Problem set 6: CLT and confidence intervals

Due November 6, 2023, at 9pm

(Your name here)

NOTE: Start with the file ***ps6_2023_more_learning_from_samples.qmd*** *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F23/tree/main/assignments). Modify that file to include your answers. Make sure you can "render" the file (e.g. in RStudio by clicking on the* ***Render*** *button). Submit both the qmd file and the PDF via Canvas.*

Question 0: Loading the data

You will find the dataset `cces_2012_subset.csv` in the `data` directory of the github. Download that dataset and load it into memory e.g. using `read.csv()`.

The dataset contains a subset of the variables from the 2012 Cooperative Congressional Election Survey (CCES), including two variables you analyzed in problem set 5 (`aa` relating to affirmative action and `env` relating to environmental/jobs tradeoffs). It also contains variables on gender, education, and race of respondents.

`gender` is 1 for men and 2 for women.

```
# your code here
```

Question 1: CLT for a regression slope

1a) As you did last week, use `lm()` to regress the affirmative action response (`aa` in this dataset, dependent variable) on the environment/jobs response (`env` in this dataset, independent variable). Report the regression coefficients from the regression output using the `coef()` function. (If you store the regression output in a variable called `my_lm`, then `coef(my_lm)` returns the regression coefficients.)

(1b) Using a for-loop, take 5000 samples of 500 rows from the dataset. In each sample, compute the regression slope coefficient and store it. Report the mean and standard deviation of your 1000 slope coefficients, and plot a histogram of the results. (HINT: before writing your for-loop,

make sure you can get a single sample of 500 rows, run a regression in this sample, and extract the coefficient; then embed your code in a for-loop.)

(1c) If the sampling distribution you obtained in the previous sub-question were perfectly normal, with mean and standard deviation equal to the mean and standard deviation of your coefficient estimates, what proportion of estimates would be above .4? What proportion of coefficient estimates are actually above .4?

(1d) Carry out the same exercise as in 1b, except now sample only 10 rows.

(1e) If the sampling distribution you obtained in the previous sub-question were perfectly normal, with mean and variance equal to the mean and variance of your coefficient estimates, what proportion of estimates would be above .6? What proportion of coefficient estimates actually are above .6?

(1f) What does the above exercise tell you (if anything) about the validity of approximating the sampling distribution of regression coefficients with a normal distribution?

Question 2: confidence intervals

(2a) Compute the average response to `env` (i.e. preference for protecting jobs over environment) separately for men and women in the dataset.

(2b) Compute a valid 90% confidence interval for the average response to `env` among women in the US population. Does the confidence interval include the average response for men in the sample?

(2c) Suppose again that the dataset contains the whole population. You will run a simulation to confirm that you can produce a 90% confidence interval for the population mean that has the correct coverage. Specifically, take 10000 samples of 200 women from the data, generate a 90% confidence interval in each sample for the average response to `env` among women in the population, and confirm that around 90% of the confidence intervals contain the estimand. Include comments in your code to explain each step.