A/B Testing: Final Project

Experiment Overview: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the course overview page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead. This screenshot shows what the experiment looks like.

The hypothesis was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

Experiment Design

Metric Choice

Metric	Invariant	Evaluation	Comment
cookies	x		Doesn't vary when assuming random allocation between control and experiment group. Is therefore a good invariant metric but a bad evaluation metric (no difference between control and experiment groups).
uid			We deal about uid that **start** the free trial. For users that do not enroll, their user-id is not tracked in the experiment.Iit measures the number of enrollments: we could have used as an evaluation metric because it is likely to change between control and experiment group. But we will use the gross conversion ("enrollment probability") instead because it normalizes between different group sizes. So this metric is neither an appropriate invariant nor evaluation metric.
clicks	x		Happens before the change is shown! So doesn't vary: is therefore a good invariant metric but a bad evaluation metric.
Click-through probability	x		Happens before the change is shown! So doesn't vary: is therefore a good invariant metric but a bad evaluation metric.
Gross conversion (d_{min}= 0.01)		x	Metric that will show us whether the screener warning did make more student choose to access the free materials without starting the free trial because they realized they don't have enough time. → **expected to decrease.** As this metric will change, it a good evaluation metric but a bad invariant metric.
Retention (d_{min}=0.01)			Not considered: see the decision taken in the duration vs exposure section.
Net conversion (d_{min}= 0.0075)		x	This is precisely what we **expect to see increase,** so we need to evaluate it. We all expect to see the difference between retention and gross

			conversion decrease: the change is about ensuring the students are aware of the time investment and making them take a thoughtful decision when choosing starting the free trial. As this metric will change, it a good evaluation metric but a bad invariant metric.

In order to launch the experiment we are looking for a **decrease in gross conversion and an increase in net conversion**. These changes will have to be both statistically (alpha=0.05) and practically (see dmin levels above) significant.

Measuring Standard Deviation

We are rather calculating the “standard error” (standard deviation of the sampling distribution). Assuming the assumptions of the Central Limit Theorem are valid, we will then approximate the distribution of our metrics by a normal distribution.
As we deal with probability distributions, we will then apply the following formula:

$$standard\ error\ = \sqrt{\frac{p(1-p)}{n}}$$

For the p values, refer to the baseline values table

- **Gross conversion:**

p=probability of enrolling, given click

$$n = \frac{cookies\ sample\ size}{Unique\ cookies\ to\ view\ course\ overview\ page\ per\ day} * Unique\ cookies\ to\ click\ "Start\ free\ trial"\ per\ day\ = \frac{5000}{4000} * 3200\ =\ 400$$

SE=0.02023

For this metric, the analytical standard deviation will likely match the empirical standard deviation as both the unit of diversion and unit of analysis are cookies.

- **retention**

Not used in the test (see sizing/duration vs exposure section)

- **Net conversion**

p=Probability of payment, given click
n=400 (see gross conversion=
SE=0.01560
Same situation as for the gross conversion: the unit of diversion and unit of analysis are the same (cookie). So the analytical and empirical standard deviations are likely to match!

Sizing

Number of Samples vs. Power

I won't use the Bonferroni correction because the chosen metrics are quite correlated, especially gross and net conversion. Moreover we've seen this correction tend to be too conservative.

Using the sample size calculator already mentioned during the lesson, we get the following sample sizes, for alpha=0.05 and beta=0.2 (also using the dmin levels given):

Metric	Sample size for one group	Total size (X2)	Corresponding number of cookies
Gross conversion	25,835	51,670	51,670/0.08=645,875
Retention	39,115	78,230	78,230/(660/40000)=4,741,212
Net conversion	27,413	54,826	54,826/0.08=685,325

An alternative method to calculate the required sizes could have been to use the script seen in the lesson and take the standard deviations previously calculated as parameters:

```
## Strategy: For a bunch of Ns, compute the z_star by achieving desired alpha, then
## compute what beta would be for that N using the acquired z_star.
## Pick the smallest N at which beta crosses the desired value

# Inputs:
#   The desired alpha for a two-tailed test
# Returns: The z-critical value
get_z_star = function(alpha) {
    return(-qnorm(alpha / 2))
}

# Inputs:
#   z-star: The z-critical value
#   s: The standard error of the metric at N=1
#   d_min: The practical significance level
#   N: The sample size of each group of the experiment
# Returns: The beta value of the two-tailed test
get_beta = function(z_star, s, d_min, N) {
    SE = s /  sqrt(N)
    return(pnorm(z_star * SE, mean=d_min, sd=SE))
}

# Inputs:
#   s: The standard error of the metric with N=1 in each group
#   d_min: The practical significance level
#   Ns: The sample sizes to try
#   alpha: The desired alpha level of the test
#   beta: The desired beta level of the test
# Returns: The smallest N out of the given Ns that will achieve the desired
#         beta. There should be at least N samples in each group of the experiment.
#         If none of the given Ns will work, returns -1. N is the number of
#         samples in each group.

required_size = function(s, d_min, Ns=1:20000, alpha=0.05, beta=0.2) {
```

```
    for (N in Ns) {
        if (get_beta(get_z_star(alpha), s, d_min, N) <= beta) {
            return(N)
        }
    }

    return(-1)
}
```

Keeping only the maximum of these 3 sizes, the needed number of cookies for our experiment is: 4,741,212.

Duration vs. Exposure

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?
The 2 main risks we can possibly face for an experiment are possible physical damage and sensitive data processing. They are not present in this experiment. So we can without risks divert 100% of the traffic.
Assuming that we would divert 100% of the traffic, we would need 4,741,212/40000=119 days! This is an excessively long duration.

I see two options to deal with this issue:

1. Dismiss the retention metric

In this case, I would choose not to consider the retention metric anymore. In this case the net conversion metric would determine the required size. And we would need:
685,325/40000 = 17,1 rounded to 18 days, which sounds like a much more reasonable duration for our experiment.

2. Increase the practical significance level

If one still wants to keep the retention as metric, one could choose to increase the minimum Detectable Effect or practical significance level in order to reduce the needed sample size:
For instance increasing the level from 1% to 3% would give a required size of 526,909 and a corresponding duration of 22 days.

Let's keep the first option: number of samples= 685,325, experiment duration = 18 days, experiment exposure = 100%.

Experiment Analysis

Sanity Checks

	pageviews	clicks	CTR
p	0.5000	0.5000	CTR_cont =

			0.0821
SE = $\sqrt{\frac{p(1-p)}{n_{cont}+n_{exp}}}$	0.0006	0.0021	0.0005
ME = $z_{crit}(\alpha = 0.05,\ 2tails) * SE$ ME = 1.96 * SE	0.0012	0.0041	0.0009
Low = p - ME	0.4988	0.4959	0.0812
High = p + ME	0.5012	0.5041	0.0830
p_obs	$\frac{x_{cont}}{n_{cont}+n_{exp}}$ = 0.5006	$\frac{x_{cont}}{n_{cont}+n_{exp}}$ = 0.5005	CTR_exp = 0.0822

All sanity checks pass.

Result Analysis

Effect Size Tests

	Enrollment (Gross conversion)	Payments (net conversion)
p_pool = $\frac{x_{cont}+x_{exp}}{n_{cont}+n_{exp}}$	0.2086	0.1151
SE_pool = $\sqrt{p_{pool}(1-p_{pool})(\frac{1}{n_{cont}}+\frac{1}{n_{exp}})}$	0.0044	0.0034
ME	0.0086	0.0067
d_obs	-0.0206	-0.0049
low	-0.0291	-0.0116
up	-0.0120	0.0019
dmin	0.01	0.075
statistically significant (0 not contained in CI)	yes	no

practically significant (+- dmin not in CI)	yes	no (-dmin in CI)

Sign Tests

Here are the results obtained using this calculator, after counting the days of negative and positive change for the gross and the net conversion respectively:

metric	Gross conversion	Net conversion
p_sign_test	0.0026	0.6776
Statistically significant (alpha = 0.05)	Yes	No

Summary

An experiment where visitors of the Udacity website are diverted by cookie between a control and an experiment group is conducted. The experiment consists in after having clicked on “Start trial”, showing the experiment group a screen asking them how much time they can dedicate to learning per week.

We selected 3 invariants metrics:

- Number of clicks (on the “start free trial button”)
- Number of cookies
- Click-through-rate

These metrics were used for validation and perform sanity checks which were successfully passed.

After dismissing the retention metric because of a corresponding required experiment time excessively long, we selected the 2 following metrics:

- Gross conversion: enrollment/clicks
- Net conversion: payments/clicks

These metrics were used for evaluation:

- Statistical significance:
    - Null hypothesis: no difference regarding the considered evaluation metric between control and experiment group.
    - Alpha = 0.05
- Practical significance: for each metric a significance level is defined

We will launch the tested feature if we reject the null for both our evaluation metrics and if the practical level of significance is reached or exceeded.

So because (and also for the reasons explained at the beginning of the Sizing part) this statistical significance is required for both metrics, I don't use the Bonferroni correction

I didn't use the Bonferroni correction because the chosen metrics are quite correlated, especially gross and net conversion. Moreover we've seen this correction tend to be too conservative. Finally we are a small number of comparisons (2), as per this paper, it is not recommended to apply the Bonferonni correction in such a case:
While the Bonferonni does help with the fact that more tests increase chances of making type I error (rejecting a true null), it also increases chances of making a type II error (false negative, keeping a false null)

Results showed a statistical and practical significance for the gross conversion but neither a statistical nor a practical significance for the net conversion.

Recommendation

The results summed up in the previous tells us that although showing the additional free trial screener did reduce the gross conversion or proportion of students enrolling, as expected, it was not combined with an increase in the net conversion, saying a proportion of students paying after the free trial, unlike hoped.
For this reason I wouldn't recommend launching: the tested features doesn't lead to the desired effects.
One should try other experiments.

Follow-Up Experiment

A reason for cancelling can be the excessive time investment, this is what was tested in the experiment.
And we saw that asking only about the time investment didn't have a significant effect on the cancellation, one could try asking about other things e.g. pre-requisite skills. Indeed one might cancel because one misses essential basic programming to follow along a course.
Without running another experiment at the same time, the parameters and conditions of this experiment would be the same as the one we just evaluated:

- Unit of diversion: cookie. Like in the initial experiment we will use the gross and net conversion as evaluation metrics, so cookie will be an appropriate unit of diversion. And if we need to approximate the analytical standard deviation with the empirical one, this approximation is likely to be correct because our unit of diversion and unit of analysis will be identical.
- Evaluation metrics: gross and net conversion
- Null hypothesis for each of the 2 test: no change
- No particular risks: we can divert 100% of the traffic

If at the end of this experiment, our control and experiment groups for both metrics are statistically and practically significant, we would launch this feature.