1. <https://github.com/shubhamlal11/Udacity-AB-Testing-Final-Project>
2. <https://www.kaggle.com/code/mariusmesserschmied/udacity-a-b-testing-final-course-project>
3. <https://zacks.one/udacity-a-b-testing-by-google/>

In this experiment, if it is successful in setting clearer expectations for students upfront and reducing the number of frustrated students who leave the free trial, the Gross Conversion is expected to decrease.

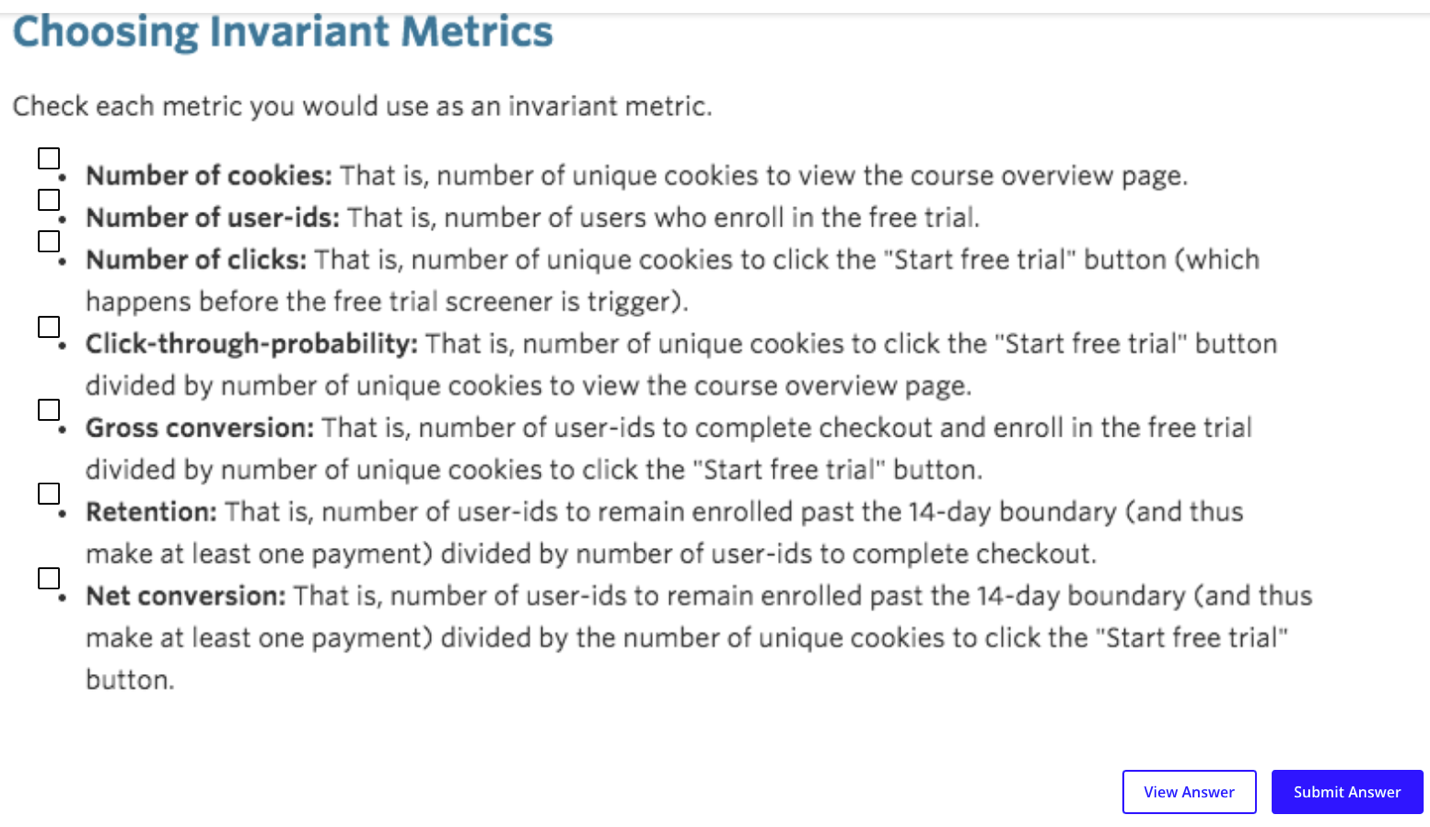
Here's the reasoning:

Gross Conversion is defined as the number of user-ids that complete checkout and enroll in the free trial divided by the number of unique cookies that click the "Start free trial" button.

The change in the experiment involves asking students how much time they have available to devote to the course before proceeding with the free trial enrollment process. If students indicate fewer than 5 hours per week, they are given the option to access the course materials for free instead of enrolling in the free trial.

By introducing this step, the experiment aims to deter students who are unlikely to commit enough time from enrolling in the free trial. This should lead to a decrease in the number of user-ids that complete checkout and enroll in the free trial.

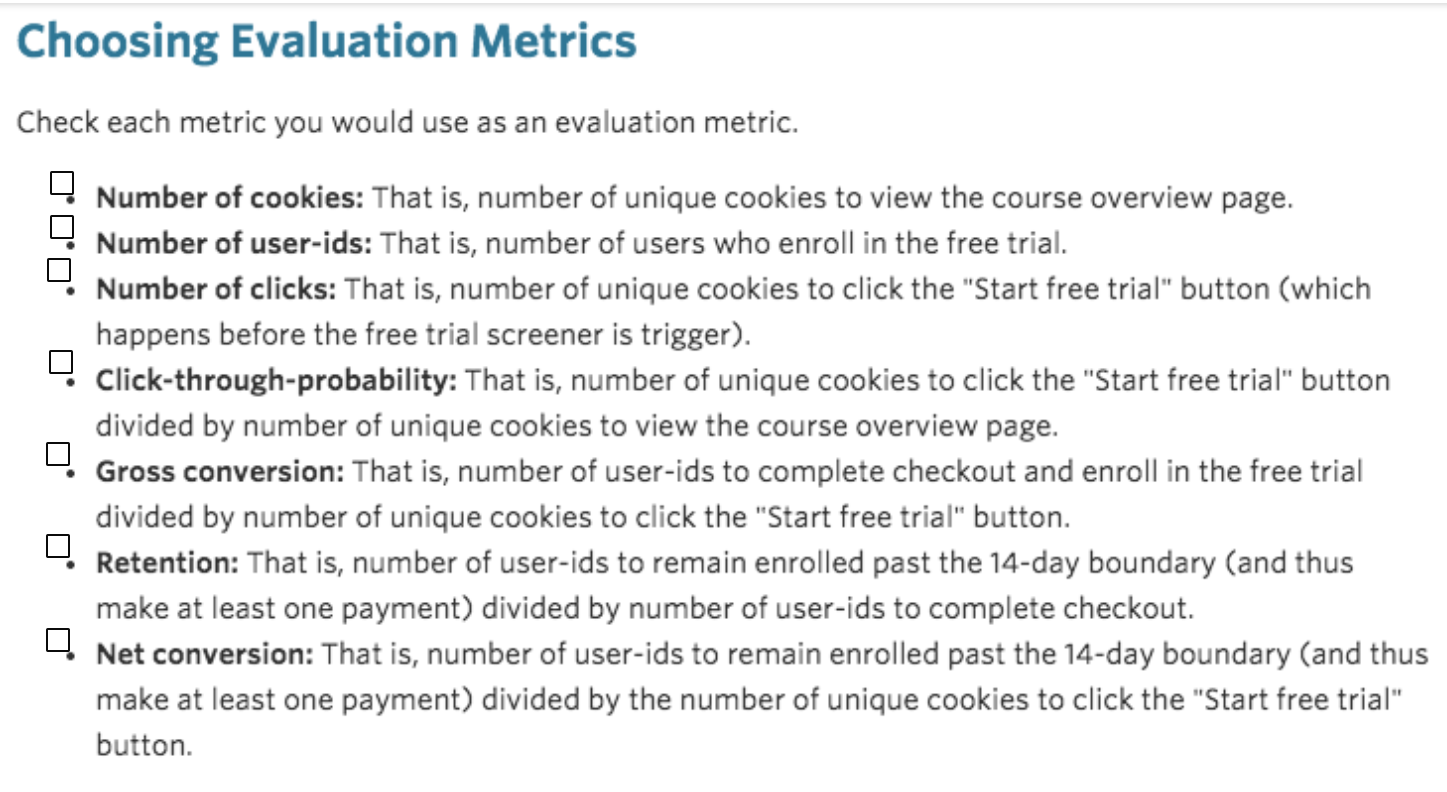
The hypothesis behind this change is that it will reduce the frustration of students who later realize they don't have enough time, ultimately leading to a decrease in the Gross Conversion.



Number of cookies

number of clicks

click-through-probability.

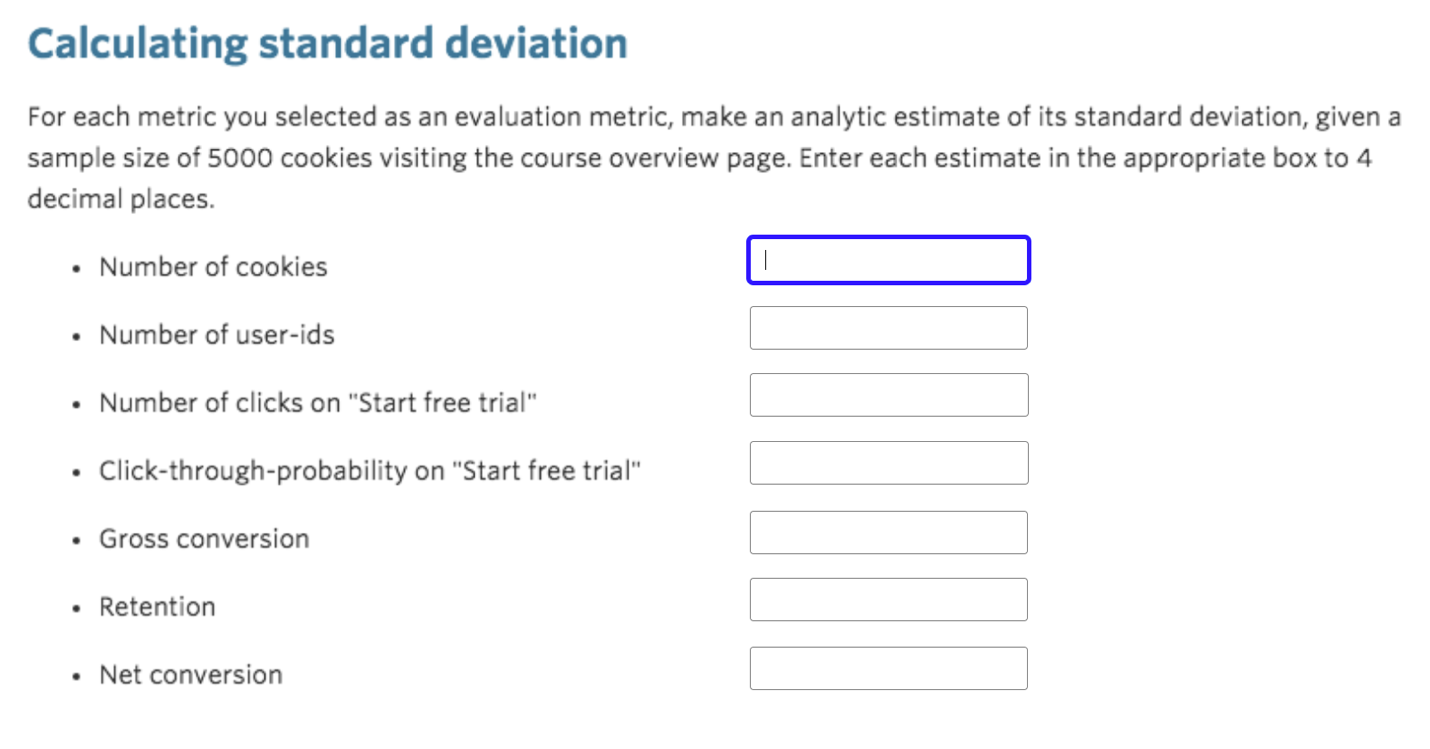


Number of user-ids: if the experiment works, then decrease. However, we would also expect the number of user-ids (i.e. the number of users who enroll in the free trial; dmin=-50) to decrease. However, the metric is not normalized and would not provide any information we are not already capturing with gross conversion (as the number of clicks will be controlled for). Thus, we will not use it as an evaluation metric.

Gross conversion: if works, then decreases. The denominator should be the same cross control and treatment but the numerator should decrease, which lead the gross conversion to decreases.

Retention: if works, increases. The denominator should decreases as the enrollment should decreases, but the numerator should increases, which lead the retention to increases

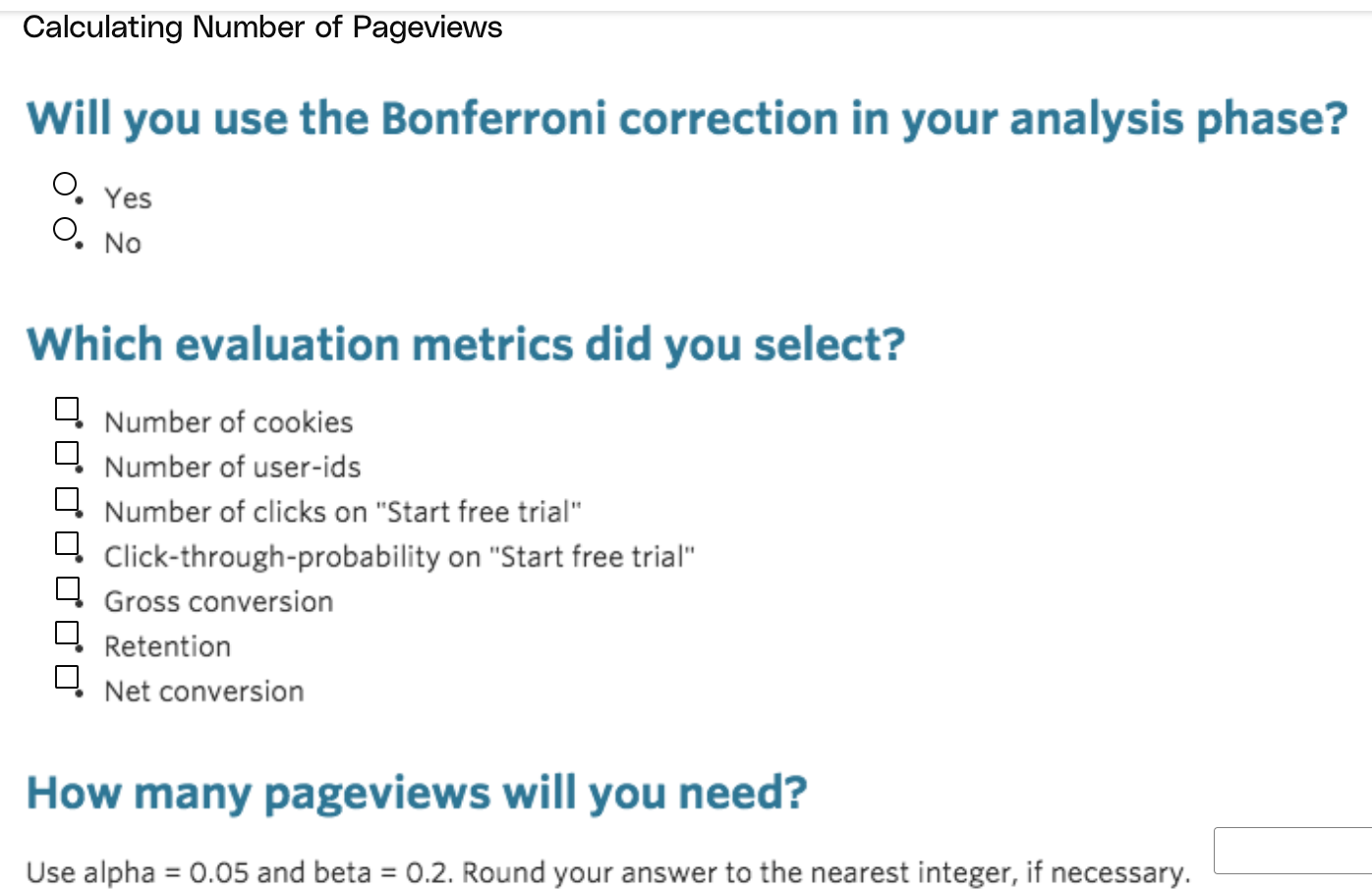
Net conversion: if works, then increases; the denominator should be the same cross the control and treatment but the numerator should increases in the experiment



|  |  |
| --- | --- |
| Unique cookies to view course overview page per day: | 40000 |
| Unique cookies to click "Start free trial" per day: | 3200 |
| Enrollments per day: | 660 |
| Click-through-probability on "Start free trial": | 0.08 |
| Probability of enrolling, given click: | 0.20625 |
| Probability of payment, given enroll: | 0.53 |
| Probability of payment, given click | 0.1093125 |

Hint 1: Make sure you are only using information given in the [**table of baseline values**](https://docs.google.com/a/knowlabs.com/spreadsheets/d/1MYNUtC47Pg8hdoCjOXaHqF-thheGpUshrFA21BAJnNc/edit#gid=0). Do not use the results of the experiment, since this step should be done before the experiment is run.

Hint 2: Make sure you figure out how many units of analysis will correspond to 5000 pageviews for each metric. Again, use the given [**baseline values**](https://docs.google.com/a/knowlabs.com/spreadsheets/d/1MYNUtC47Pg8hdoCjOXaHqF-thheGpUshrFA21BAJnNc/edit#gid=0).



<https://stats.stackexchange.com/questions/605466/how-to-get-python-statsmodels-to-match-evan-millers-famous-ab-test-sample-size>

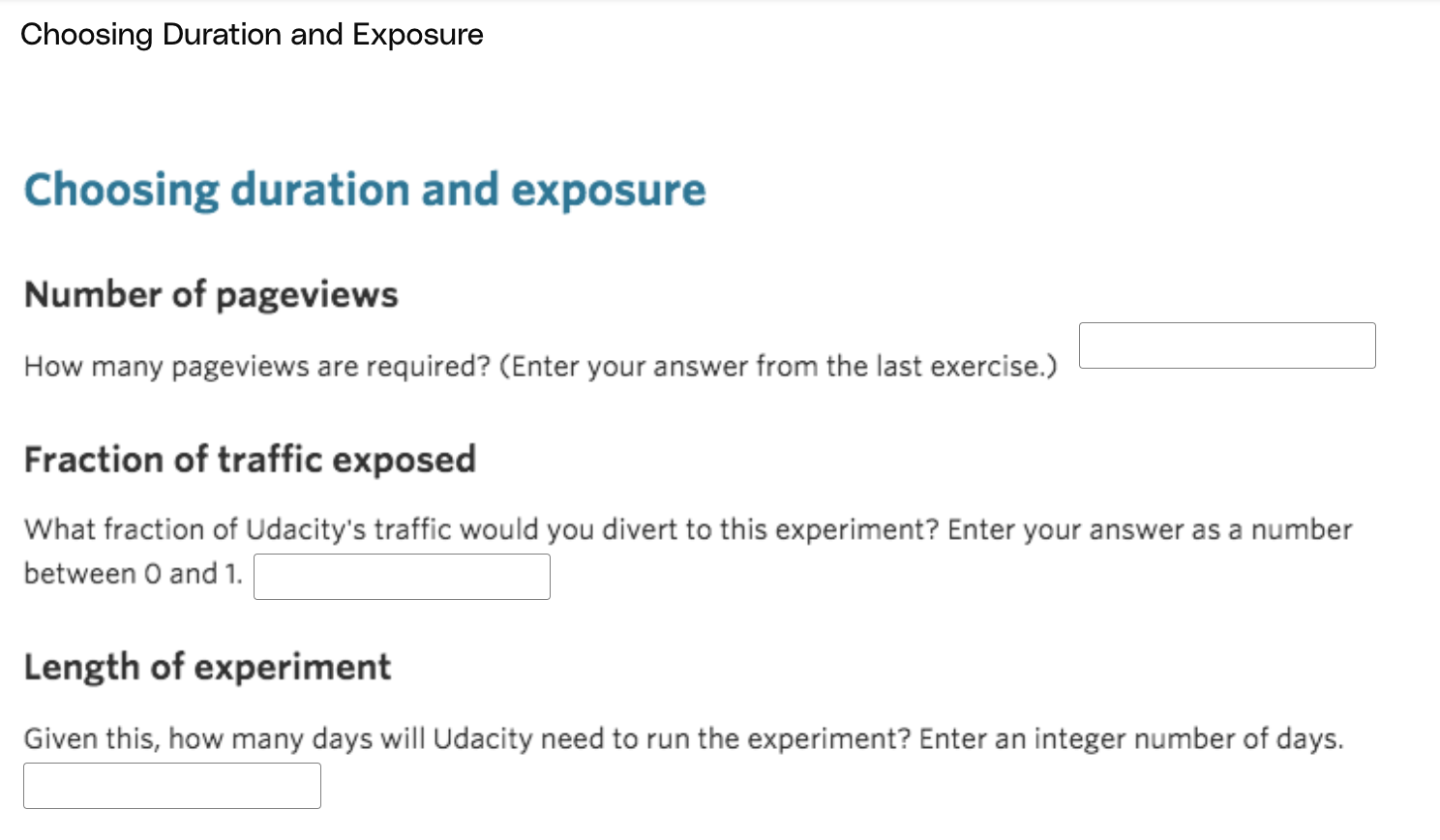
Use this code to replicate the evans millers awesome calculator

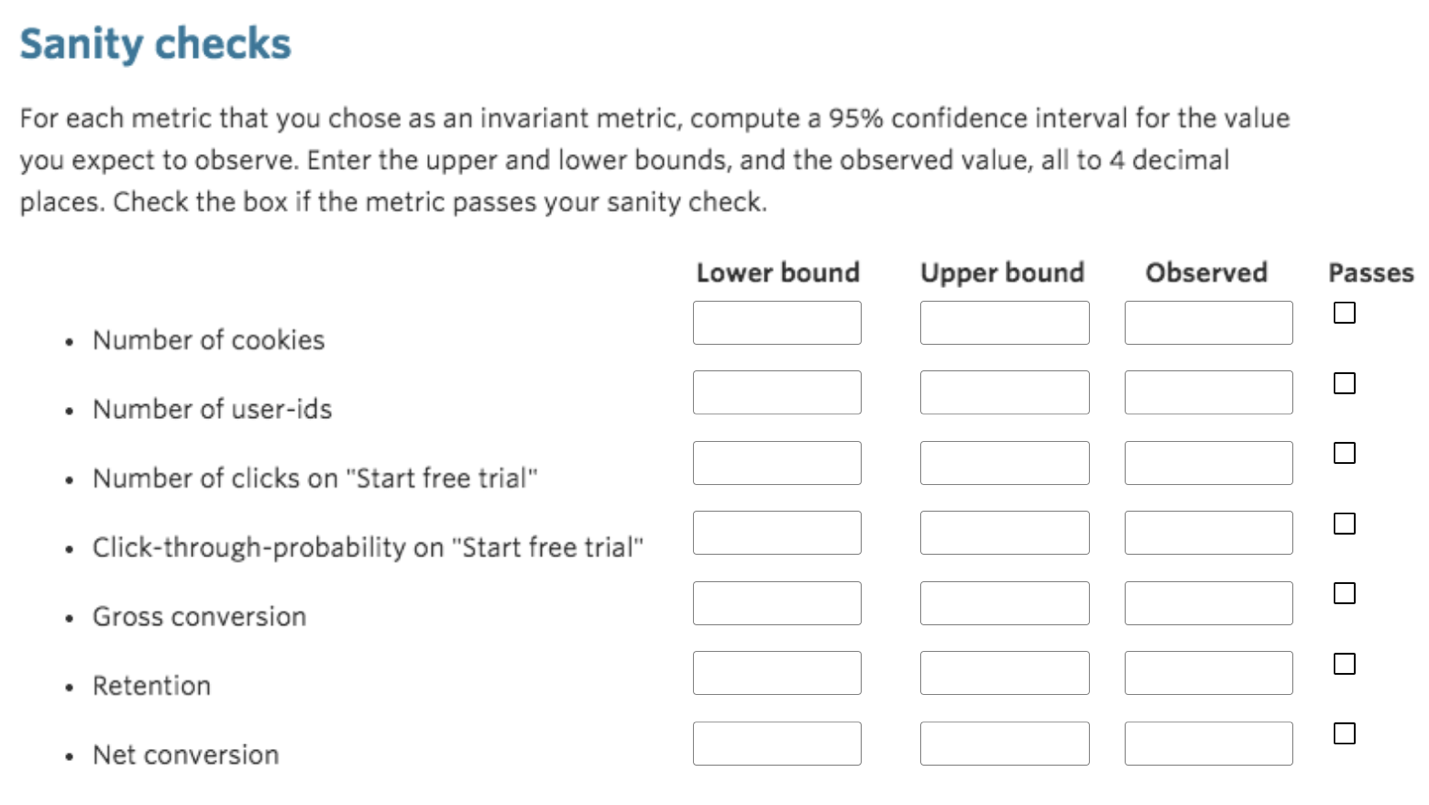
As we now have more than one hypothesis, the chance to get false positives increases. However, our metrics are not fully independent which is why the true probability for false positives will still be lower than 9.75% (that's the case for independent metrics). We could use Bonferroni but it is too strict <https://multithreaded.stitchfix.com/blog/2015/10/15/multiple-hypothesis-testing/>

given that the chance to get more false positives is only slightly increased in this case, we won't control for multiple hypothese here.

Given our calculations, we would need around 638,940 pageviews (cookies) to test the first hypothesis (given our assumptions on alpha, beta, baseline conversions and dmin). To additionally test the third hypothesis, we would need a total of 685,336 pageviews. And, in case we would like to also test the second hypothesis, we would need a total of around 4,737,771 pageviews. Need too many sample to test out the retention, so only include the GC(gross conversion), and NC(net conversion)

Used evan’s awesome calculator: https://www.evanmiller.org/ab-testing/sample-size.html





Abandon the retention because it needs too many pageviews. In the end, we need 685,325 page views. Each day Udacity has 40,000 page views. *45% and 50%, choose 49%*

685325/(40000\*0.49)=35, which is the multiple of 7. Assuming the pageview behavior is seasonal, we want to have the number of days be the multiple of 7.

We see that we would need to run the experiment for about 119 days in order to test all three hypotheses (and this does not even take into account the 14 additional days (free trial period) we have to wait until we can evaluate the experiment). Such a duration (esp. with 100% traffic diverted to it) appears to be very risky. First, we cannot perfom any other experiment during this period (opportunity costs). Secondly, if the treatment harms the user experience (frustrated students, inefficient coaching resources) and decreases conversion rates, we won't notice it (or cannot really say so) for more than four months (business risk). *Consequently, it seems more reasonable to only test the first and third hypothesis and to discard retention as an evaluation metric.* Especially since net conversion is a product of rentention and gross conversion, so that we might be able to draw inferences about the retention rate from the two remaining evaluation metrics.

So, how much traffic should we divert to the experiment? Given the considerations above, we want the experiment to run relatively fast and for not more than a few weeks. Also, as the nature of the experiment itself does not seem to be very risky (e.g. the treatment doesn't involve a feature that is critical with regards to potential media coverage), we can be confident in diverting a high percentage of traffic to the experiment. Still, since there is always the potential that something goes wrong during implemention, we may not want to divert all of our traffic to it. Hence, 80% (22 days) would seem to be quite reasonable. *However, when we look at the data provided by Udacity (see 4.1) we see that it takes 37 days to collect 690,203 pageviews (sum of all the treatment and control group pageviews), meaning that they most likely diverted somewhere between 45% and 50% of their traffic to the experiment*

*If 45% traffic, then*

\frac{690203}{37\cdot40000}

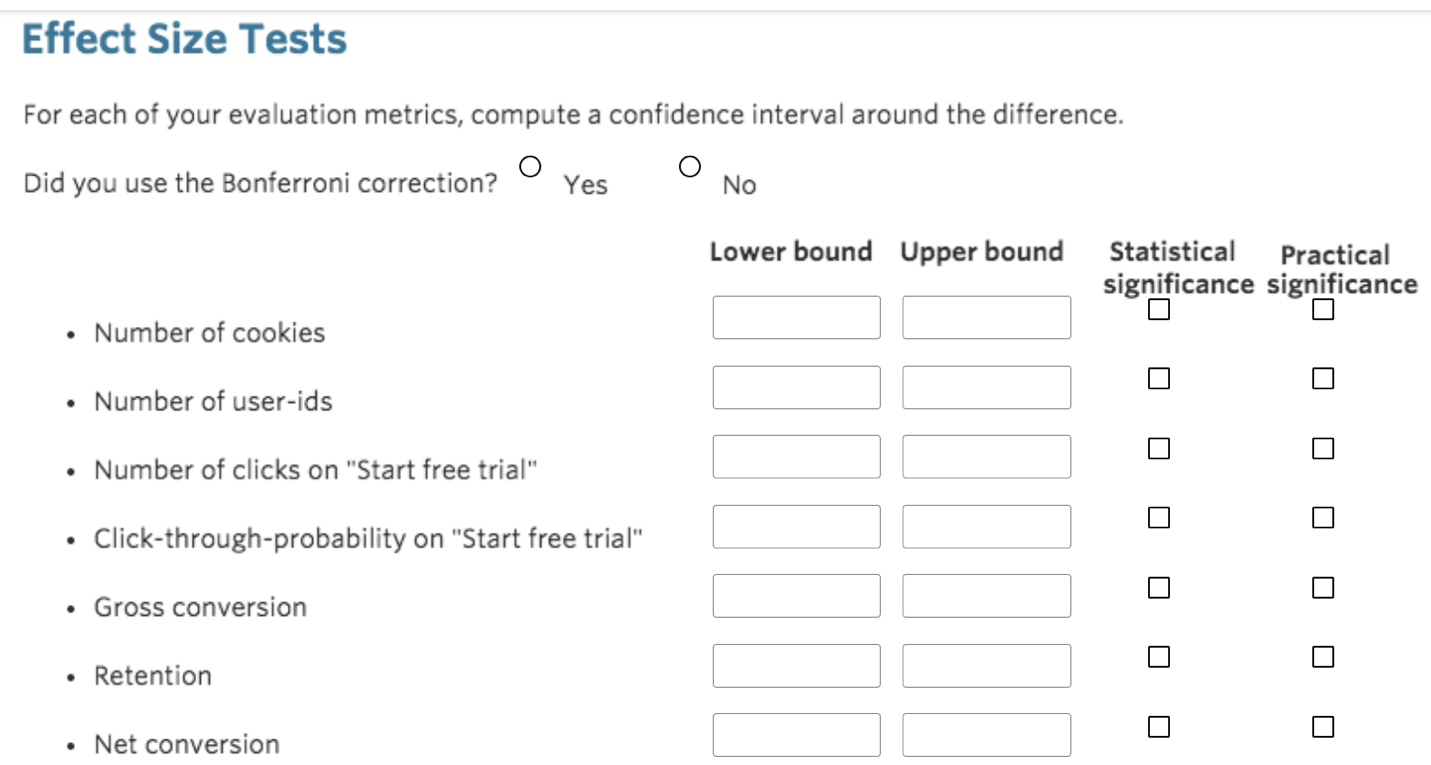
To get the upper and lower bound, To check whether the click-through probabilites in the control and treatment groups are significantly different from each other, we conduct a two proportion z-test with a click being interpreted as a success. We thereby assume that the two populations have normal distributions but not necessarily equal variances (hence p is not pooled below). To perform the test, we can calculate a confidence interval around the expected difference of the two metrics which is 0. Alternatively, we can calculate the Z-test-statistic and then check the corresponding p-value.

Further instructions

For a count (such as the first three metrics), you should calculate a confidence interval around the fraction of events you **expect** to be assigned to the control group, and the observed value should be the actual fraction that was assigned to the control group. This is the process that was shown in Lesson 5.

For any other type of metric, (such as the last four metrics), you should construct a confidence interval for a difference in proportions using a similar strategy as in Lesson 1, then check whether the difference between group values falls within that confidence level.

If a metric is not an invariant metric, leave the corresponding boxes blank.



Significance definitions

A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)

