

We can conclude that we have been able to decrease the gross conversion by setting clearer expectations for students upfront. However for the net conversion, the confidence interval [-0.0116 , 0.00186] does include the negative of the practical significance boundary (-0.0075). So there is a risk that the net conversion metric decreases by an amount that would matter for the business and negatively impact revenue. For this reason, we shouldn't launch this 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.

In this experiment, we will test whether or not an email showcasing a pertinent forum discussion from the course they are taking can improve their retention. The hypothesis would be that students who are struggling would find useful information that would help them better understand concepts and ideas in the course, or additional explanations on exercises. This should reduce their frustration and increase their motivation.

Our evaluation metrics would be the 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. To have a practical significant result we would need to have a positive practical boundary of 0.01.

Our unit of diversion would be user-id because we only care about students who already enrolled in the free trial.