Public Policy 529
Fall 2023 Problem Set #5

Francisco Brady

2023-10-11

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. Using a higher of confidence level results in a wider confidence interval. This is because the degree of confidence used is translated into a Z score which is used to construct the interval. For example, for a 95% level of confidence, the z score used is 1.96, which is then multiplied by the standard error of the point estimate to construct the interval. For a 99% level of confidence, the z score used is 2.576, which creates a wider interval when multiplied by the standard error.

(b) The sample standard deviation (s). As the standard deviation increases, confidence interval increases. This is because the standard error is calculated using the standard deviation in the numerator $\frac{\sigma_{\bar{x}}}{\sqrt{n}}$. This is then multiplied by the z statistic in order to create a confidence interval.

(c) The sample size (n). Increasing the sample size of the estimate will decrease the width of the confidence interval. This is because larger sample sizes decrease the size of the standard error ($\frac{\sigma}{\sqrt{n}}$), which is used to construct the confidence interval.

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?
Degrees of freedom: n - 1 = 30.
Critical value of t: 2.042

(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?
The area in the upper tail would be between .005 and .001.

(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.
Between .005 and .010

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.
Since the sample is larger than 30 observations, and we are dealing with means, we can use a t statistic ($t_{.025} = 1.96$), and use that to construct the interval:

$$CI = \bar{y} \pm t \cdot \frac{s}{\sqrt{n}} = 8.2 \pm 1.96 \cdot \frac{5.5}{\sqrt{650}}$$

```r
moe <- 1.96 * (5.5 / sqrt(650))
CI <- round(c(8.2 - moe, 8.2 + moe),3)
cat(c("lower:",CI[1], "\nupper:", CI[2]))
```

```
## lower: 7.777
## upper: 8.623
```

The interpretation is: Following these procedures, if we take repeated random samples of $n = 650$ and calculate the upper and lower bound of the estimates each time, 95% of the time it would capture the true population mean years of education. The confidence interval in this instance is between 7.777 and 8.623.

(b) Now, construct a 90% confidence interval and interpret the result.
For a 90% confidence interval, we can use a t statistic of 1.645 ($t_{.050}$). Following the same calculations as above:

$$CI = \bar{y} \pm t \cdot \frac{s}{\sqrt{n}} = 8.2 \pm 1.645 \cdot \frac{5.5}{\sqrt{650}}$$

```r
moe <- 1.645 * (5.5 / sqrt(650))
CI <- round(c(8.2 - moe, 8.2 + moe),3)
cat(c("lower:",CI[1], "\nupper:", CI[2]))
```

```
## lower: 7.845
## upper: 8.555
```

Interpretation: Following these procedures, if we took repeated random samples of $n = 650$ and calculate the upper and lower bound of the estimates each time, 90% of the time it would capture the true population mean years of education. The confidence interval in this case is between 7.845 and 8.555.

(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?
Since the mean from the national census is within both the 90% confidence interval and the 95% confidence for the (new) sample estimate, it is plausible that the population sample mean is still 7.9 years of education.

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.

To construct a confidence interval for proportions, we need to calculate the standard error $\hat{\sigma\pi} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$, which we then use in combination with the Z-score, to construct the intervals. At the 90% confidence level, $z_{.05} = 1.645$.

$$CI = \hat{\pi} \pm z \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.15 \pm z \cdot \sqrt{\frac{0.15 \cdot (1-0.15)}{110}}$$

Provided that there are at least 15 cases in each category (failing standards and not failing standards). $110 \cdot 0.15 = 16.5$, so that is true. Using the z table to find a score, we can use 1.29 (.985). Plugging into the standard error formula:

```r
pi_hat <- .15
n <- 110
z <- 1.29
standard_error <- sqrt((pi_hat * (1 - pi_hat)) / n)
CI <- round(c(pi_hat - (z * standard_error), pi_hat + (z * standard_error)),3)
cat(c("lower:",CI[1], "\nupper:", CI[2]))
```

```
## lower: 0.106
## upper: 0.194
```

So the confidence interval is 0.106 and 0.194.

(b) Interpret this confidence interval properly.
Following these procedures, if we took repeated random samples of $n = 110$ and calculate the upper and lower bound of the estimate, 90% of the time it would capture the true proportion of cars that failed to meet the emissions standard. The confidence interval in this case is between 0.106 and 0.194

(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.

i. Assumptions: We assume that the sample is random. We also assume that the sample is approximately normally distributed and that since the sample is larger than 30, the central limit theorem applies. We are dealing with proportions, so we should use a z-statistic for our significance test.

ii. Hypothesis:
$H_0 : \pi = .06$
$H_a : \pi \neq .06$

iii. Calculate test statistic: Using the formula for the z-statistic.

$$z = \frac{\hat{\pi} - \pi_0}{se} = \frac{.15 - .06}{\sqrt{\frac{.06(1-.06)}{110}}} = \frac{0.09}{0.02264348} = 3.974654$$

For $\alpha = .05$, the critical value of the z statistic is 1.96.

iv. Calculate a p-value: Our z statistic is ≈ 3.97, which is higher than the critical value. The right hand tail for z $= 3.97$ is $\approx .0000317$, this makes the p-value: $p = 2 \cdot .0000317 = .0000634$.

v. Conclude: Since $p < \alpha$, we can reject the null hypothesis.

```r
load('anes2020subset.RData')
```

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

```r
table(anes2020$BAplus, useNA = 'a')
```

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

```
##
## Does not have BA           Has BA             <NA>
##             4502             3647              131
```

```r
# NOTE: Not using NA values to calculate proportion
denom <- (4502 + 3647)
num <- (3647)
prop <- num / denom
prop
```

```
## [1] 0.4475396
```

```r
#  code for question 5.b
pi_hat <- prop
n <- denom
z <- 1.96
standard_error <- sqrt((pi_hat * (1 - pi_hat)) / n)
CI <- round(c(pi_hat - (z * standard_error), pi_hat + (z * standard_error)),3)
cat(c("lower:",CI[1], "\nupper:", CI[2]))
```

(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.

```
## lower: 0.437
## upper: 0.458
```

(c) Interpret this confidence interval.

If we took repeated random samples of $n = 8149$ and calculated the upper and lower bound of the estimates, 95% of the time it would capture the true population proportion of people with a BA. The confidence interval in this case is between 0.437 and 0.458.

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

1. It is possible that this sample is one of the 5% of random samples in the sampling distribution that does not capture the true proportion.
2. There could be issues with the data. In this variable, there were 131 records whose response to this question was NA. It's also possible this is weighted data, and we may not be using the data correctly.

```r
# code for 5.e
prop.test(3647, 8149, conf.level = .90, correct = FALSE)
```

(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.

```
## 
##  1-sample proportions test without continuity correction
## 
## data:  3647 out of 8149, null probability 0.5
## X-squared = 89.707, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.4384982 0.4566158
## sample estimates:
##         p 
## 0.4475396
```

If we took repeated random samples of $n = 8149$ and calculate the upper and lower bound of the estimates each time, 95% of the time it would capture the true population proportion of people with a BA. The confidence interval in this case is between 0.438 and 0.457.

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.

```r
# code for 6.a
cat('mean:', mean(anes2020$UnionsTherm, na.rm = T))
```

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

```
## mean: 58.33511
```

```r
n <- length(na.omit(anes2020$UnionsTherm))
cat('n:', n)
```

```
## n: 7326
```

```r
cat('sd:', sd(anes2020$UnionsTherm, na.rm = T))
```

```
## sd: 24.0318
```

```r
# code for 6.b
moe <- 1.96 * (24.0318 / sqrt(7326))
CI <- round(c(58.33511 - moe, 58.33511 + moe),3)
cat(c("lower:",CI[1], "\nupper:", CI[2]))
```

(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.

```
## lower: 57.785
## upper: 58.885
```

```r
# code for 6.c
t.test(anes2020$UnionsTherm, conf.level = .99)
```

(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.

```
##
##  One Sample t-test
##
## data:  anes2020$UnionsTherm
## t = 207.77, df = 7325, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  57.61170 59.05852
## sample estimates:
## mean of x
##  58.33511
```

Across repeated random samples of $n = 7326$, if we calculated the upper and lower bound of the estimates, 99% of the time it would capture the true population mean of respondents sentiment on labor unions. The confidence interval in this case is between 57.612 and 59.059.