Design an A/B Test

Experiment Design

Metric Choice

- `Number of cookies`: invariant metric

> As the *start free trial* page changes, the number of user that visit the website is unlikely to vary as that page has not been seen yet and should not affect users visiting the page.

- `Number of user-ids`: none

> Since the enrollment depends on the rendering of *start free trial* page, I would expect to see discrepancies in the control and experimental group. As such, it cannot be an invariance metric. On the other hand, it makes for a poor evaluation metric as it is redundant compared to the other metrics. The number of user-ids or enrolled users can fluctuate a lot with respect to the number of *start free trial* clicks on a given day, and thus not a good proxy for this experiment. Instead, the number of user-ids divided by the number of *start free trial* clicks, which is the gross conversion, is a better metric as it marginalizes variances in the empirical count of user-ids.

- `Number of clicks`: invariance metric

> This metric does not depend on how the *start free trial* page is rendered, much like the number of cookies.

- `Click through probability`: Invariance metric

> Similar to number of cookies and clicks, since the users have not seen the *start free trial* page before they decide the click on the button, the click through probability also is not dependent on the test being carried out.

- `Gross conversion`: evaluation metric

> The rendering of the *start free trial* page influences the number of users signing up for the free trial. That is, is the *5 or more hours per week* suggestion likely to affect conversion rates - this is one question we would like to understand through this A/B test. Therefore, this is a good evaluation metric.

- `Retention`: evaluation metric

  Likewise, it can be presumed that prompting users about the *5 or more hours per week* will have an effect on the ratio of users who make payments versus those who finish the free trial, and thus making this metric good for evaluation. However, this evaluation metric is discarded as it would have taken too long to gather data on this to support or deny the stated hypothesis of this A/B test. A complete analysis of this rationale is given below.

- `Net conversion`: evaluation metric

  Since this metric is the product of the previous two metrics, it can be simply derived to be an evaluation metric as well. The ratio of users who make payment over those who see the *start free trial* page is dependent on the rendering of that page and the *5 or more hours per week* suggestion. Hence, being a good overall goal of the A/B test and a good evaluation metric.

Measuring Standard Deviation

To evaluate whether the analytical estimates of standard deviation are accurate and matches the empirical standard deviation, the unit of analysis and unit of diversion are compared for each evaluation metric. A Bernoulli distribution is assumed here with probability `p` and population `N` where the standard deviation is given by `sqrt(p*(1-p)/N)`.

Gross conversion

```
p = 0.20625 (given)
N = 5000 * 0.08 = 400
std dev = sqrt(0.20625 * (1-0.20625) / 400) = 0.0202
```

Gross conversion is the 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. Since the unit of analysis and the unit of diversion is the cookie, the analytic variance is likely to match the empirical variance. Thus, the analytical estimate is used for this evaluation metric.

Net conversion

```
p = 0.1093125
N = 5000 * 0.08 = 400
std dev = sqrt(0.1093125 * (1-0.1093125) / 400) = 0.0156
```

Net conversion is the 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 analytical estimate is likely accurate as both the unit of analysis and unit of diversion are cookies as is with gross conversion.

Sizing

Number of Samples vs. Power

As the metrics used in this experiment are highly correlated, I decided against using the Bonfessoni correction as it will be too conservative in the figures calculated. This online calculator was used to generate the number of samples needed with `alpha = 5%` and `1-beta = 80%`.

Evaluation metric	Baseline conversion rate	d_min	Sample size needed	Number of pageviews needed
Gross conversion	20.625%	1%	25,835	645,875
Retention	53%	1%	39,115	4,741,212
Net conversion	10.93125%	0.75%	27,413	685,325

To achieve the necessary number of pageviews for the retention metric, it would take 117 days of complete site traffic, which is too long for an A/B test. Thus, only gross conversion and net conversion are used as evaluation metrics with the required number of pageviews being 685,325 and taking only 18 days at a 100% site traffic percentage.

Duration vs. Exposure

The fraction of Udacity's site traffic to be redirected for this experiment is purely driven by the risk tolerance of the experimenter. I feel that a 100% share towards this experiment gives a good and tolerable balance between the length of the experiment of 18 days (17.1 days rounded up) and the risk tolerance of exposing users to uncertain changes. That is, with all traffic redirected as test subjects, there will be 50-50 split between the control and experimental groups where each have 50% of the overall site traffic. If this experiment turns out to have a negative impact on the business, only 50% of all site visitors are at risk. Reducing this risk will require lengthening the duration of the experiment from 18 days, which is not desirable for an A/B test.

Experiment Analysis

Sanity Checks

Number of cookies

```
control group total = 345543
experiment group total = 344660
standard deviation = sqrt(0.5 * 0.5 / (345543 + 344660)) = 0.0006018
margin of error = 1.96 * 0.0006018 = 0.0011796
lower bound = 0.5 - 0.0011797 = 0.4988
upper bound = 0.5 + 0.0011797 = 0.5012
observed = 345543 / (345543 + 344660) = 0.5006
```

The observed value is within the bounds, and therefore this invariant metric passes the sanity check.

Number of clicks on "start free trial"

```
control group total = 28378
experiment group total = 28325
standard deviation = sqrt(0.5 * 0.5 / (28378 + 28325)) = 0.0021
margin of error = 1.96 * 0.0021 = 0.0041
lower bound = 0.5 - 0.0041 = 0.4959
upper bound = 0.5 + 0.0041 = 0.5041
observed = 28378 / (28378 + 28325) = 0.5005
```

The observed value is within the bounds, and therefore this invariant metric passes the sanity check.

Click-through-probability on "start free trial"

```
control value = 0.0821258
standard deviation = sqrt(0.0821258 * (1-0.0821258) / 344660) = 0.000468
margin of error = 1.96 * 0.000468 = 0.00092
lower bound = 0.0821258 - 0.00092 = 0.0812
upper bound = 0.0821258 + 0.00092 = 0.0830
observed = 0.0821824 (given)
```

The observed value (experiment value) is within the bounds, and therefore this invariant metric passes the sanity check.

Result Analysis

Effect Size Tests

Gross conversion

```
p = (3785 + 3423) / (17293 + 17260) = 0.2086
se = sqrt(0.2086 * (1-0.2086) * (1/17293 + 1/17260)) = 0.00437
d = 3423/17260 - 3785/17293 = -0.02055
lower bound = -0.02055 - 0.00437 = -0.0291
upper bound = -0.02055 + 0.00437 = -0.0120
```

This metric is statistically significant as the interval does not include zero, and is practically significant as it also does not include the practical significance boundary.

Net conversion

```
p = (2033 + 1945) / (17293 + 17260) = 0.1151
```

```
se = sqrt(0.1151 * (1-0.1151) * (1/17293 + 1/17260)) = 0.00343
d = 1945/17260 - 2033/17293 = -0.0048
lower bound =  -0.0048 - 0.00343 = -0.0116
upper bound =  -0.0048 + 0.00343 = 0.0019
```

This metric is not statistically significant as it included zero, and therefore not practically significant either.

Sign Tests

I used this online calculator to perform the sign tests.

	Number of days to see an improvement out of 23 total days	p-value	Statistically significant (< alpha)
Gross conversion	4	0.0026	Yes
Net conversion	10	0.6776	No

Summary

I decided not to use the Bonferroni correction as the metrics are already highly correlated and the correction would only make the resulting figures more conservative than needed. It might be useful to apply the Bonferroni correction if we decide to do post-test-segmentation on the results, such as browser-based analysis or demographic analysis. Based on the practically significance of the effective size and sign tests, gross conversion will decrease while net conversion will not be significantly impacted.

Recommendation

I recommend that we do not adopt the proposed changes of including the *5 or more hour* suggestion to the *start free trial* page as the A/B test shows that this will not have a practical significant effect on all evaluation metric, in particular, the net conversion. This change will not meet its business goal of increasing the number of paid users, and therefore this feature cannot be shipped.

Follow-Up Experiment

This experiment was focused on acquiring new users who are more qualified and would thus convert better. A potential follow-up experiment could be testing *enroll now with a discount*. Each nanodegree take a good part of a year on average and there is currently an offer of receiving 50% of tuition paid back if the program is completed within a year. The proposed discount will be applicable if the student finishes a course within a set time frame, say a month. This feature will allow users who are committed on the long-term to sign-up right away without entering a free trial and study from various nanodegree curriculum or standalone courses as they please and at a set time frame.

This option will appear after the *start free trial* button. This feature will allow users to skip the 14-day free trial portion and in exchange they get a tuition discount. This feature will be potentially compelling to users who are already determined to take the course and ready to jump in directly.

The hypothesis is that by providing this direct enrollment with a discount for completion in a set time frame feature, the number of signups will increase as students will understand the long-term nature of Udacity's course learning and will not have to make a rushed decision in just 14 days. They will have an expectation set on what an average completion time frame looks like and work towards that. Following are components for this analysis:

- *unit of diversion*: user_ids

  > This follow-up experiment can use user ids when they sign-up as the unit of diversion. This ensures that a signed-in user is not both in the control and experimental group.

- *invariant metric*: number of user_ids

  > As the course page changes with half the population only seeing the *enrol with a discount* option after selecting the *start free trial*, the number of user that visit the website is unlikely to vary as that page has not been seen yet and should not affect users visiting the page.

- *evaluation metric*: net conversion rate

  > This will provide data to test whether this new feature boosts enrollment. It is a good evaluation metric as it is directly dependent on the effect of the experiment.

If the evaluation metric is practically significant and better than the control group at the end of the experiment, we can launch the new feature. Further experimentation could be done on the discount value offered.

References

- What you really need to know about mathematics of A/B split testing
- Statistical Analysis and A/B Testing
- 3 Real-Life Examples of Incredibly Successful A/B Tests