Template Format

This template can be used to organize your answers to the final project. Items that should be copied from your answers to the quizzes should be given in blue.

Intro: General description of the experiment

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

Experiment Design

1. Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

1. Answer student metric choice

In order to answer this question let us remind ourselves of the definition:

- **Invariant metric** is a metric that should not be affected by the changes in the experiment
- **Evaluation metrics** are metrics used to measure the impact or the changes made in the experiment

Invariant metrics:

- Number of cookies: That is, number of unique cookies to view the course overview page.
    - As this is the unit of diversion, we can expect an even distribution amongst the control and experiment groups. Hence, it is an appropriate invariant metric.
- Number of clicks: That is, number of unique cookies to click the "Start free trial" button.
    - As this happens before the free trial screener is triggered, we can again expect an even distribution amongst the control and experiment groups or in other words: at this stage of the funnel the experience is the same for all
- 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.
    - Same reasoning as for cookies and clicks: we can again expect an even distribution amongst the control and experiment groups or in other words: at this stage of the funnel the experience is the same for all

Evaluation metrics

- 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.
    - Clearly these are metrics measuring the impact of the experiment or in other words the success (checkout, enrollment) or failure of the experiment. Hence, it an evaluation metrics
    - **This is the metric we want to decrease as it would be connected to less student frustration and less students leaving after the free trial**
- con: 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.
    - Clearly these are metrics measuring the impact of the experiment or in other words the success (remaining enrollment and hence pay) or failure (not enrolled more than 14d) of the experiment. Hence, it an evaluation metrics
    - **This is the metric we want to increase as it would be connected to less student frustration and more students staying after the free trial**

For which results would we wish to launch the experiment?

- Gross conversion: This is the metric we want to decrease as it would be connected to less student frustration and less students leaving after the free trial
- Net conversion: This is the metric we want to increase as it would be connected to less student frustration and more students staying after the free trial

2. Comment on remaining metrics

In the frame of the Final Project quiz two metrics seem unused so we disregard them for now:

- Number of user-ids: That is, number of users who enroll in the free trial.
- 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.
    - Could probably also be used as an evaluation metric

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

1. Measuring Standard Deviation

Comment: We are calculating the analytical standard deviation for our two evaluation metrics: **Gross conversion and Net conversion**

We are using the following formula:

$$GrossConvSTDV = \sqrt{Probabilityofenrolling * (1 - Probabilityofenrolling)/(Pageviews * Uniquecookiestoclick\text{ " }Startfreetrial\text{ " }p.d.) /Uniquecookiestoviewcourseoverviewpagep.d.)}$$

$$NetConvSTDV = \sqrt{Probabilityofpaymentgivenclick * (1 - Probabilityofpaymentgivenclick)/(Pageviews * Uniquecookiestoclick\text{ " }Startfreetrial\text{ " }p.d.) /Uniquecookiestoviewcourseoverviewpagep.d.)}$$

Additionally we calculate the standard dev of the retention.

Results of Standard deviations rounded:

Gross conversion Std. DEV	Net conversion Std. DEV	Retention Std. DEV
0.0202	0.0156	0.0549

2. Empirical versus analytical standard deviation of evaluation metrics

In this case I expect the analytics versus empirical values to be quite different. The main reason for that is the small sample size. In the control and experiment data there are only 23

data points, being quite a small sample size. Having more data points in the sample would probably improve the gap.
Also, other impacts like the distribution:
As the analytical standard deviation assumes a specific distribution, in other words the normal distribution the distribution of the empirical data should also follow the normal distribution.

Calculating the empirical standard deviation:

- We are also checking the empirical variability of the gross and net conversion to check against. To do so, we need to import control and experiment data.
- After this step, we calculate the gross and net conversion per day and determine the standard deviation of this result for all 23 data points. This is to say: we calculate the average of the std dev per day, calculate the delta between the std dev per day and the average, which we take times and sum up for all days. After this, we divide by the number of samples (23) and take the square root.

Control group: Gross conversion Emp. Std. DEV	**Control group:** Net conversion Emp. Std. DEV	**Experiment group:** Gross conversion Emp. Std. DEV	**Experiment group:** Net conversion Emp. Std. DEV
0.044	0.0294	0.0475	0.0322

Interpretation of the empirical standard deviation:

- We find that values between control and experimental group do match quite well
- As expected in the reasoning, analytical and empirical results do not match very well
    - e.g. Gross conv. STD DEV: 0.0202 vs empirical value 0,044 or 0,0475 → twice as high
    - e.g. Net conv. STD DEV: 0.0156 vs empirical value 0.0294 or 0.0322→ twice as high

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

1. Answer student: Bonferroni correction

Usually the Bonferroni correction is used to adjust p-values in order to control the family-wise error rate (FWER) in multiple hypothesis testing or in other words the purpose is to reduce the probability of a type I error (false-positive) when making a larger number of hypothesis tests.

See: https://www.investopedia.com/terms/b/bonferroni-test.asp

In this case we are only a control and experimental group with 23 observations. **Hence, the number of tests is fairly small and I will not use the Bonferroni correction.**

Besides, the Bonferroni correction is known for some quite heavy assumptions:

- Conservative approach resulting in quite low adjusted significance levels, which can make it quite difficult to detect true effects
- Independence assumption: Bonferroni correction assumes tests to be conducted independent of each other. This is not necessarily the case for Gross and Net Conversion
- Type II error rate: Bonferroni correction can indirectly affect the Type 2 error rate

1. **Calculation of the sizing algorithm:**
    1. First, we will need the **analytical standard deviations** from part 3 for gross and net conversion --> given from 3
    2. **Minimum Detectable Effect** that is given by dmin=0.01 and dmin= 0.0075
    3. We also need the **baseline conversion rates** for both metrics --> given by Probability of enrolling, given click and Probability of payment, given click
    4. **Sample Size**: using the inputs from 2 and 3 allows us to calculate the sample size using a calculator like https://www.evanmiller.org/ab-testing/sample-size.html
    5. **Total sample size:** Sample size * number of groups (2: control and experiment)
    6. **Number of page views:** = Total Sample Size / / Click-through-probability on "Start free trial" for each metric
    7. **Maximum number of page view**: take the higher number in 6.
    8. **Traffic required with 100%:** Maximum number of page view (from 7) / Unique cookies to view course overview page per day
2. **Summary of the metrics**

Metric	Gross Conversion	Net Conversion
Analytical standard deviations	0.0202	0.0156
Minimum Detectable Effect	0.01	0.0075
Baseline conversion rates	0.20625	0.109313
Sample Size	25835	27413
Total sample size	51670 (=Sample Size *2)	54826 (=Sample Size *2)

Number of page views	645875	685325
Maximum number of page view	685325	

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

3. General introduction

After having calculated the max number of pages views, we are checking the required traffic needed with 100% of the traffic and will then reflect on the risk and feasibility of the result

4. Calculation when using 100% of the traffic

Traffic required with 100% = Maximum number of page view (from 7) / Unique cookies to view course overview page per day

Using 100% of the traffic it would take this many days to run the experiment: 18 days

5. Reasoning on traffic selection

Using 100% of the traffic would be quite risky for different reasons:

- Other experiments: Udacity might have other experiments to do
- Risk: using 100% of the traffic would mean that we are using the new version for the complete traffic without having tested for bugs, problems on a smaller scope
- Technical problems: we might still have bugs or smaller issues on the new page and hence running it on 100% of the traffic seems very risk

Bottom line:

- It seems quite risky to use the full traffic. Hence, we will use **50% of the traffic**

6. Calculation of required traffic for 50%

Traffic required with 50% = Maximum number of page view (from 7) / (Unique cookies to view course overview page per day * 0,5)

Using 50% of the traffic it would take this many days to run the experiment: 35 days

Experiment Analysis

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.)

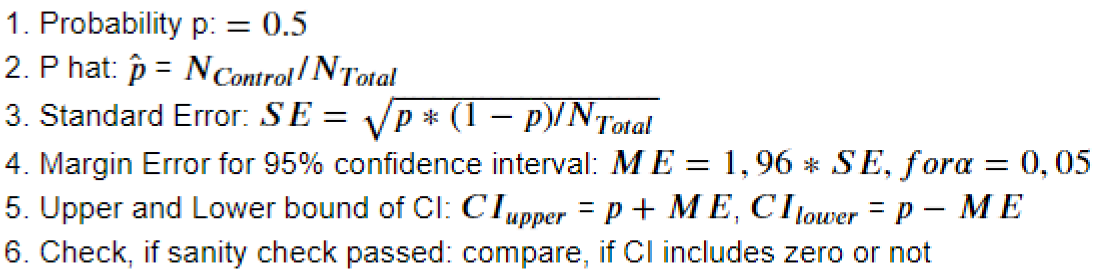
For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.**

7. Purpose:

In this section we will perform some sanity checks. The general purpose of this is to ensure that our experiment has been set up correctly and the data is reliable.
In other words: we can check for issues or even biases this way and assure that data between control and experimental group are statistically similar and comparable before proceeding with the analysis.
We will carry out these checks ofr our invariant metrics, as defined:

- Number of cookies: That is, number of unique cookies to view the course overview page.
- Number of clicks: That is, number of unique cookies to click the "Start free trial" button.
- 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.

8. Calculation of sanity checks algorithm



1. Probability p: $= 0.5$
2. P hat: $\hat{p}$ = $N_{Control} / N_{Total}$
3. Standard Error: $SE = \sqrt{p * (1 - p) / N_{Total}}$
4. Margin Error for 95% confidence interval: $ME = 1,96 * SE, for \alpha = 0,05$
5. Upper and Lower bound of CI: CI_{upper} = $p + ME$, CI_{lower} = $p - ME$
6. Check, if sanity check passed: compare, if CI includes zero or not

9. Calculation of sanity checks summary

Metric	Probability p (expected value)	P hat (observed value)	Standard Error	Margin Error for 95% CI	Lower bound of CI	Upper bound of CI	Check, if sanity check passed
Number of cookies	0.5	0.5006	0.0006	0.0012	0.4988	0.5012	YES, PASSED ✔️
Number of clicks	0.5	0.5005	0.0021	0.0012	0.4959	0.5042	YES, PASSED ✔️
Click-through-probability	0.0821	0.0822	0.0005	0.0009	0.0812	0.0831	YES, PASSED ✔️

10. Sanity check passed reasoning

Cookies
--> PASSED, CI does not include zero ✔️

Number of clicks
--> PASSED, CI does not include zero ✔️

CTR
--> PASSED, CI does not include zero ✔️

11. Sanity check Summary ✔️

All sanity checks for our three invariant metrics:

- Number of cookies
- Number of clicks: That is, number of unique cookies to click the "Start free trial" button.
- Click-through-probability:

have been checked and passed the sanity check as all observed values were between the upper and lower bounds of the CI. **Hence, we can proceed with the Effect Size Tests.**

Result Analysis

Effect Size Tests

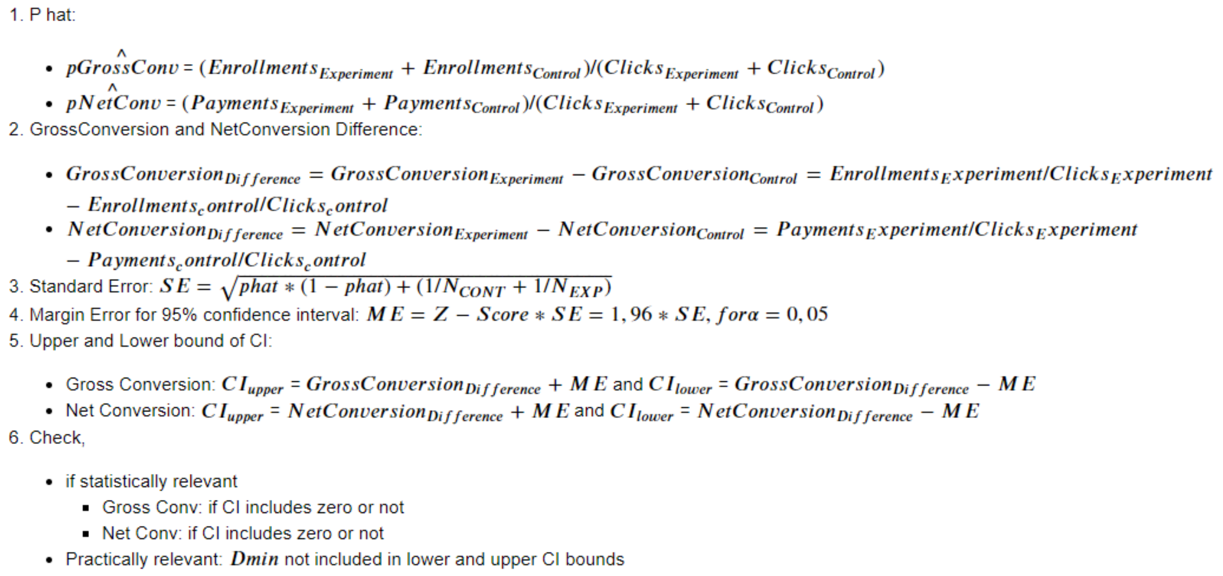
For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)

1. Purpose

The Effect Size test tells us how meaningful the relationship between variables or the difference between groups is or in other words it tells us, if a difference between a control and experiment group is also practically significant.

Sources:
https://www.scribbr.com/statistics/effect-size/#:~:text=Effect%20size%20tells%20you%20how,size%20indicates%20limited%20practical%20applications.

2. Calculation of Effect Size Test algorithm



1. P hat:
   - $pGro\hat{s}sConv = (Enrollments_{Experiment} + Enrollments_{Control})/(Clicks_{Experiment} + Clicks_{Control})$
   - $pN\hat{e}tConv = (Payments_{Experiment} + Payments_{Control})/(Clicks_{Experiment} + Clicks_{Control})$
2. GrossConversion and NetConversion Difference:
   - $GrossConversion_{Difference} = GrossConversion_{Experiment} - GrossConversion_{Control} = Enrollments_Experiment/Clicks_Experiment - Enrollments_control/Clicks_control$
   - $NetConversion_{Difference} = NetConversion_{Experiment} - NetConversion_{Control} = Payments_Experiment/Clicks_Experiment - Payments_control/Clicks_control$
3. Standard Error: $SE = \sqrt{phat * (1 - phat) + (1/N_{CONT} + 1/N_{EXP})}$
4. Margin Error for 95% confidence interval: $ME = Z - Score * SE = 1,96 * SE, for \alpha = 0,05$
5. Upper and Lower bound of CI:
   - Gross Conversion: CI_{upper} = $GrossConversion_{Difference} + ME$ and CI_{lower} = $GrossConversion_{Difference} - ME$
   - Net Conversion: CI_{upper} = $NetConversion_{Difference} + ME$ and CI_{lower} = $NetConversion_{Difference} - ME$
6. Check,
   - if statistically relevant
     - Gross Conv: if CI includes zero or not
     - Net Conv: if CI includes zero or not
   - Practically relevant: $Dmin$ not included in lower and upper CI bounds

3. Calculation of Effect Size Test summary

Metric	Dmin	P hat	Difference (observed)	Standard Error	Margin Error for 95% CI	Lower bound of CI	Upper bound of CI	Result
Gross Conversion	0,01	0.2086	-0.0206	0.0044	0.0086	-0.0291	-0.012	OK, practically and staticallly significant

								✓
Net Conversion	0,0075	0.1151	-0.0049	0.0034	0.0067	-0.0116	0.0019	KO, NOT practically and staticallly significant ✗

4. Reasoning Effect Size Test

1. Gross Conversion

- Statistical: **--> PASSED, CI does not include zero** ✓
- Practical: $Dmin = 0.01$ not included in lower and upper CI bounds

--> OK, practically and statically significant ✓

2. Net Conversion

- Statistical: **NOT PASSED as zero is included in CI** ✗
- Practical: $Dmin = 0.0075$ included in lower and upper CI bounds

--> KO, NOT practically and statically significant ✗

5. Effect Size Test Summary

We have seen that our Gross Conversion Metric **is very well practically and statistically relevant**. However, the Net Conversion Metric is **NOT practically and statistically**.

This is interesting and surprising as Gross and Net Conversion should correlate with each other.

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

1. Purpose

"The Sign test is a non-parametric test that is used to test whether or not two groups are equally sized." In other words this means we want to check, if in the case of two groups there is a difference that is significant.

Sources:
https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/sign-test/

2. Calculation of Sign Test algorithm

1. Create a clean version of control and experiment data frame without null values
2. We need to join our two dataframe control and experiment group so we get the following rates. Some renaming will have to be done
    - Gross conversion rate (Enrollments/Clicks)
    - Net conversion rate (payments/clicks)
3. On the resulting dataframe we can calculate the daily Difference for Gross/Net Converion rate between control and experiment group:
    - GrossConversionDiff = Gross Conversion Rate Control - Gross Conversion Rate Experiment
    - NetConversionDiff = Net Conversion Rate Control - Net Conversion Rate Experiment
4. Get the number of trials and successes to run a sign and binomial test:
    - Trials = number of samples in new data frame
    - Successes:
        - Number of daily GrossConversionDiff < 0
        - Number of daily NetConversionDiff < 0
5. Use a binomial test to calculate the p-value. Either using https://www.graphpad.com/quickcalcs/binomial1.cfm or using scipy.stats import binom_test

3. Result Sign Test

Metric	**Assumed significance level alpha**	**p-value**	**Result**
Gross Conversion	alpha = 0,05	0.0026	OK, statistically significant alpha 0,05 → Reject H0
Net Conversion	alpha = 0,05	0.6776	KO, statistically NOT significant alpha 0,05 → Fail to reject H0

4. Interpretation

- For the Gross Conversion:
    - The two-tail P value is 0.0026 being smaller (<) than the significance level alpha 0,05. This means that we have **strong evidence to reject null hypothesis H0 and the effect for the gross conversion is statistically relevant**. In other words: the effect between the control and experimental group can be considered as significant and the observed effect is not random.
- For the Net Conversion:
    - "The two-tail P value is 0.6776 being higher (>) than the significance level alpha 0,05. This means that we **have not enough evidence to reject null hypothesis H0 and the effect for the gross conversion is statistically NOT relevant**. In other words: the effect between the control and experimental group can be considered as insignificant and the observed effect is by random chance.

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

General setup:
Udacity carried out by Udacity in order to separate students into a control and experimental group. The idea behind it was that the experimental group was asked about the time they would be able to spend on the courses when clicking the stage “start free trial”. However, the control group did not receive this question.
The hypothesis was that this setup in the experimental group would result in less student frustration as there would be a clearer separation between students spending enough time and potentially not leaving after the free trial and therefore increase payments to Udactiy, whereas the students without enough time would be directed towards free content.

Metrics:
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

1. We first divided our metrics into **invariant metric** being a metric that should not be affected by the changes in the experiment and **evaluation metrics,** which are metrics used to measure the impact or the changes made in the experiment

    We chose the following invariant and evaluation metrics:

    - **Invariant metrics:**
        - Number of cookies
        - Number of clicks:
        - Click-through-probability

- **Evaluation metrics**
    - Gross conversion:
    - Net conversion

2. Secondly, we calculated the analytical standard deviation for our two evaluation metrics: **Gross conversion and Net conversion**

3. After this, we looked at the sizing and used the analytical standard deviation as well as Minimum Detectable Effect and baseline conversion rates to calculate the necessary sample size and total sample size.
   Once we had this we were able to calculate the **number of pages views = Total Sample Size / / Click-through-probability**

4. Bonferroni correction correction: we chose not to use the Bonferroni correction for the following reasons

- usually the Bonferroni correction is used to adjust p-values in order to control the family-wise error rate (FWER) in multiple hypothesis testing or in other words the purpose is to reduce the probability of a type I error (false-positive) when making a larger number of hypothesis tests. In this case we are only a control and experimental group with 23 observations. **Hence, the number of tests is fairly small and I will not use the Bonferroni correction.**

Besides, the Bonferroni correction is known for some quite heavy assumptions:

- Conservative approach resulting in quite low adjusted significance levels, which can make it quite difficult to detect true effects
- Independence assumption: Bonferroni correction assumes tests to be conducted independent of each other. This is not necessarily the case for Gross and Net Conversion
- Type II error rate: Bonferroni correction can indirectly affect the Type 2 error rate

5. In the next step we checked the Duration vs. Exposure. Here we found that allocating 100 % of the traffic to the experiment was too risky and decided to go with 50% of the traffic resulting in 35 days to run the experiment.

6. After this, we performed Sanity checks for our three invariant metrics:
    - Number of cookies
    - Number of clicks:.
    - Click-through-probability

   We found that all metrics passed the sanity check as all observed values were between the upper and lower bounds of the CI.

7. After this, we performed the Effect Size Tests to see if a difference between a control and experimental group is also practically significant.

We have seen that our **Gross Conversion Metric is very well practically and statistically relevant. However, the Net Conversion Metric is NOT practically and statistically.**

8. Finally we ran a Sign test, calculated the p-value to check, if we had evidence to reject null hypothesis and the effect was statistically relevant
   We found that:
   - Gross Conversion: The effect between the control and experimental group can **be considered as significant** and the observed effect is not random.

   - Net Conversion: the effect between the control and experimental group **can be considered as insignificant** and the observed effect is by random chance.

9. No discrepancies between the effect size test and the sign test were found

Recommendation

Make a recommendation and briefly describe your reasoning.

To give a proper recommendation, let us first recall the initial goal of the experiment. We wanted to reduce student frustration and decrease the number of students leaving after the free trial to ultimately increase the enrollments and payments.

Our results suggest that we found a **statistically and practically significant decrease in Gross Conversion**, which is good.
However, we did not see any **statistically and practically significant increase of the Net Conversion** being the metrics that we care about as it signifies students staying longer than the free trial and ultimately increasing payments.

Hence, **I recommend not launch the changed version** of the Udacity site but rather perform more experiments to get more clarity.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

To make a suggestion for a proper follow-up experiment, let us focus on our defined goals once more:

- We want to reduce student frustration
- decrease the number of students leaving after the free trial
- Increase the enrollments and payments meaning the net conversion

Ideas: One can think of different ways to do so: improve the content, improve the user experience, enhance visibility on where students stand compared to where they should be and additional benefits.

We want to focus on the additional benefits:

My suggestion is to decrease the number of students leaving after the trial by encouraging them to stay longer once they want to leave after the free trial. This could happen through **discounts** for the first three months compared to the original price. This way, Udacity **might earn less in the short-term, but motivate and persuade students to stay long-term**. To do so, a modified version of the homepage after the free trial would be needed offering the discussed discount when trying to end the free trial.

Experiment design:

- For students that have just started the free trial we would separate them between a control and an experimental group randomly. The experimental group would get to see the discount features, whereas the control group would still get the old homepage

Null Hypothesis:

- The new discount feature **will not increase** the retention of students (remain enrolled past the 14-day boundary)

Alternative Hypothesis:

- The new discount feature **will increase** the retention of students (remain enrolled past the 14-day boundary)

Unit of Diversion:

- As we would need to control the retention between experimental and control group, the user-id seems an appropriate unit of diversion

Invariant Metric:

- The invariant metric is a metric that should not be affected by the changes. Hence, the **number of user-ids** having enrolled in the free-trial could be suitable. In this case a cookie will be necessary in the unit of diversion.

Evaluation Metric:

- Evaluation metrics are metrics used to measure the impact or the changes made in the experiment. Hence, the **Retention** would make sense in order to see if students stay enrolled longer. We could also use the **net conversion** again to check for the improvement of the financial performance.

External resources

- Bonferroni Correction: https://www.investopedia.com/terms/b/bonferroni-test.asp
- Sample Size calculator: https://www.evanmiller.org/ab-testing/sample-size.html
- Effect size test: https://www.scribbr.com/statistics/effect-size/#:~:text=Effect%20size%20tells%20you%20how,size%20indicates%20limited%20practical%20applications
- Sign and binomial test:
  - https://www.graphpad.com/quickcalcs/binomial1.cfm
  - https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.binom_test.html