Assignment 3

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

Problem 1: Confidence Intervals in R. 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?

Problem 2: Performing 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 "rdchem" data for 32 firms in the chemical industry, and consider the relationship between expenditures on research and development (R&D) captured in the variable *rdintens* (this is measured as a percentage of sales) and sales, captured by *sales*. Split the firms into those with above-average and below-average R&D spending. Are the sales significantly different? Interpret your findings in the context of a research statement.

b. What is the 95% confidence interval for the proportion of firms spending over 6% of sales on R&D? Is this fraction statistically different from zero? What does this mean, and why might you be observing this?

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!). This problem will help extend the testing framework to that problem.

a. Consider two groups $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$. We suppose that $X_i \sim_{\text{i.i.d.}} f(\mu_1, \sigma_1)$ and $Y_i \sim_{\text{i.i.d.}} f(\mu_2, \sigma_2)$. Additionally, we suppose that X and Y are independent samples.

We are trying to estimate the difference in means, $\mu_1 - \mu_2$. What sample estimator should we use? (You can use the method of moments in a pinch, but trust your gut.)

- Extra credit: Prove that your estimator is unbiased (so that its expected value is the true difference in means), and that it has a standard deviation of

$$\sigma_T = \sqrt{\left\{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right\}}$$

b. Now use your estimator to write a test statistic for the following test:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = \Delta_0 \\ H_1 &: \mu_1 - \mu_2 > \Delta_0 \\ \alpha &= \alpha_0. \end{aligned}$$

Recall that the general version of a test statistic is

$$T = \frac{\text{estimator} - \text{assumed value in } H_0}{\text{s.d. of your estimator}}$$

This test statistic follows the same rules as other statistics—if we know σ^2 or have a large enough sample, $T \sim N(0, 1)$. If instead, we use our estimate s^2, the test statistic uses a t distribution (whose degrees of freedom is a function of n and m).

c. Suppose that we are trying to evaluate the effect of job market training programs on wages. Use the “jtrain98” dataset to assess this question. Approximately 7% of employees in this sample received job market training in 1998 (the “train” variable). 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 research setting. What does this show? Is it something we hope would be true, and why?

d. 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.

e. Is this sufficient to argue that wages increased because of job training? Why or why not? To prove your point, select one or more other demographic variables included in the dataset and examine how these are different across the treated/control groups.