Experiment Design

Metric Choice

Invariant Metrics

Invariant Metrics are sanity check means, to make sure that the distribution of both samples are similar. When comparing two samples, we want them to be comparable from the first place. These variables must not be affected by the experiment, the must remain constant and hence the name invariant.

In our experiment, we have three given variables that seem to be immune to the experiment:

1- Number of cookies
   - If we are correctly distributing users over both groups, we must end with two roughly equal number.
   - It does not fit as an evaluation metric, since its value is predetermined by our testing framework. Also, the experiment is not about measuring if the users open the page or not, thus it would be irrelevant as a metric of evaluation.

2- Number of clicks (Of the start free trial button)
   - As described, this is the initial click over the button, before seeing the experimental part about asking the number of hours. Therefore, it must be similar in both groups if our assignment was truly random.
   - Since this value is captured before the experiment, it does not capture any phenomena we are interested about, and therefore is not an evaluation metric.

3- Click-through-probability (Of the start free trial button)
   - It is the probability that the button was clicked, it is simply derived from the previous two metrics. If both samples were selected from the same population, this rate must be similar in both groups.
   - Since this one is composed of the previous two, its rationale for not being an evaluation metric is the same as the previous two.

Any significant difference in the previous metrics means that our experiment was not correctly designed.

For the number of user-ids, it cannot be used as an invariant metric since should the screener be effective, it must change between the two groups. User-ids also do not fit as an evaluation metric, because we lack a solid metric to compare against. I think if we wanted to use it as an evaluation metric, we should have chosen to track the number of user-ids who clicked the button so we can get the rate of conversion of users-ids between both groups. I would have personally chose to track right from the beginning as I think it is more robust than cookies (Cookies can get cleared, or user might sign in using different devices\browsers), unless of course the user can run through the experiment before signing in.

Evaluation Metrics

All the remaining given metrics will be used for evaluation. These metrics are measured post the experiment part, and they are actually a measure of how effective was the idea to confirm how much time can the student consecrate to their studies

1- Gross conversion

2- Net conversion

Here, I have dropped retention from getting used as an evaluation metric because its unit of analysis is not the unit of diversion. Their variances will be different, and we risk underestimating the results analytically, which is an unnecessary hassle since we have two direct measurements already. Also, using it would have required a much longer experiment time to acquire enough samples.

Gross Conversion
- If the screener is effective, I would expect this number to change in the experiment group, therefore it is not suitable as an invariant metric.
- To launch the experiment, I want to see this number go down, i.e. the screener did filter users who will not study hard enough from using the paid service for two weeks.

Net Conversion
- If the screener is effective, AND the hypothesis is correct, then this number should not decrease significantly.
- To launch the experiment, this metric should not significantly change. Remaining as is or increasing is a green light to launch the experiment.

Note: The relationship between both metrics are a logical AND, not an OR. In other words, to launch the experiment:
- The Gross Conversion must significantly decrease
- The Net Conversion must not significantly decrease

Measuring Standard Deviation

From the baseline spreadsheet, we have:
- 40,000 cookies visiting the course overview page.
- 3200 cookies click the "start free trial", or 8%
- 660 enrollments, or 1.65% of the total (20.625% of those who clicked the button)

From the quiz, it says given a sample size of 5000 cookies visiting the course overview page.

➔ Our sample is about $\frac{5000}{40000}$ = 12.5% of the data.

So, assuming that the sampling was correct (i.e. representative of the population), we would expect that it will contain:
- ~ 8% will click "Start free trial", 0.08 x 5000 = 400 unique cookies clicking the button.

So now that we know what numbers N to expect, and also given the probabilities from the baseline sheet, we can use the binomial distribution to estimate the standard deviation for our evaluation metrics:

$$\sigma = \sqrt{\frac{p * (1-p)}{N}}$$

$\sigma_{gross} = \sqrt{\frac{0.20625*(1-0.20625)}{400}}$ = 0.0202

$$\sigma_{net} = \sqrt{\frac{0.1093125*(1-0.1093125)}{400}} = 0.0156$$

I would expect that the values hold between the analytical and the empirical values, since we use the same unit of diversion. If this was not the case, then the analytical results would likely underestimate the empirical values. But just as a side note, the actual data have a daily sample that is much higher, so it is safe to expect that the standard deviations (Both analytical and empirical) will be different than the numbers provided above (They should be smaller, since we have a higher count).

Sizing

Number of Samples vs. Power

As specified, we are going to use:

- $\alpha = 0.05$
- $\beta = 0.2$
- Gross Conversion
  - $d_{min} = 0.01$
  - Baseline Conversion rate = 0.20625
- Net Conversion
  - $d_{min} = 0.0075$
  - Baseline Conversion rate = 0.1093125

Using the online calculator mentioned in the course (www.evanmiller.org/ab-testing/sample-size.html), we get:

- Gross Conversion: 25,835 per group
- Net Conversion: 27,413 per group

Since we are going to use both metrics, it only makes sense to use the larger number. So we need at least 54,826 total samples to be able to trust our results. By sample here, we mean people who do actually click the button, which is 8% of the total traffic. So all in all, we need at least 685,325 total, assuming that our 8% estimate was correct.

There is no need of using a Bonferroni correction here, the two metrics are required together to take a decision, and so there is not a big risk of launching due to chance alone.

Duration vs. Exposure

I will start first with reasoning about the exposure. Generally speaking, the risks in this experiment is minimal, there are no sensitive information or any potential psychological\physical harm I can think of, and so we can go with a very high number.

On the other hand, the potential risk is more from our side, the business:

- The new feature may have a bug which might hurt the company's image, so still I would not go with a 100% exposure.
- An important risk to think about too: what if the experiment is a terrible idea from the first place, and it will drive away the students? i.e. it have a negative significant impact. Maybe

the wording is too negative, or maybe some students would get offended from Udacity underestimating their capabilities, who knows.

- The other side benefit from not using all the traffic will be slightly prolonging the time of the experiment, so any special patterns coming from holidays and the likes will be leveled out.
  ➔ Arbitrarily, I selected 80% of the traffic, which means that less than half of the users will be exposed to the new feature.

Now that we have set the exposure, we can estimate the duration based over our estimated "Unique cookies to view page per day", which is equal 40,000.

➔ 80% of 40,000 = 32,000 views per day
➔ With the total of 685,325 views needed, we will need 22 days to get the needed samples (We will actually get around 704,000 views)

Experiment Analysis

Sanity Checks

Our Invariant Metrics:

- Number of cookies
- Number of clicks on "Start free trial"
- Click-through-probability on "Start free trial"

We assume that members of both groups had an equal chance of being assigned to either groups, so our p = 0.5
We are asked to use α = 0.05

➔ Z = 1.96 for a two tailed
➔ Computations will be based over the control group, as observed in the experimental group

Metric	Total Number	Standard Deviation	Standard Error	Lower CI bound	Upper CI bound	Actual Value	Status
Number of cookies	690203	0.000601	0.0011779 6	0.4988	0.50117	0.50063967	PASS
Number of clicks on "Start free trial"	56703	0.002099	0.0041140 4	0.4958	0.50411	0.50046735	PASS
Click-through-probability on "Start free trial"	(In Control group) p = 0.08212581	0.000467	0.0009153 2	0.0812104	0.083041	(In experimental group) p = 0.082182	PASS

All invariant metrics seem to be equal in both groups, we can proceed with our analysis.

Result Analysis

Effect Size Tests

First, we can only work over the first 23 entries, since these are the ones that contain data for the conversion rates.

Gross Conversion

Let us compute the pooled mean p so that we can compute the pooled standard error

$$\hat{p} = \frac{X_{control}+X_{experiment}}{N_{control}+N_{experiment}} = \frac{3785+3423}{17293+1726} = 0.2086$$

Normality Check: is $\hat{p} \times N > 5$?

➔ Yes, then we can assume normal distribution (That was expected anyway given the large number of samples)

$$SE_{pooled} = \sqrt{(1 - 0.2086) * 0.2086 * (\frac{1}{17293} + \frac{1}{17260})} = 0.00437162$$

Margin of error $m = SE_{pooled} \times Z_{95\%} = 0.00437162 \times 1.96 = 0.00856837$

Let our hypothesis equation: $\hat{p}_{experiment} - \hat{p}_{control}$ be $\hat{d}_{gross\_conv}$. To disprove our null hypothesis, $\hat{d} \neq 0$

$$\hat{d} = \frac{3423}{17260} - \frac{3785}{17293} = -0.020554875$$

So, our confidence interval is $\hat{d}_{gross\_conv} \pm$ margin of error = [-0.02912, -0.01198]

Since our population's mean is zero, and zero is not part of our confidence interval, we can reject our null hypothesis. That may be good, since the question was to filter out students who are not likely to dedicate enough time to their studies.
From the business point of view, the results are practically significant as well.

Net Conversion

$$\hat{p} = \frac{2033+1945}{17293+1726} = 0.11512749$$

$$SE_{pooled} = \sqrt{(1 - 0.11512749) * 0.11512749 * (\frac{1}{17293} + \frac{1}{17260})} = 0.00343413$$

Margin of error $m = SE_{pooled} \times Z_{95\%} = 0.00343413 \times 1.96 = 0.00673$

Let our hypothesis equation: $\hat{p}_{experiment} - \hat{p}_{control}$ be $\hat{d}_{net\_conv}$. To disprove our null hypothesis, $\hat{d} \neq 0$

$$\hat{d} = \frac{1945}{17260} - \frac{2033}{17293} = -0.004873723$$

So, our confidence interval is $\hat{d}_{net\_conv} \pm$ margin of error = [-0.0116, 0.001856]

The population mean is within our confidence interval, so statistically we accept that the net conversion rate was unaffected. However, the negative boundary of the practical significance is well within the confidence interval. That means that there is a good chance that we might have decreased the actual enrollments.

Sign Tests

Online Calculator Used: https://graphpad.com/quickcalcs/binomial1.cfm
α = 0.05
Size = 23
Probability = 0.5

Gross Conversion
Number of success: 4
p = 0.0026
p < α, therefore statistically significant

Net Conversion
Number of success: 10
p = 0.6776
p > α, therefore statistically insignificant

The sign tests match the effect size tests

Summary
Statistically and practically, the free trial screener reduced the number of students who proceeded with the enrollment, which is the goal.
The net conversion statistically did not change. However, the net conversion showed a practical risk that it might have decreased. It would be too risky to launch the improvement with such results.

Bonferroni correction is not suitable to be used in this test, in fact it will be too conservative. Bonferroni's purpose is to address the risk of type I errors, which is not a concern since the relationship between our metrics is an AND. In our case, the higher risk is type II errors and this is not what Bonferroni correction addresses.

Recommendation
My recommendation is not to launch the screener. It reduced the number of students who started the free trial, which is part of the desired outcome, but there is a risk that it negatively affected the number of students who remained after 14 days. It seems that it discouraged some of the students who would have potentially continued past the trial period from trying the service from the first place.

Follow-Up Experiment
To follow-up, we need an experiment that allows us to filter out students who would potentially drop without affecting the total enrollment. We now know that a screener is effective, but the current one potentially filters out the wrong crowd. Maybe a slight improvement can give it the necessary boost required.

Since Udacity's payment scheme is rather monthly, not a lump sum for the degree (Like the newer AI and self-driving car programs), then the perceived length of the time it would take the student to complete the degree can be important. I would change the screener to be instead of the

total time devoted for the studies per week, I would ask at the screener if the student possesses the prerequisites for the degree. For example, in the data analyst nanodegree, Python knowledge is certainly helpful. Basic high school mathematics are important as well (It is alright if the student needs a small refresher, but previous exposure is important).

So my new experiment would be a screener that is a small quiz: a programming question, a basic question about relevant mathematics and perhaps a question about the student's motivation to take the course (If the student's goal is to just expand their knowledge, I assume they will be less motivated for the commitment than a student who is unemployed and is seriously looking to acquire new skills that increase their hiring potentials).

The new experiment:

- Will continue using cookies as a unit of diversion.
- Will continue using the same invariant metric as a sanity check. After the previous experience, we can have more confidence that they work.
- I would use the same two group settings for an overall analysis, but I will want to segment the experiment group into two subgroups to understand for a post experience analysis to understand more the effect of the experiment:
    - Segment of students who performed below the average on the quiz
    - Segment of students who scored above the average on the quiz

  ➔ This segmentation part is not really a part of the main experiment, but I am just stating that I would ask the engineering team to keep track of the student's results in the quiz.
- For the evaluation metrics, I would keep the gross and net conversions.
    - The gross conversion will be our indicator about the effectiveness of the screener as a way to reduce the number of enrollments.
    - The net conversion will be our indicator that the segment of students who do actually enroll are not discouraged by the screener, only those who drop-out are.

H_0: Number of students who enroll in the free trial is not affected by a screening quiz on the prerequisites overall.

H_A: The amount of students who enroll in the free trial will decrease after passing the screening quiz. The filtered out students do not contain a substantial amount of those who were to continue past the 14 days, thus the net conversion rate will not be negatively affected (No change or positive change are both good).