1. Answer these questions before loading any data.
   1. How large a sample size would be needed to detect an increase in click throughs equal to 10% of the mean, with 90% power and 5% significance? Assume the mean click through rate is 5%.
      1. Sample size per group: 840593.5527752708
      2. Total sample size: 1681187.1055505415
   2. If the typical treatment arm has 3200 observations, what is the minimum detectable effect (not standardized), given 5% significance ?
      1. Minimum detectable effect: 0.08104803053685428
      2. MDE: 0.024314409161056285
   3. If the typical treatment arm has 3200 observations, what is our power to detect an effect equal to 10% of the mean, with 5% significance?
      1. Power to detect an effect equal to 10% of the mean: 0.05459330260010606
   4. Your organization has a website which gets 4,000 views/hour. You are considering running an A/B test of headlines to determine headline effectiveness. For how long would you have to run the test to be able to detect an increase in the click through rate from 5% to 5.5% (with 90% power and 5% significance)?  ("A/B test" means you have 1 control arm and 1 treatment arm.)
      1. Required sample size per group: 840593.5527752729
      2. Total required sample size: 1681187.1055505457
      3. Total hours needed to run the test: 420.29677638763644
2. Use the data for this assignment. In this question, you’ll treat the data as if they come from one giant experiment. Calculate the total number of impression and total number of clicks in each arm of the experiment, and then calculate the overall clickthrough rate in each arm.
   1. What is the difference in means?

A screenshot of a computer

Description automatically generated

* 1. Form a 95% confidence interval for the difference in means. Is the difference statistically significant?
     1. 95% CI for difference in means:
        1. ( -0.0012004499012693817 , -0.0010815149025597787 )
     2. Yes, the difference is statistically significant

[Hints: To make the standard errors, use the formula for the standard deviation of a 0/1 variable, and  treat impressions as the sample size.]

1. Now treat each experiment as a separate observation. For each experiment, calculate the effect of question mark on the outcome. This will give you 6,939 ATEs.
   1. What is the average ATE, averaging over all the experiments?
      1. Average ATE over all experiments: -0.0012118772185663103
   2. How many of the ATEs are statistically significant? (hint: consider making a confidence interval for each experiment)
      1. Number of statistically significant ATEs: 2315
   3. What is the mean ATE among experiments where the ATE is statistically significant?
      1. Mean ATE among statistically significant experiments: -0.003049721899747998
   4. Why is the answer in (c) different than in (b)?
      1. In part c, I calculated ATE only of the statistically significant experiments rather than every single observation like in b by filtering out any ATEs that were not statistically significant. This means that only the ATEs where the confidence intervals do not include zero are considered in the mean calculation.
   5. If you were an organization that ran many experiments and only paid attention to significant ones, would your answers be reliable? Explain briefly.
      1. I believe that you will have a problem if you are only looking at the statistically significant results for every experiment. First of all, statistical significance does not mean practical significance. This means that from getting rid of tons of information because of it not being statistically significant will cause the organization to miss out on small but meaningful effects that could honestly lead to decision-making improvements. This would also lead to a lot of bias because if you are only focusing on significant results that means you’re only going to share these results to the public which may be misleading to customers and give false expectations of the overall effect.
2. Develop a hypothesis about what features of headline generate higher click through rates. Think about an experiment that would let you test it, using the Upworthy experiment database. The key challenge in testing your hypothesis is whether there are experiments that A/B test the feature, and whether you can identify those experiments. For example, we can tell if there is a question in a headline because the data has a headline column, and we can see if it contains a ‘?’. Make a pre-analysis plan at aspredicted.org and turn in your pre-analysis plan with your write-up.
   1. Hypothesis: Headlines that contain a question will result in a higher click-through rate compared to those that do not contain a question.