

Final Project of Design an A/B Test Course

Udacity Data Analyst Nanodegree Project 7

Experiment Design

Metric Choice

Invariant metrics: Number of Cookies and Number of clicks *Evaluation metrics:* Gross Conversion and Net Conversion

- **Number of cookies:** I chose it as an invariant metric because it will be consistent throughout the experiment. It is calculated before the user sees the pop-up message, so the number will be similar in experiment and control groups.
- **Number of user-ids:** I didn't choose this as an invariant metric because this value is based on whether the user decides to enroll after the pop-up message. So the number of user-ids might be different between the experiment and control groups. It is not a good evaluation metric because this value does not have a denominator so it can't be normalized. If it could normalize like net conversion and retention, then it would have been a good evaluation metric.
- **Number of clicks:** Similar to the number of cookies, this value is independent from the experiment, so the number will be similar in experiment and control groups.
- **Click-through-probability:** A good invariant metric because it is independent from the experiment. However, I didn't choose this metric as an invariant metric for the test because this value seems to be redundant to the number of cookies and the number of clicks.
- **Gross Conversion:** This is a good evaluation metric because it is dependent on the experiment (This explains why it cannot be a good invariant metric). We expect that the value will be lower in the experiment group because some people who are not able to commit more than 5 hours per week won't enroll in the classes. In the control group, users won't see any pop-up message so they will enroll without any consideration of the number of hours they can commit per week.
- **Retention:** This is a good evaluation metric because it is dependent on the experiment (This explains why it cannot be a good invariant metric). We expect this value will be higher in the experiment group because majority of the group are those who can commit 5 hours per week and these people are more likely to make the first payments for the classes. However, I didn't choose this as an evaluation metric for the test because it is redundant to Gross conversion and Net conversion; this value can be calculated using Retention and Net Conversion.
- **Net Conversion:** This is a good evaluation metric because it is dependent on the experiment (This explains why it cannot be a good invariant metric). It would be great if this value increases after the test; however, due to the expected decrease in number of students enrolling in the free trial, it is possible that net conversion might decrease after the test. However, since the majority of the experiment group are those who can commit 5 hours per week and they are more likely to make the first payments for the course. So we want this value doesn't decrease after the test.

Measuring Standard Deviation

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

```
Standard Deviation of Gross Conversion: **0.0202**  
Standard Deviation of Net Conversion: **0.0156**
```

Gross Conversion and **Net Conversion** use the **number of cookies** as their denominators. This is a unit of diversion and it is same as a unit of analysis; thus, analytical estimate is comparable to the empirical variability.

Sizing

Number of Samples vs. Power

Will you use the Bonferroni correction in your analysis phase?

No. The metrics are not independent from the others so Bonferroni correction will be conservative to use for this test.

The number of pageviews I need to power the experiment:

I calculated the pageviews from the calculator from <http://www.evanmiller.org/ab-testing/sample-size.html> with a baseline conversion rate = 10.931% (Net Conversion rate) and minimum detectable effect = 0.75%. As a result, I got 27,413. However, this number is the size we need to power the experiment for net conversion rate. To find the pageviews we need, we need to take account that among 40,000 people viewing the page, only 3,200 people clicks the "Start free trial". Thus, $27,413 * 40,000 / 3,200 = 342,662.5$ pageviews. There are two groups, control and experiment, so we need to double this value. The total pageviews are 685,325.

Duration vs. Exposure

The fraction of traffic I would divert to this experiment is 0.8 (from 0 to 1). Given this fraction, I need $685,325 / 40,000 / .8 = 21.42 \rightarrow 22$ days 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?

Only 80% will be affected and the change due to the experiment is small, so it won't cause too much trouble in the overall business. The overall experiment is not very risky since there are no sensitive information we use in this experiment. Although users have to provide their credit card numbers when they enroll the course, these information are not part of the experiment.

Duration is 22 days that is less than a month which is what the client wants. So it is good amount of duration.

Experiment Analysis

Sanity Checks

```
Number of cookies: Lower bound = 0.4988, Upper bound = 0.5012, Observed = 0.5006:
Passed
SE = sqrt( p * (1-p) * (1/Number of Control Pageviews + 1/Number of Experiment
Pageviews)
= sqrt( 0.5 * 0.5 * (1/344660 + 1/345543) ) = 0.0006
Margin of error = 0.0006 * 1.96 = 0.0012
Lower bound = 0.5 - 0.0012 = 0.4988, Upper bound = 0.5 + 0.0012 = 0.5012
Observed = 345543 / (345543 + 344660) = 0.5006

Number of clicks: Lower bound = 0.4959, Upper bound = 0.5041, Observed = 0.5005:
Passed
SE = sqrt( p * (1-p) * (1/Number of Control Clicks + 1/Number of Experiment Clicks)
= sqrt( 0.5 * 0.5 * (1/28325 + 1/28378) ) = 0.0021
Margin of error = 0.0021 * 1.96 = 0.0041
Lower bound = 0.5 - 0.0041 = 0.4959, Upper bound = 0.5 + 0.0041 = 0.5041
Observed = 28378 / (28378 + 28325) = 0.5005
```

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: Lower bound = -0.0291, Upper bound = -0.0120
Statistical Significance: Yes, Practical Significance: Yes

Net Conversion: Lower bound = -0.0116, Upper bound = 0.0019
Statistical Significance: No, Practical Significance: No
```

Sign Tests

```
Gross Conversion: p-value = 0.0026, Statistical Significance = Yes
Net Conversion: p-value = 0.6776, Statistical Significance = 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.

I didn't use Bonferroni correction because the metrics are not independent; thus, Bonferroni correction is too conservative for the test.

We would have used Bonferroni correction if we expected not all of the metrics we chose showed significant changes after the test. The probability of making an error of falsely rejecting the hypothesis goes up when the number of metrics increases. So one false positive can control the decision. In this case, type I error would be very high and Bonferroni correction can be used to reduce this type of risk by reducing *alpha*. However, in this test, we expect all of the metrics behave in ways we expect. In this case a single false negative can control the decision; thus Type II error will be high, and we don't want to reduce the *alpha*. Thus, we don't want to use Bonferroni correction.

Both effect size tests and sign tests result in having statistical significance in **Gross Conversion** but no statistical significance in **Net Conversion**. This means that although the experiment affects the users' decision to enroll the online courses but it didn't affect much to the enrolled users to pay the courses.

Recommendation

Make a recommendation and briefly describe your reasoning.

I would recommend not to launch the experiment. The result of Gross Conversion showed statistically and practically significant changes after the pop-up message. This means that only those who are likely to commit more than 5 hours per week enroll the classes and they are more likely to make the first payments; this will reduce the costs of having those who are only taking free courses without making any payment. However, Net Conversion don't show statistically or practically significant changes after the test. This means that showing pop-up message to users before they enroll the classes don't affect the number of users who make their first payments after the free trial. Furthermore, the confidence interval of the net conversion includes the negative of the practical significance boundary which suggests the risk of hurting the Udacity's business.

Follow-Up Experiment: How to Reduce Early Cancellations: If you wanted to reduce the number of frustrated students who cancel early in the course, what experiment would you try?

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.

I believe that one of the main reasons why students cancel early in the course is due to the uncertainty of getting jobs after attaining the degree. Only if they have some guarantees of getting jobs, they would definitely continue enrolling the courses. Unfortunately the statistics of students of getting jobs after the course is difficult to measure because there are not enough amount of data and it takes too long to gather such data. However, Udacity can give students some inspirations by showing some clips of students who enjoy taking the Nanodegree. For example, if users complete checkout of the Data Analyst course, they will see some clips of students who enjoy taking Data Analyst course. Clips can show students who don't have any experience in computer science or students who are just trying to learn more about the certain area of studies. In current Data Analyst Nanodegree, there is no such clip specific to data analyst nanodegree when users

complete checkout. I believe that showing the clips can give some courage and inspirations to those who are considering to start the courses and it might increase a chance of reducing the number of students who cancel early in the course.

- **Null Hypothesis:** Showing a clip of enrolled students when users complete checkout will not increase the net conversion significantly.
- **Alternate Hypothesis:** Showing a clip of enrolled students when users complete checkout will increase the net conversion by a practically significant amount.

Clips will be assigned to users randomly by dividing users into two groups: control and experiment groups. In control group, everything stays the same and no clip will be shown in the course overview page. In experiment group, users will see clips when they visit the course overview page.

- **Unit of Diversion:** User-id is a good for a unit of diversion because users will be assigned with user-id once they decide to enroll the course after they watch the clips.
- **Invariant Metric:** I can use an invariant metric same as the unit of diversion. The number of user-ids will be a good invariant metric because the value won't be affected by the test because users see the clips after they complete checkout.
- **Evaluation Metric:** Retention will be a good evaluation metric because this value tells us the ratio of the number of users who make first payments to the number of users who complete checkout "start free trial" button. We want to see statistically and practically significant increases in experiment group.