## **PART ONE: Introduction**

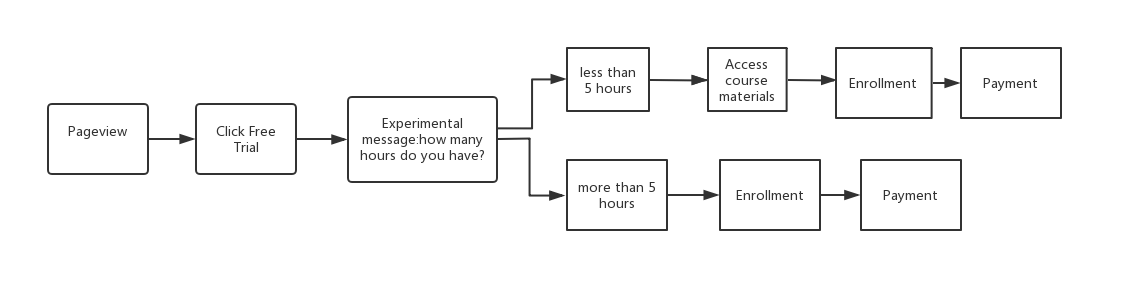
In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how many time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 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.

**Null hypothesis():** This change cannot result to the significant change and may not effective in helping to reduce early course cancellation.

**Alternative hypothesis ():**This change can help reduce 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.

After using this suggestion page, the process of students who click this page is like:



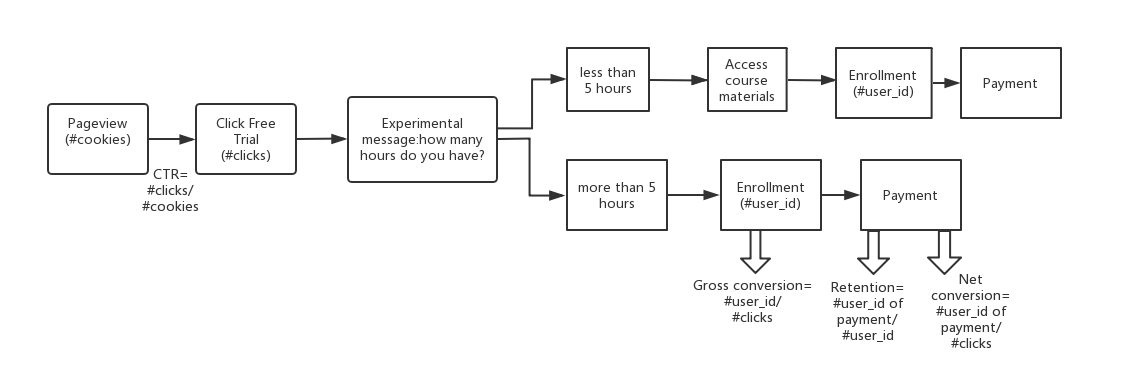
## **PART TWO: Metric choice**

For metrics listed below, which of the following metrics would you choose to measure for this experiment and why?

Hint: Any place "unique cookies" are mentioned, the uniqueness is determined by day. (That is, the same cookie visiting on different days would be counted twice.) User­ids are automatically unique since the site does not allow the same user­id to enroll twice.

* **Number of cookies:** That is, number of unique cookies to view the course overview page. (dmin=3000)
* **Number of user ­ids:** That is, number of users who enroll in the free trial. (dmin=50)
* **Number of clicks:** That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (dmin=240)
* **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)
* **Gross conversion:**That is, number of user­ids 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)
* **Retention:**That is, number of user­ids to remain enrolled past the 14­day boundary (and thus make at least one payment) divided by number of user­ids to complete checkout. (dmin=0.01)
* **Net conversion:** That is, number of user­ids to remain enrolled past the 14­day 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)

For all the metrics mentioned, we can show it in the flowchart about how to calculate.



After analysis, we conclude the invariant metrics and variant metrics.

**Invariant metrics:** number of cookies, number of clicks, click-through-probability

**Variant metrics for evaluation:** gross conversion, retention, net conversion

Here are the reasons:

1. **Number of cookies:** In experiment, we would like to make sure the cookies are even distributed between the control group as well as experiment group.
2. **Number of user\_ids:** this metrics cannot be regarded as an invariant metrics since it has changed if we introduce the experimental message. However, it is not also a good metric for evaluation.we would also expect the number of user-ids 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.
3. **Number of clicks:** Because the click free trial movement happens before introducing the experimental message, the elements of the experiment would not be expected to impact clicking the “start free trial” button. And we also want the number of clicks divided the same to make sure the effective analyze procedure.
4. **Click-through-probability:** number of cookies to click “start free trial”/number of cookies to view the course overview. All the two processes happen before introducing the experimental message. Hence, we expect equal distribution in both groups.
5. **Gross conversion:** This is a good metric for analyzing. If the experimental change is effective, the gross conversion of experimental group will be smaller than the control group. This is because we use the time recommendation to filter students who have less time to enroll classes and to increase students who paid the enrollment after free trial.
6. **Retention:** This is also a good metric for analyzing. We expect the retention rate to increase as the treatment should have filtered out users who are likely to churn.
7. **Net conversion:** We also want to see more students remained and paid the course after 14 days tree trial. Therefore, from a business perspective, we hope to see an increase we want it to increase in the experimental group, but it could vary in each direction or stay constant depending on gross conversion and retention effects.

**Summary:** For our experiment, gross conversion, retention and net conversion are good metrics for evaluation. If the change we made is effective, there is highly likely that the gross conversion will decrease, retention and net conversion will increase.

## **PART THREE: Measuring Variability**

For each metric you selected as an evaluation metric, estimate its**standard deviation analytically.** We need to calculate the **standard deviation of the sampling distribution** of the sample mean (**standard error**, in short) for each of the evaluation metrics. To be more precise, in this case we calculate the estimated standard errors of the sample proportions as our evaluation metrics are probabilities. The standard error is an estimate of how far the sample proportion is likely to be from the population proportion.

In this part, we use the Jupyter Notebook to calculate variability.Here are the steps:

1. Do scaling. We use 5000 cookies to estimate the variability, however, the baseline data contains 40000 cookies. Therefore, the use\_ids and clicks also change with the total cookies. We need to shrink the clicks and cookies to the same scale. The scaling feature is 5000/40000
2. Because the unit of diversion is cookie which is the same of the unit of analyze, we can use the data to analyze variability instead of empirical analysis.
3. The formula is ,we can calculate the variability of gross conversion, retention and net conversion

## **PART 4: Size**

Page views required for each metric can be calculated separately through online calculator or use the code. The experiment has already set the .

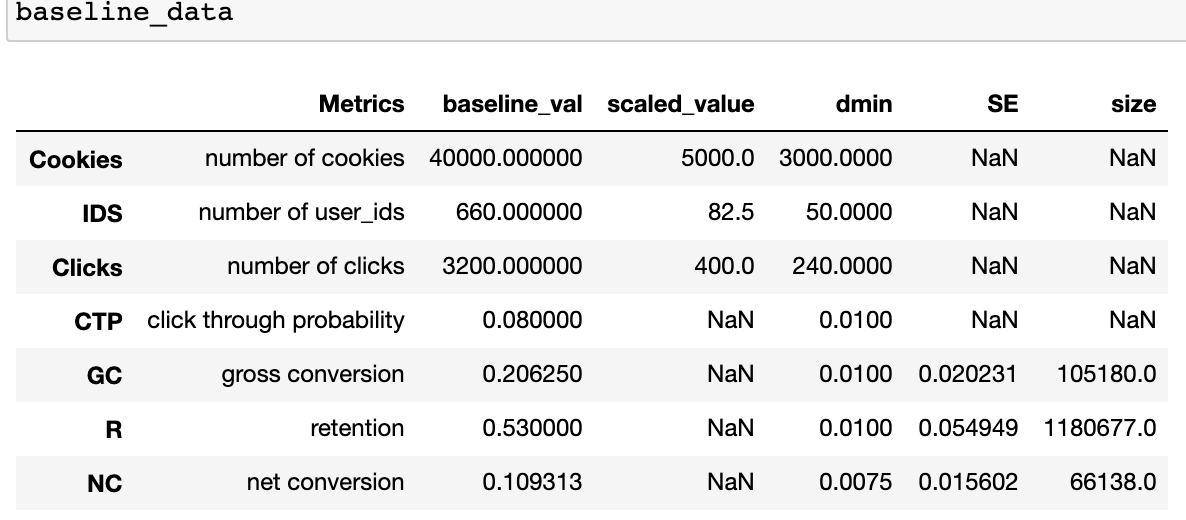
If we assume that the standard deviations of the population proportion and the sample proportion are equal and that also the sample sizes of treatment and control group are the same, then the required experiment sample size per group can be determined through another method. We can use the formula and Jupyter Notebook to analyze it.For the basic sample size formula for two groups, with a two-sided alternative, normal distribution with homogeneous variances and equal sample sizes.Here is another method to calculate n

can also be expressed as ,and ,so we can calculate size by using this formula. And the whole inference of the formula is attached as the supplementary materials.

When we calculate the experiment sample size we have to keep in mind that n gives us the sample size per group. For a classical A/B test, there are two groups. Further, we want to calculate the experiment sample size in terms of cookies that visit the page. Thus, we also need to account for the circumstance that our evaluation metrics' units of analysis are clicks and user-ids, respectively.

The total experiment sample size per evaluation metric is hence given by:

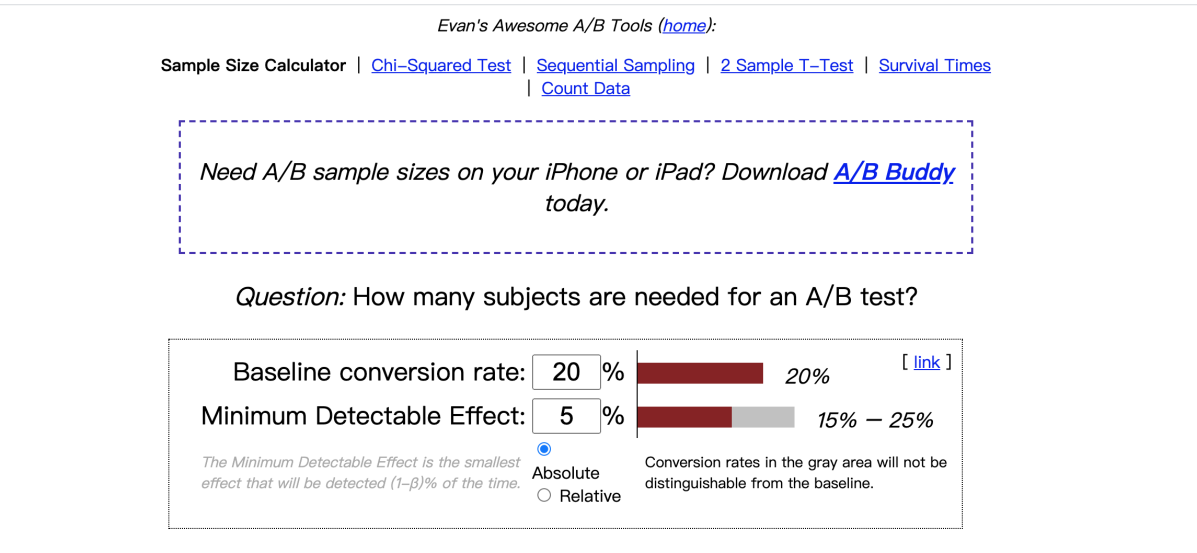
After using the Jupyter Notebook, we can conclude the results:

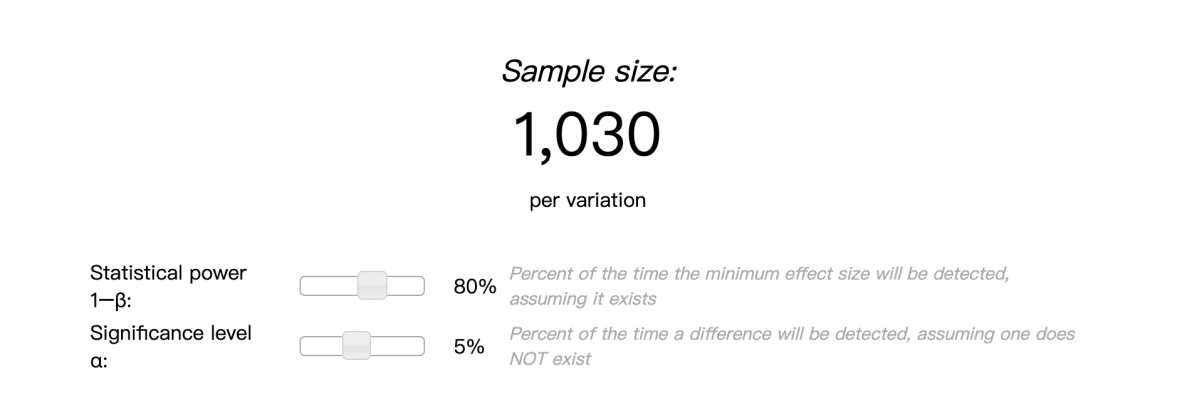


Given our calculations, we would need around 105,180 pageviews (cookies) to test the first hypothesis (given our assumptions on alpha, beta, baseline conversions and dmin). To additionally test the third hypothesis, we would need a total of 66,138 pageviews. And, in case we would like to also test the second hypothesis, we would need a total of around 1,180,677 pageviews.

If we do not assume common standard deviations then a more precise way to determine the required sample size would be more complex, and the online calculator of Evan Miller uses this method. Here is the screenshot of the online calculator:

(link: <https://www.evanmiller.org/ab-testing/sample-size.html>)





From the online calculator, we need to provide some information: baseline conversion rate, minimum detectable effect, statistical power and significance level. For different metrics, the baseline conversion rates are different, so we need to calculate each of metrics separately.

* Gross conversion

Baseline conversion rate: 0.206250

Minimum detectable effect (dmin): 0.01

Statistical power: 0.8

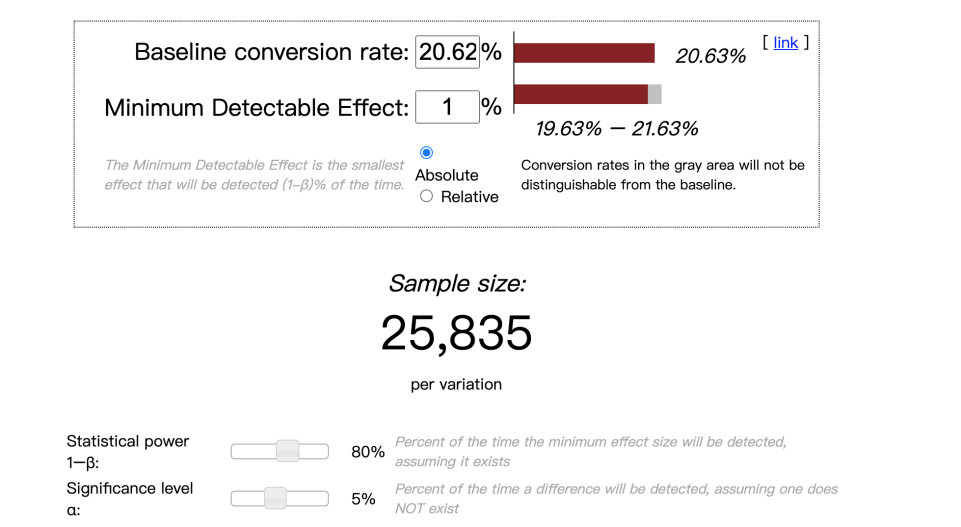
Significance level: 0.05

Size: 25,835

Since we have two groups (experimental and control group), the total size is 25,835\*2=51,670;

Clicks/pageviews=3200/40000=0.08

Then, the pageviews required is 51,670/0.08 = 6,45,875



* Retention

Baseline conversion rate: 0.53000

Minimum detectable effect (dmin): 0.01

Statistical power: 0.8

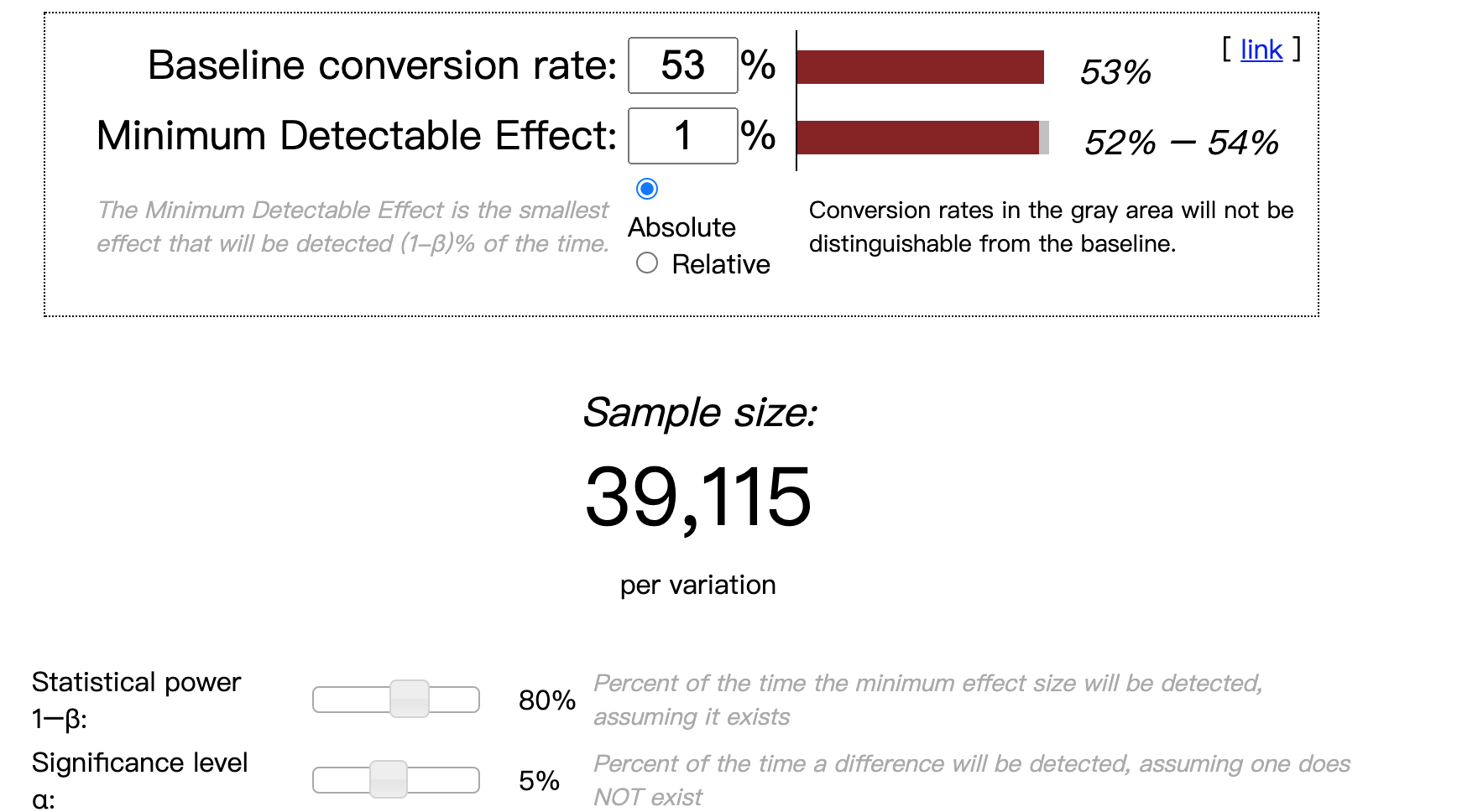
Significance level: 0.05

Size: 39,115

Since we have two groups (experimental and control group), the total size is 39,115\*2=78,230;

Enrolls/pageviews=660/40000=0.0165

Then, the pageviews required is 51,670/0.0165 =4,741,212



* Net conversion

Baseline conversion rate: 0.109313

Minimum detectable effect (dmin): 0.0075

Statistical power: 0.8

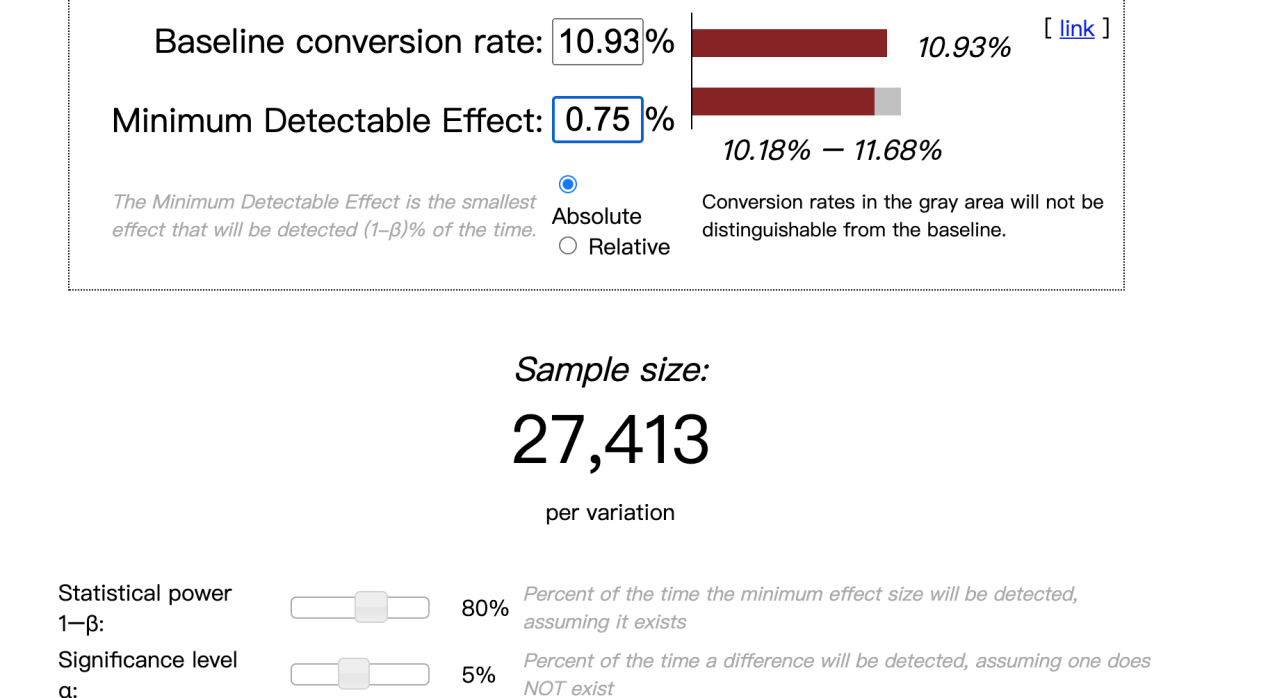
Significance level: 0.05

Size: 27,413

Since we have two groups (experimental and control group), the total size is 27,413\*2=54,826;

Clicks/pageviews=3200/40000=0.08

Then, the pageviews required is 54,826/0.08 = 6,85,325



**Summary:** Pageviews required is the maximum of pageviews required for gross conversion, retention as well as net conversion. Hence, the size required is 47,41,212.

## **PART 5: Duration and Traffic**

After determining the size, next step is to ensure duration, how many days do we need to run the experiment to reach the size? According to the given information, we can assume there is no other experiments we want to run simultaneously. We use the first result of online calculator.

Therefore, normally we can divert 100% of traffic in our experiment. Given our estimation that there are about 40,000 unique pageviews per day, this would result in:

For gross conversion:6,45,875/40,000=16.15

For gross conversion+net conversion:6,85,325/40,000=17.13

For gross conversion+retention+net conversion: 47,41,212/40,000=118.53

From the result, you can see that we need at least 119 days to run this experiment. And the duration doesn’t contain 14 trial days for us to analyze results. The 119-days duration is very risky. First, during these days, we cannot run other experiment which has opportunity costs. Second, if this change does harm to our customers, such as frustrated students and inefficient teaching resources, we won’t even notice the affection in first 4 months which is 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 retention and gross conversion, so that we might be able to draw inferences about the retention rate from the two remaining evaluation metrics.

Next question is what’s the suitable diverting proportion? How much traffic should we divert to the experiment? From the results we calculated before, if we divert 100% of the traffic, we will run the experiment at least 18 days. However, since there is always the potential that something goes wrong during implementation, 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 we see that it takes 37 days to collect 690,203 pageviews, meaning that they most likely diverted somewhere between 45% and 50% of their traffic to the experiment.

## **PART 6: Analyze**

## **6.1 Sanity check:Clicks & Cookies**

After running the whole experiment, we need to analyze the result. The first step is always doing sanity check to ensure we have done the experiment and the result is validated. As stated earlier, we would expect that these metrics do not differ significantly between control and treatment group. Otherwise, this would imply that something is wrong with the experiment setup and that our results are biased.

For our invariant metrics cookies, we would expect that the total amount of cookies in the control group is the same as the total amount in the experimental group, each group accounts for 50% of the combined number of total cookies (control+experimental) as they should be assigned randomly. we can use the binominal distribution to model the number of successes in the given sample (treatment+control) and perform a binomial test/one-proportion z-test as a sanity check.

Here are steps about sanity check of cookies, clicks:

1. Sum the total number of control and experimental group
2. Calculate the standard error of the cookies, we use this formula to calculate

And we have assumed the event of customer view the pages is a binomial distribution, therefore, the formula can be expressed like

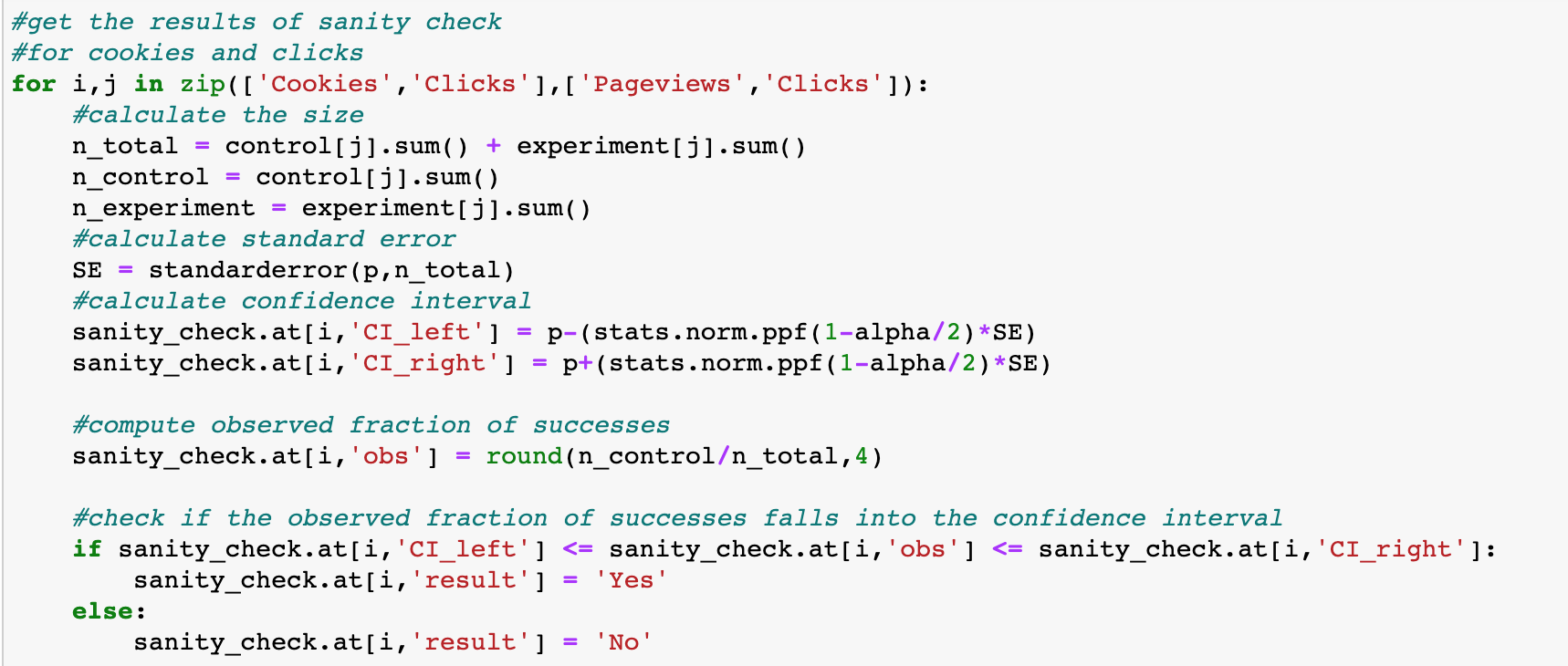
3) Since N is large, we can further assume that the sampling distribution of the sample proportion approaches a normal distribution (Central Limit Theorem)

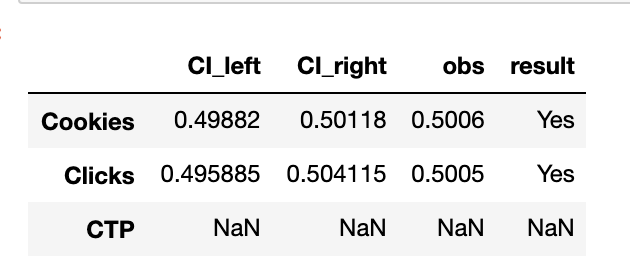
Use the SE to multiply z-score, we get the the margin,

1. When we have the margin, we can get the confidence interval

5) Compare the cookies in the control group, if the falls into the confidence interval, which means that this metric pass the sanity check hypothesis, we can regard the metric is an invariant metric.

Here we use the Jupyter Notebook to calculate the results.





From the results, we can see that for the invariant metrics: cookies, the confidence interval is [0.49882,0.50118],and the observed fraction of successes is 0.5006, here is the obs falls into the confidence interval, and pass the sanity check, and also for the clicks metrics, it also passes the sanity check.

**6.2 Sanity check: Cookie through probability**

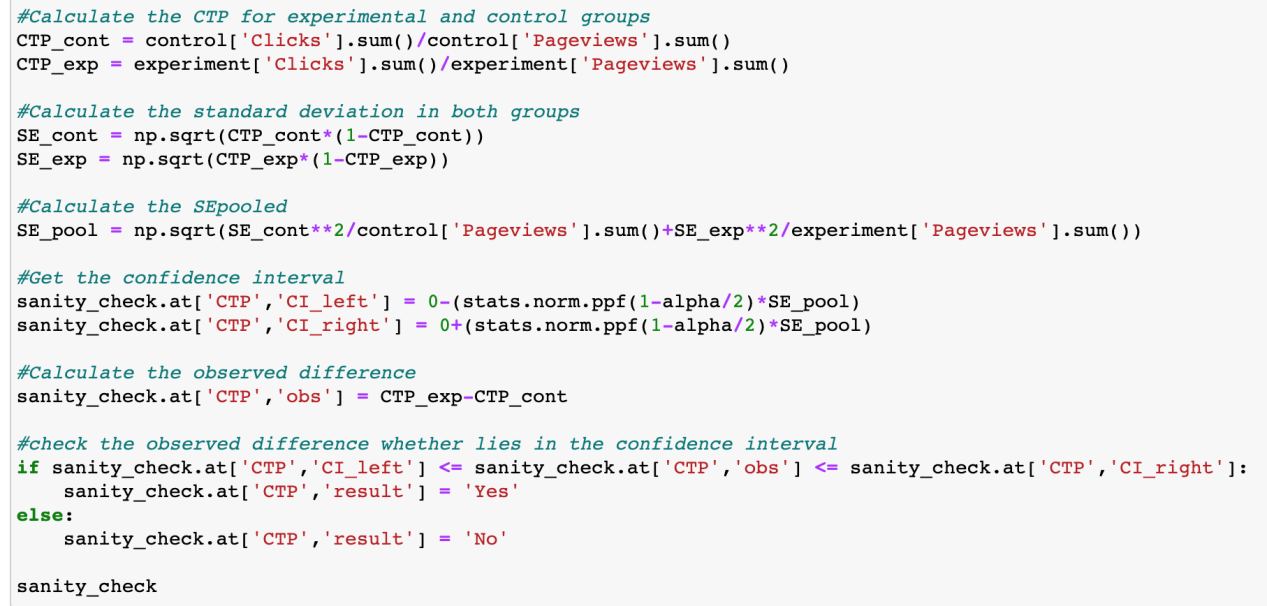
To check whether the click-through probabilities in the control and treatment groups are significantly different from each other, we thereby assume that the two populations have normal distributions but not necessarily equal variances (hence p is not pooled below).

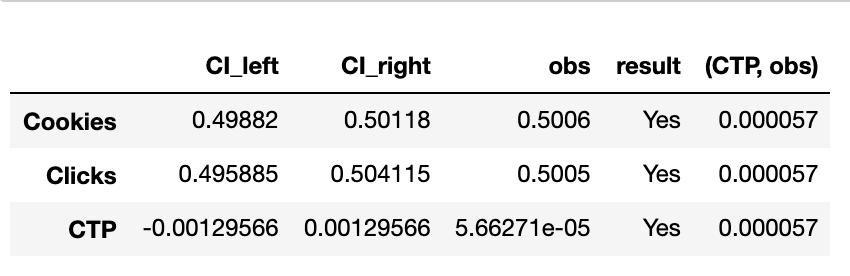
We conduct a two proportion z-test with a click being interpreted as a success. And in this method, to perform the test, we can calculate a confidence interval around the expected difference of the two metrics which is 0.

Here are steps about sanity check about CTP:

1. Calculate the CTP for both experimental and control group,using the formulas
2. Calculate the standard deviation for both two groups,by using
3. Calculate the ,by using
4. Get the confidence interval,
5. Compute the observed difference between the two metrics d and check whether d lies within CI, where

Here are the codes and results of sanity check about CTP





From the results, we know that all three invariant metrics have passed the sanity check, therefore, we can move to the result analyze.

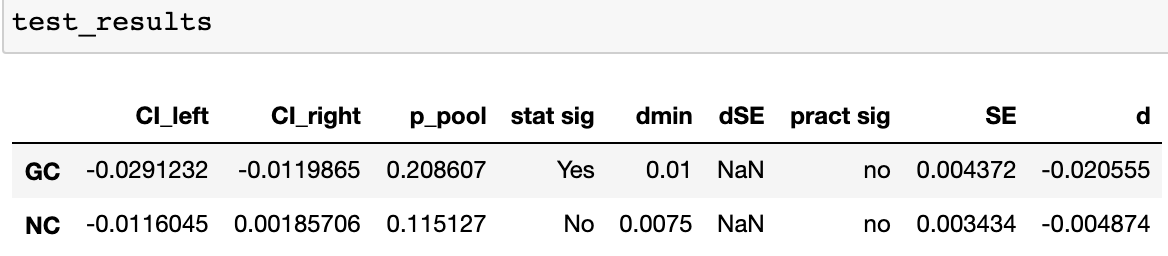
## **6.3 Results analyze**

Finally, we move to the step of analyzing results. We would like to the change we have implemented has an effective affection or not. We need to analyze results of three evaluation metrics one by one, and notice that we need to rule out the days which don’t have enough data. In this experiment, we don’t count the days after Nov,3rd, since there are no data about enrollments and payments. As could be seen in the table when we loaded the data, "payments" (and strangely also "enrollments") were only tracked for 37 days (23+14 days) and not for 51 days (37+14 days) which would have been necessary in order to fully account for the 14-day trial period. Consequently, in our actual A/B test, the true sample size is lower (n\_true = 423,525) than we initially aimed for (n = 685,336)

Here are the steps about analyzing results:

1. For each metric, calculate the ,
2. Use the results in the first step to calculate the observed difference
3. Calculate the
4. Calculate the )
5. Calculate the confidence interval
6. Check whether the confidence interval contains 0 or not to test the statistical significance and whether the lower bound of confidence interval larger than the dmin to test practical significance.

Here are the results about Jupyter Notebook



We can see the gross conversion metric has the statistical significance, however, the net conversion metric doesn’t have statistical significance. And they both do not have practical significance.

**6.4 Validate results**

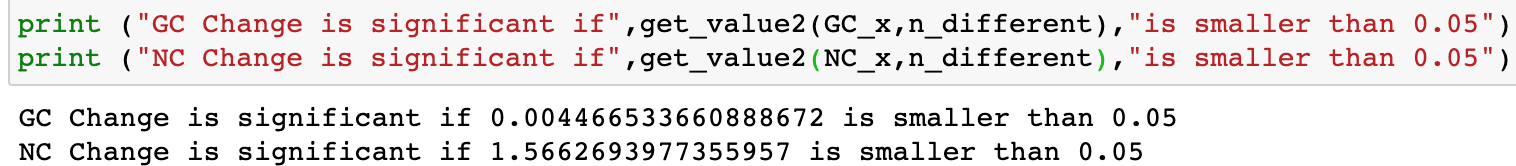
Sign test is an another method to validate the results obtained above. We can use the sign test the test our results.

Under normal circumstances, the sign test method is used to detect whether there is a difference between the means of two samples. It is assumed that this result is random every day, and there are only two results of success or failure, so it can be set as a binomial distribution. Here are the steps about sign test.

1. We use ,...,) to test the null hypothesis(there is no difference between two groups).
2. Define m is the number of ,...,) which are not 0.
3. Calculate the probability of success. If ,)

4) Compare the p of success with alpha=0.05, if the result is larger than alpha, we accept the null hypothesis, regarding there is no significant difference between two groups, whereas, if p is smaller than 0.05, we reject null hypothesis.

Here are the results of sign test.



Hence, we can conclude that for gross conversion, the p is 0.0045 which is smaller than alpha, we reject the null hypothesis. However, for net conversion the p is 1.566 which is much larger than 0.05, we accept the null hypothesis.

## **PART 7: Summary**

An experiment was conducted in which potential Udacity students were divided into two groups: experimental and control group. This experiment is used to test one change made by Udacity platform can help reduce the number of frustrated students who left the free trial because they didn't have enough time. The experiment group was asked to input more time to study, after clicking the ‘start free trail button’, whereas the control group was not. **First** we define the null hypothesis and alternative hypothesis by carding whole process. **Second,** after determining the test, we also choose three invariant metrics(cookies,clicks and click through probability) while three variant metrics for evaluation (gross conversion,retention and net conversion). **Third,** we measure the variability about three evaluation metrics by calculating analytically instead of empirical calculating since the unit of diversion is cookie which is the same of the unit of analyze. **Fourth,** we need to calculate the size of testing, we use the online calculator to ensure the different sizes for three metrics. We also introduce the Jupyter Notebook to use another method to calculate the size. **Fifth,** after determining the size, if we assume there is no other experiment works simultaneously, if we use the size which meet three metrics and 100% traffic, the duration will be 119 days. This duration is unrealistic since it will result to some business costs and risks, therefore, we choose the size which meet two metrics (gross conversion,net conversion). In this case, the duration is much shorter(18 days). **Sixth,**since in the most of circumstances, we can’t divert 100% of the traffic due to realistic implementations and risks, if we divert 80% of the traffic, the duration is 22 days. **Seventh,** we move to the process of analyzing. We want to do the sanity check first to ensure that our experiment works correctly**.Finally,** after passing the sanity check, we can analyze our results about different metrics and then we can also use the sign test to validate results.

Because our acceptance criteria requires statistically significant differences for all metrics, the use of the Bonferonni correction is not appropriate. The Bonferonni correction is a method for controlling type I errors (false positives) when using multiple metrics in which relevance of any of the metrics matches the hypothesis. In this case the risk of type I errors increase as the number of metrics increases (significance by random chance). In our case in which all metrics must be relevant to launch, the risk of type II errors (true negatives) increases as the number of metrics increase, so it stands to reason that controlling for false negatives is not consistent to our acceptance criteria.

## **PART 8: Recommendation**

Through comprehensive sanity check, result analyze and sign test, we can conclude there is a statistical significance, sign test significance in gross conversion. However, we cannot get the statistical significance, sign test significance in net conversion. And there are not practical significance in both conversions. Hence, a decrease in enrollment not coupled to an increase in students staying for requisite 14 days to trigger payment.

Considering this, my recommendation is not to launch this change, but rather to pursue other experiments.