Public Policy 529
Fall 2023 Problem Set #5

Due on Wednesday, October 11, end of day

For this problem set, I will start promoting the use of Markdown in both Stata and R, but especially for R users. It is not necessary for you to use Markdown, so do not feel the need to do so. In fact, it is a bit clunky to set it up in Stata, especially on a university lab machine, so feel free to just paste your output into your answers like you have been doing. While it requires a bit of learning, I think the results are worth it in the long run.

I have posted Markdown template files, a help guide, and demonstration videos in Canvas. After you have completed the setup, the easiest way to proceed is to start with the problem set template and insert what you need in that document.

1. Explain how each of these things affect the size (i.e. width) of the confidence interval?

    (a) The desired level of confidence.

    (b) The sample standard deviation (s).

    (c) The sample size (n).

2. Find the following using the t-table handout.

    (a) If $n = 31$, what critical value of t would you would use to make a 95% confidence interval for a sample mean?

    (b) How much area is there in the upper tail of the t-distribution for a t-statistic of 2.799 when degrees of freedom are 26?

    (c) If there are 60 degrees of freedom and your t-statistic is -2.42, how much area is there in the lower tail of the t-distribution? A range is okay.

3. The Millennium Development Goals aimed to improve human welfare by setting goals for improvement in a range of human development indicators. One goal was to increase the level of education for the people in a country. To test progress, officials take a survey with a sample of 650 and find that the mean years of education is 8.2 with a standard deviation (s) of 5.5.

(a) Construct a 95% confidence interval around the sample mean and interpret the result.

(b) Now, construct a 90% confidence interval and interpret the result.

(c) Suppose that, one decade earlier, the country performed a full national census and found that the mean years of education was 7.9 years. Drawing upon your answers above for insight, how plausible is it that 7.9 years is still the mean of the population?

4. Suppose that the Environmental Protection Agency tests a random sample of 110 cars from a particular manufacturer and found that a proportion of .15 failed to meet emission standards.

(a) Based upon this study, construct a 90% confidence interval associated with the estimated proportion of .15.

(b) Interpret this confidence interval properly.

(c) At $\alpha = .05$, perform a significance test versus the null hypotheses that the true proportion of the cars that fail emission standards is .06. Be sure to follow all the steps: consider assumptions (i.e. is your sample size adequate, what is the appropriate test statistic, etc.), state hypotheses, identify the critical value of the test statistic, calculate the value of the test statistic, make a decision on whether to accept or reject the null hypothesis, and calculate the p-value.

5. Use the anes2020subset dataset for the following questions. The variable BAplus asked respondents whether they have a college degree: yes or no.

(a) What proportion of respondents responded "yes"? Note: get the frequency distribution and use the information to calculate the proportion.

(b) Using the formulas to perform the calculations yourself, construct a 95% confidence interval for this estimated proportion. Note: you will need to work this problem using at least three decimal places.

(c) Interpret this confidence interval.

(d) Describe two reasons why this interval might not contain the true proportion of people in the population that have a college degree.

(e) Now use your software to make a 90% confidence interval. In Stata, the command is `ci prop BAplus, level(90)`. In R, the command takes the following form: `prop.test(#cat, n, conf.level = .90, correct = FALSE)`. In place of "#cat," which stands for number in the category, you should put the number of people who say they have a college degree. In place of "n," you should put the total number of cases.

Interpret this confidence interval properly.

6. In the 2020 National Election Study (anes2020subset), "thermometer scores" are used to measure how much the respondent likes or dislikes a person, group, or organization. The scale goes from 0 to 100. The respondent is told that 50 degrees on the scale means neutral feelings, while higher scores mean warmer feelings and lower scores mean colder feelings. Treat this scale as an interval-level variable.

    (a) Find the mean thermometer score for labor unions (UnionsTherm), the sample size, and the standard deviation of this variable.

    (b) With the information you received from your command, use the formulas to construct a 95% confidence interval for this score. Do this yourself using the formulas and show your work.

    (c) Now use your software to make a 99% confidence interval for UnionsTherm. In Stata, the command is: ci mean UnionsTherm, level(99)

    In R, the command is: t.test(anes2020$UnionsTherm, conf.level = .99)

    Interpret the resulting confidence interval properly.