Assignment 3

Note: Unless otherwise specified, data sets come from Wooldridge's econometrics textbook, *Introductory Economics*. To use them, use "library(wooldridge)".

Note: To ensure that you and your group members have replicable results, make sure to use the `set.seed()` command at the start of your file. If you want to have the same answers as in my solution set, use 5772 as the seed.

Problem 1: Confidence Intervals. Suppose we sample *n* data points from a distribution $N(\theta, 36)$, where the value of the central mean θ is unknown.

a. Suppose that $n = 100$ and the sample mean is estimated to be $\bar{x} = 25$. What is the 90% confidence interval for θ? The 95%? The 99%?
b. Repeat the exercise for $n = 1{,}000$. Why are the confidence intervals always narrower?
c. Create a dataset comprised of 1,000 observations drawn from the $N(25{,}36)$ distribution. Use that data set to identify a 90% confidence interval. Then repeat this 100 times. How many of those confidence intervals contain the sample mean of 25? What principle of a confidence interval is this exercise highlighting?

Problem 2: Performing Simple Hypothesis Tests. This problem asks you to perform simple hypothesis tests for sample means and populations. For each test, make sure to (i) state the null and alternative hypotheses and the chosen level of significance, (ii) define the test statistic, (iii) calculate the value of the realized statistic with its corresponding *p*-value, and (iv) decide whether or not to reject the null hypothesis. All of this should be explained clearly in the context of the problem.

a. Use the "wooldridge::rdchem" data for this problem. This data contains information on 32 firms in the chemical industry. First, consider the average profit rate among these firms, captured in dollars with the *profits* variable. The data provides information on sales, captured by *sales,* as well as the percentage of sales converted to profits (*profmarg).* Write and perform a test of the hypothesis that the average of the *profits* variable is equivalent to the average of *sales* times the average of *profmarg* (in decimals). What are you testing here? What do you conclude with $\alpha = 0.05$?
b. What is the 95% confidence interval for the <u>proportion</u> of firms spending over 6% of sales on R&D (*rdintens)*? Is this fraction statistically different from zero? What does this mean, and why might you be observing this?
c. Test the correlation between rdintens and profits. What do you conclude?
d. *Extra credit:* How do your results for (c) change if you use the logarithm of R&D spending (*lrd*)? What does this mean in the context of the problem?

Problem 3: Testing a difference in means. Frequently, we care about whether two groups have the same mean in a given outcome (this is the entire basis of estimating the effect of a treatment on a group relative to a control!).

a. Suppose that we are trying to evaluate the effect of job market training programs on wages. Use the "wooldridge::jtrain98" dataset to assess this question. Approximately 7% of employees in this sample received job market training in 1998 (*train*). First, test if the trained group had significantly different earnings in 1996 (*earn96*) than the nontrained group. Interpret your results in the context of a researc~~h~~ setting. What does this show? I~~s~~ it something we hope would be true, and why?

b. Now test the difference in 1998 wages (*earn98*) across groups. Are the results surprising to you? Would you argue that they are economically meaningful? Defend your answers.
c. Is the test in (b) sufficient to argue that wages increased because of job training? Why or why not?
d. *Extra credit:* Provide supporting evidence for your answer to (c) by examining demographic variables included in the dataset. What do you find and why might this influence the causal relationship between job training and wages? Is there a way to design a research question that avoids this/these problem(s)?