1. (3 pts) Suppose we have discovered an association between two variables in a dataset. Which of the following would be the BEST way to test whether it is causal? Choose ONE.

   ◯ Use hypothesis testing to check whether the association is statistically significant.

   ◯ Run a randomized controlled experiment.

   ◯ Brainstorm some potential confounding factors and test whether any of them has an association with both variables.

   ◯ If both variables are numerical, use a scatter plot to check for a trend.

2. (3 pts) Suppose you perform a randomized controlled experiment where people in the treatment group take vitamin C regularly and people in the control group do not take vitamin C. In each group you measure the proportion that get sick over the next year. Which of the following statements is correct (choose one).

   ☐ If the control group has a higher proportion of sickness, then it is reasonable to conclude that vitamin C causes a reduction in the chance of getting sick.

   ☐ If the control group has a higher proportion of sickness and a hypothesis test finds that this difference is statistically significant, then it is reasonable to conclude that there is an association between taking vitamin C and not getting sick but we can't reasonably conclude that vitamin C causes a reduction in the chance of getting sick.

   ☐ If the control group has a higher proportion of sickness and a hypothesis test finds that this difference is statistically significant, then it is reasonable to conclude that there is an association between taking vitamin C and not getting sick AND furthermore that vitamin C causes a reduction in the chance of getting sick.

3. Each person in a random sample of 1000 U.S. adults was asked if they agreed with the statement, "News organizations are growing in influence." Among the sampled men, 39% agreed. Among the sampled women, 43% agreed. Data scientists decide to use an A/B test with a significance level of 1% to see whether or not the observed difference is due to chance.

   (a) (3 pts) The null hypothesis is one of the statements below. Pick the right one.

   ◯ In the sample, the percent of women who agree is the same as the percent of men who agree. The observed difference is due to chance.

   ◯ In the U.S., 39% of the men agree and 43% of the women agree, due to chance.

   ◯ In the U.S., the percent of men who agree is the same as the percent of women who agree. The difference in the sample is due to chance.

   ◯ In the U.S., the percent of women who agree is different from the percent of men who agree, due to chance.

   (b) (3 pts) If the null hypothesis is true, what is the probability that their hypothesis test will **fail to reject** the null hypothesis?

   (c) (2 pts) Which of the following test statistics is best for choosing between the null and alternative hypotheses in this situation? (Choose one).

   ◯ The absolute difference between the proportion of men who agree and the proportion of women who agree.

   ◯ The absolute difference between the proportion of men who agree and 0.5

   ◯ The absolute difference between the proportion of women who agree and 0.5

   (d) (3 pts) Which of the following is the best method for simulating the null hypothesis?

   ◯ Simulate a Binomial distribution with n=1000 and p=0.5 and calculate the test statistic on this distribution.

   ◯ Repeat the following process: Randomly sample 390 responses from the original sample and calculate the test statistic on those responses.

   ◯ Repeat the following process: randomly shuffle the gender labels on the sample data and recalculate the test statistic using the shuffled labels.

   (e) (3 pts) Suppose they run the hypothesis test and the empirical p-value comes out to be 0.005, that is, 1 in 200. Which of the following statements are true? Select ALL that apply.

   ☐ The data scientists will fail to reject the null hypothesis.

   ☐ There is only a 1 in 200 chance that the null hypothesis is true.

   ☐ There is a 199 in 200 chance that the alternative hypothesis is true.

   ☐ 0.005 of the simulated values of the test statistic were greater than or equal to the observed value.

   ☐ The data scientists have proven the null hypothesis is true.

   ☐ The data scientists can reject the null hypothesis.

   ☐ None of the above statements is true.