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.

```r
dat <- read.csv("cces_2012_subset.csv")
```

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.)

Answer:

```r
reg <- lm(aa ~ env, data = dat)
coef(reg)
```

```
(Intercept)         env 
  1.7728994   0.3678295 
```

As we saw in the last problem set, respondents who prioritize jobs more (higher values of `env`) also tend to be more opposed to affirmative action.

(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.)

Answer:

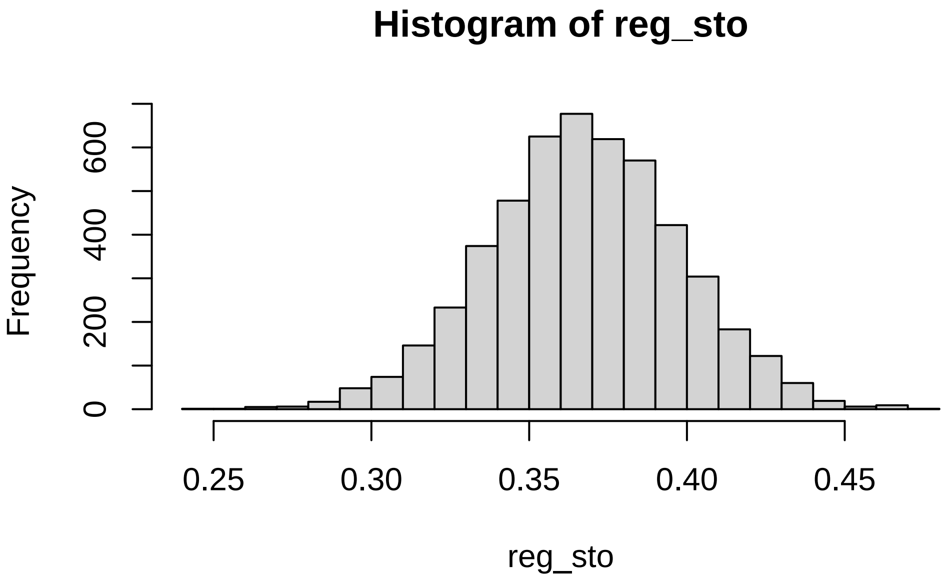
```r
n <- 500
m <- 5000
reg_sto <- rep(NA, m)
for(i in 1:m){
  this_dat <- dat[sample(1:nrow(dat), size = n), ]
  reg_sto[i] <- coef(lm(aa ~ env, data = this_dat))[2]
}
mean(reg_sto)
```

```
[1] 0.3674154
```

```r
sd(reg_sto)
```

```
[1] 0.03030196
```

```r
hist(reg_sto, breaks = 20)
```



(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?

Answer

If it were perfectly normal, then the proportion of draws above .4 would be 1 minus the CDF at .4 for a normal distribution centered at `mean(reg_sto)` with standard deviation `sd(reg_sto)`. In `R`, we can compute this as

```
1 - pnorm(.4, mean = mean(reg_sto), sd = sd(reg_sto))
```

```
[1] 0.1411137
```

The actual proportion of draws above .4 is

```
mean(reg_sto > .4)
```

```
[1] 0.1408
```

which is very close.

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

Answer:

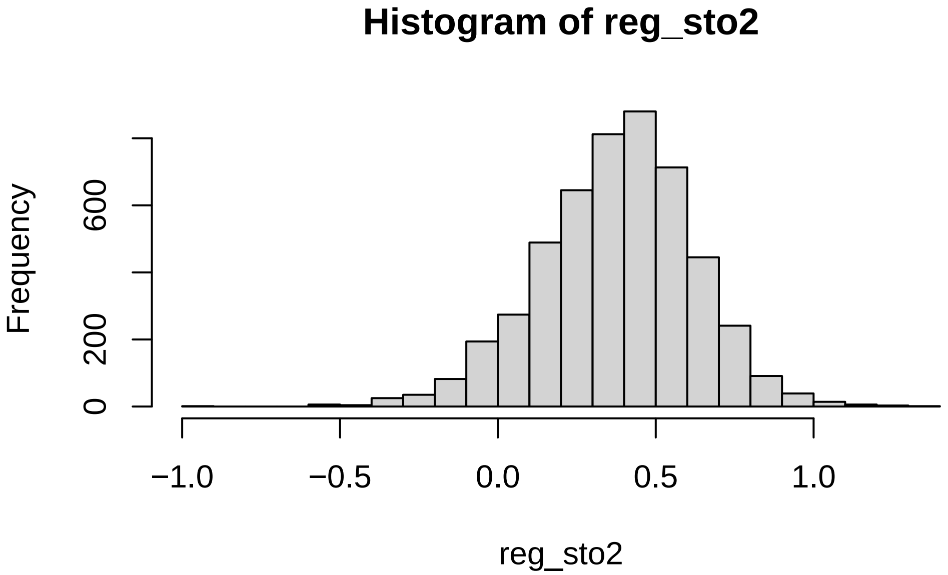
```
n <- 10
m <- 5000
reg_sto2 <- rep(NA, m)
for(i in 1:m){
  this_dat <- dat[sample(1:nrow(dat), size = n), ]
  reg_sto2[i] <- coef(lm(aa ~ env, data = this_dat))[2]
}
mean(reg_sto2)
```

```
[1] 0.3792559
```

```
sd(reg_sto2)
```

```
[1] 0.2432753
```

```
hist(reg_sto2, breaks = 20)
```



(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?

Answer

If it were perfectly normal, then the proportion of draws above .6 would be 1 minus the CDF at .6 for a normal distribution centered at `mean(reg_sto2)` with variance `var(reg_sto2)`. In `R`, we can compute this as

```
1 - pnorm(.6, mean = mean(reg_sto2), sd = sd(reg_sto2))
```

```
[1] 0.182102
```

The actual proportion of draws above .6 is

```
mean(reg_sto2 > .6)
```

```
[1] 0.1692
```

which is close, though not as close.

(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?

Answer

It suggests that the sampling distribution of regression coefficients can be pretty well approximated by a normal distribution even with small samples, at least in some cases. The approximation may be less good if the distribution of the variables is very skewed (e.g. with some very large or very small values in the population); in such cases a larger sample may be necessary for the sampling distribution to be approximately normal.

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.

Answer

```
# average for women
mean(dat$env[dat$gender == 2], na.rm = T)
```

```
[1] 3.156821
```

```r
# average for men
mean(dat$env[dat$gender == 1], na.rm = T)
```

```
[1] 3.227057
```

(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?

Answer: A valid 90% confidence interval for a sample mean $\hat{\theta}$ is $[\hat{\theta} - 1.64\hat{\sigma}, \hat{\theta} + 1.64\hat{\sigma}]$, where $\hat{\sigma}$ is the estimated standard deviation of the sample mean.

Our sample mean $\hat{\theta}$ is just the average response among women, given above.

Our unbiased estimate of the standard deviation in the population is the sample standard deviation:

```r
std_dev <- sd(dat$env[dat$gender == 2], na.rm = T)
```

We divide that by $\sqrt{n}$ to get the standard error, i.e. the standard deviation of the estimator:

```r
hat_sigma <- std_dev/sqrt(sum(dat$gender == 2 & !is.na(dat$env)))
```

Note that our n here is the number of valid (non-NA) responses from women, which can be written as `sum(dat$gender == 2 & !is.na(dat$env))`.

So our 90% confidence interval is:

```r
mean(dat$env[dat$gender == 2], na.rm = T) + 1.64*hat_sigma*c(-1, 1)
```

```
[1] 3.121727 3.191915
```

The interval does not include the average response for men, which was 3.23.

(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 from each sample, and confirm that around 90% of the confidence intervals contain the estimand. Include comments in your code to explain each step.

Answer:

```r
m <- 10000 # number of samples
n <- 200 # size of samples
ci_lower <- ci_upper <- rep(NA, length = m) # storage
# the data from which we will sample
dat_env_women <- dat$env[dat$gender == 2 & !is.na(dat$env)]
for(i in 1:m){
  samp <- sample(dat_env_women, size = n) # take a sample
  samp_mean <- mean(samp) # sample mean
  sd_of_samp_mean <- sd(samp)/sqrt(200) # estimated sd of sample mean
  ci_lower[i] <- samp_mean - 1.64*sd_of_samp_mean # store lower
  ci_upper[i] <- samp_mean + 1.64*sd_of_samp_mean # store upper
}
estimand <- mean(dat_env_women) # our target
contains <- estimand > ci_lower & estimand < ci_upper
mean(contains) # coverage
```

```
[1] 0.9064
```

The coverage is close to 90%, so the confidence interval appears to be valid.