Experiment Design

Among the prospective students of Udacity's Nanodegrees, there are some unrealistic assumptions about the necessary time to complete one of these programs. Some students begin the free trial, but become frustrated because they don't have enough time at hand to progress successfully. Many of these students soon drop out, but they are still taking a lot of attention of the coaches during their presence.

Udacity is trying to improve the experience and support for those students thoroughly dedicated to working on a ND, and at the same time create a more realistic image of the requirements for being successful in a ND.

To this end, they are experimenting with including a modal pop-up box after students click the *Start free trial* button, that informs them about time requirements and asks them whether they have at least 5 hours available each week.

With this change, Udacity is hoping to reduce the amount of students that start the free trial without the prospect of actually working on their ND intensively. Thereby increasing the Student-to-Coach ratio. This should be visible in a significant reduction of the **Gross conversion** metric.

At the same time Udacity does not want to lose on actual customers. In theory, those not signing up for the free trial because of the experimental change, would be students who would anyways drop out of the program before the first payment rolls in. Therefore, we're expecting the **Net conversion** rate to remain unaffected.

Metric Choice

Invariant Metrics

Chosen: Number of cookies:

That is, number of unique cookies to view the course overview page. (dmin=3000)

A cookie can be taken as a proxy for unique users. The course overview page gets accessed by both groups, control and experiment, therefore the **number of cookies** to view the course overview page should not be significantly different between them. This makes it a good invariant metric to perform a sanity check on.

Chosen: Number of clicks:

That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is triggered). (dmin=240)

I chose also **number of clicks** as an invariant metric. Since the pop-up modal appears only *after* the click, the number of clicks should be similar between control and experiment.

If it's not, then something went wrong in the experimental setup or in my expectation of how the experiment will proceed - therefore it is a good invariant metric to double-check what is

and has been going on.

Click-through-probability:

That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. (dmin=0.01)

Click-through-probability, calculated for unique cookies clicking the *Start free trial* button divided by unique cookies viewing the course overview page, could also be considered a useful invariant metric.

However, I chose number of clicks over click-through-probability because I found that metric to be more straightforward and easier to handle.

Evaluation Metrics

Number of user-ids:

That is, number of users who enroll in the free trial. (dmin=50)

This is also a possible evaluation metric. I chose, however, to stick with the two metrics mentioned further down, because they represent *ratios* which are more reliable.

Chosen: Gross conversion:

That is, 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. (dmin= 0.01)

This tests the effect that the dialogue box has on the users. In the experiment, I would expect it to be lower for the experiment group than for the control group - but we never know, so we'll still be doing it two-tailed.

Gross conversion is therefore a metric that, according to our tested hypothesis, should show a significant change between the different groups - in case there is a change to be observed.

Retention:

That is, 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. (dmin=0.01)

Retention measures a similar aspect as *Net conversion*. However, the unit of analysis for Retention is *user-ids*, which is different from the unit of diversion we'll be using in this experiment (unique cookies). With these two being different, it makes it more difficult to do a sound analysis and additional aspects and transformations need to be considered. Because I think that the aspect I am interested in, is measured well enough with *Net conversion*, I therefore limit the experiment to this.

Chosen: Net conversion:

That is, 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. (dmin= 0.0075)

Net conversion is maybe one of the more frequent business metrics, since it tracks users actually becoming paying customers. Since this is also interesting for Udacity, I'll make this one of my Evaluation Metrics.

Here we are however not looking for a change, but instead would be already fine if it will stay the same. The point of the proposed change, is to warn and keep away people with too little time to be successful in a paid program, from sticking around in the free trial and eventually getting frustrated and dropping out anyways.

I will use this metric, because with it we can measure whether the proposed change puts off not only the right kind of people, but *also* puts off those that should be staying around.

Basically, we hope that the experimental change will not lead to a significant decrease in paying customers (people remaining over the 14-days-boundary).

Measuring Standard Deviation

Gross conversion: **0.0202**

Net conversion: **0.0156**

I expect the analytical estimate of both metrics' standard deviations to be comparable to the empirical variability. This is, because the unit of diversion in our experiment are unique cookies - which is also the unit of analysis for both evaluation metrics. Since unit of diversion and unit of analysis coincide, I can expect the estimates to converge.

However, since both metrics are ratios, it would be maybe of help to also do an empirical estimate to come to a more secure conclusion in the Effect Size tests.

Sizing

Number of Samples vs. Power

I decided not to use the Bonferroni correction. There is no need to peek because I know how long I have to run the experiment. The justifications follow here, when I calculated the required sample size and duration in advance. Therefore I can let the experiment run to the end with the calculated and pre-defined sample size, and then go to check the results in the end.

The number of page-views I will need in order to power my experiment is: **685 275**

I calculated this using the online sample size calculator (screenshots attached: `gross_conversion.png` and `net_conversion.png`). The results were the necessary clicks per group. I chose the higher of the two numbers, because this will allow me to have enough power for the other metric as well. Then I had to adapt the number to account for pageviews (/ 0.08) and for both groups (* 2).

Duration vs. Exposure

Since this is a change that is currently intended to be launched globally, and there are no specific cultural population of users that are to be tested out, I would stick with launching the change globally instead of only on a specific *target population*. Also, the change is not very browser intensive and should run fine on all browsers and devices, so I believe limiting the population according to this aspect is also unnecessary.

According to the baseline values provided, I can expect ca. 40 000 unique cookies to visit the Udacity site per day. (Note that the rate at which unique cookies visiting the page also click on the *Start free trial* button is already included in our sample size calculation further above, therfore I can here deal with pageviews.)

I would not divert the full amount of students to this experiment. We do not know what the effect really is, and it dabbles with the sign-ups. Since the students paying for support are Udacity's income, it's at the heart of the business. Therefore I originally wanted to rather stay cautious and divert only 40% of the traffic to the experiment.

However, this would give me 16 000 pageviews diverted to the experiment per day (8 000 to the control group and 8 000 to the experiment group), which would require the experiment to run for 43 days.

In the spirit of fast iteration that I know Udacity operates on, I consider this a too long timespan. The actual risk is probably much lower than my cautious approach suggests: Apart from potentially losing some new sign-ups, there wouldn't be any longer-lasting changes that Udacity would have to deal with, and it will also not affect the already exisiting user base. I would therefore increase the number to 80% of the traffic, effectively getting results within half the time.

This means I would run the experiment on **80%** of Udacity's traffic, for **22 days** (21,4 rounded to the next full day). I am comfortable with this timespan, because it will also include a variety of different days: weekdays, weekends, and maybe even a short holiday. This will make the result more stable and realistic.

Additionally, I would take a look into the sign-in data that Udacity (probably) collected, in order to be able to take a decision about *when* to launch the experiment. I can imagine that sign-ups for Udacity NDs might be higher at the start of university quarters, or otherwise maybe during holidays - and they might be lower during e.g. exam periods. However, these are just wild guesses, and I'd first take a look at the actual data in order to find a timeframe that might result in a relative stable flow of new sign-ups.

Experiment Analysis

Sanity Checks

Number of cookies:

Cookies should have been randomly assigned between the control and the experiment group. If the experiment is properly set up, we can therefore compute the Confidence Interval around 0.5.

Number of cookies passes the Sanity Check (code used for calculation is attached: `A_B_calculations.html` and `.py`):

```
Sanity check passed
p^(control) = 0.5006 and p^(experiment) = 0.4994
Confidence Interval: 0.4988 | 0.5 | 0.5012
```

- expected value: **0.5006**
- CI lower bound: **0.4988**
- CI upper bound: **0.5012**

This means that the cookies were indeed randomly assigned between the control and the experiment group.

Number of clicks:

I expect there to be no inherent heterogenity regarding the willingness of the users in the different groups to click (aka their "loose finger" coefficient), therefore this should also pass. And it does:

```
Sanity check passed
p^(control) = 0.5005 and p^(experiment) = 0.4995
Confidence Interval: 0.4959 | 0.5 | 0.5041
```

- expected value: **0.5005**
- CI lower bound: **0.4959**
- CI upper bound: **0.5041**

These results show, that eventual differences in the number of clicks between the control and the experiment group are within the bounds of a 95% confidence interval. This means, that from 100 experiments, only 5 would show results indicating otherwise.

Result Analysis

Effect Size Tests

See the code I used for calculation in the attachment `A_B_calculations.html` and `.py` for an executable version.

Gross Conversion

practical significance boundary: dmin = 0.01

- Lower bound of the CI: **-0.029123**
- Upper bound of the CI: **-0.011986**

Gross conversion is **both statistically and practically significant**.

The confidence interval does not include the practical significance boundary (mind the minus-sign!), which means that we can be confident at a 95% confidence level, that the true change is large enough to be worth launching.

Net Conversion

practical significance boundary: dmin = 0.0075

- Lower bound of the CI: **-0.011605**
- Upper bound of the CI: **0.001857**

Net conversion is **neither statistically nor practically significant**.

Since the confidence interval includes 0, the results are not statistically significant.

Sign Tests

Analyzing the day-by-day data, I've come up with the two-tailed-p-values for the two evaluation metrics I chose. The two-tailed-p_value is the probability of observing a significant result due to chance. In order for the metric to pass the sign test, the value needs to be below the chosen alpha, which is 0.05 - if it is higher the metric fails the test.

The p-value for Gross conversion is **0.0026**, which is clearly **significant**.

The p-value for Net conversion is **0.6776**, which is **not statistically significant** at our chosen *alpha* of 0.05.

(calculations made here in the provided google spreadsheet on *Sheet 3*)

Summary

I decided **not to use the Bonferroni correction** for two reasons:

1. I am expecting for both metrics a *specific outcome regarding their significance* (gross conversion should show a significant change, net conversion should show *no* significant change, or a significant change towards an increase in net conversion). So I am expecting very specific results from both my metrics, and none of the two "doesn't matter", making them related in an "AND" condition.
2. Using the Bonferroni correction additionally to this would increase the conservativism of my results even more, and will make it unlikely to detect an actual effect.

The effect size hypothesis and the sign tests agree with each other on the fact that *gross conversion* was significantly reduced in the experiment group as compared to the control group. At the same time there was no significant change recorded between the two groups

for the *net conversion* metric.

Recommendation

From the gathered data and the statistical results of our analysis, I would recommend Udacity to **launch the change** regarding the onboarding for the free trial.

The results showed a significant decrease in the number of users that would start the free trial after clicking the *Start free trial* button. At the same time, they failed to show a significant change in the amount of users who eventually signed up for a Nanodegree.

This means that the proposed change might indeed serve as a way for people to consider whether they have enough time at hand to be successful in a ND. Since those that haven't are likely to end up leaving the program within the free trial period, the launch might prevent some frustrations by directing the prospective students to the free courses where they can study at a much slower pace.

This change could have the potential to free up more resources of the Udacity coaches to interact and support those students that currently are able to dedictate the appropriate amount of time and will therefore also be able to progress in their ND. The test's result indicate, that these students would not be scared away by the additional dialogue box.

However, there are always other factors to consider, so I still believe it would be worthwile to launch the change conservatively, maybe not affecting the whole traffic, but only half of it. That way it would be possible to continue collecting data about eventual mistakes or unexpected outcomes that weren't caught during the experiment. Some aspects, such as time period in which the experiment was launched, or also diverse cultural backgrounds of Udacity users could have unexpected effects. Some potential of the latter is discussed in my proposal for a follow-up experiment.

Follow-Up Experiment

Description and Reasoning

Primarily, Udacity aims to be an international learning site that offers its courses in English. Therefore there was no stringet reason to limit the previous experiment to a specific population according to some demographic aspects.

However, assessing the global population might have the effect of diluting the outcome and rendering it less significant than it might be for a certain target population. After all, even though students take the courses in English, they still come from different cultural and language backgrounds. It might therefore well be true that the added dialogue box could have different effects on people from different cultural backgrounds. Selecting different target populations according to countries, world regions, or language families could result in an interesting set of experiments.

This would obviously require more testing, and therefore I feel like proposing it as a **follow-up experiment**.

To limit the experimental suite in scope, in order to see whether this might be worhtwile, I'd

propose one initial experiment with US customers as target population. The largest part of Udacity's customers are US-based (I think). Therefore it could be interesting to take US customers as the target population and find out how the change affects them. If possible, we could also run the experiment on the global population and see whether the significance of the result would change, and in which direction.

If this would yield interesting results, the process could be repeated for other culturally diverging target populations.

With Udacity being an international education provider, it could be worthwile and interesting, to tailor certain aspects of the service specifically to audiences from different countries. This might make the content altogether more accessible and might make people from outside the US more likely to engage with the service, ultimately increasing Udacity's revenue.

I would run these experiments in a very similar way to this experiment, using unique cookies as the unit of diversion and measuring the gross conversion and the net conversion.

Hypothesis

My **hypothesis** for this experiment (or these groups of experiments) is, that in the experiment group, the number of users who clicked the button and then signed up for the free trial (gross conversion) will significantly decrease, while the number of users who clicked the button and stayed sign in over the payment boundary (net conversion) will not decrease.

H0: p_exp(gross_conversion) = p_cont(gross_conversion) AND p_exp(net_conversion) = p_cont(net_conversion)

Ha: p_exp(gross_conversion) < p_cont(gross_conversion) AND p_exp(net_conversion) >= p_cont(net_conversion)

Unit of diversion

As mentioned, for the **unit of diversion** I'd suggest to use an *anonymous ID*, represented by the cookie. I am aware that users might be accessing Udacity's service on different devices, or maybe from different browsers. It is also possible that some users empty their browser cache, effectively deleting the cookie and having a new one assigned on their next login. However, all units of diversion are associated with some kind of potential problems, and this one seems to be the best fitting for the experiment. This is, because I want to test *user visible changes* and have the experiment commence before users are signed up for the service.

User IDs are not feasible, because a part of what I am trying to test out happens *before* users actually sign up. So for this they wouldn't be trackable through user IDs.

Using the *event* as the unit of diversion could result in a very fragmented user experience (so that users get assigned to different groups on each login), which is something that should be avoided. However, since the event of actually signing up for the free trial happens only once - and further tracking is interesting, on one hand it sounds like this could work, on the other

hand (the longer tracking) it feels useless to take the *event* as the unit of diversion. Because I want to track users over a period of time, *event*s are not a useful unit of diversion.

Other considerations

Ethical accountability

I'd be using cookie-based diversion, which means that the users will not be identifyable. There are no email addresses or other such types of information collected during the experiment that would be associated with the cookie and eventually render it non-anonymous. Therefore, from a **ethical** point of view, the experiment should be fine.

Learning effect

Accounting for a **learning effect** would be unnecessary in this experiment, because users only sign up once. So everyone exposed to the dialogue box either decides to sign up or not. Some might return and be confronted with the info and subsequent choice again. However, it's not that every time they interact with the site, they will be exposed to the dialogue box and it wouldn't each time have an influence on how they would continue. Therefore any *learning effect* should be miniscule and can be omitted.

Resources

A/B Tests

- http://www.evanmiller.org/how-not-to-run-an-ab-test.html
- http://www.exp-platform.com/Documents/2014%20experimentersRulesOfThumb.pdf

Tools

- http://www.evanmiller.org/ab-testing/sample-size.html
- http://graphpad.com/quickcalcs/binomial1.cfm

Forum

- https://discussions.udacity.com/t/calculating-page-views/27070
- https://discussions.udacity.com/t/help-final-project-calculate-number-of-pageviews/27038/
- Bonferroni correction: https://discussions.udacity.com/t/when-to-use-bonferroni-correction/37713/5
- https://discussions.udacity.com/t/why-dont-we-use-empirical-ways-to-calculate-se/157958
- https://discussions.udacity.com/t/final-project-hypothesis/37176