Problem set 5: Learning from samples

Due November 4, 2024, at 10am

(Your name here)

NOTE: Start with the file `ps5_2024_learning_from_samples.qmd` (available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F24/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 1: Estimating the sample mean (theory)

(1a) Below is the proof of Theorem 3.2.4, "Sampling Variance of the Sample Mean", from Aronow & Miller:

$$
\begin{aligned}
\mathrm{V}\left[\overline{X}\right] &= \mathrm{V}\left[\frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)\right] && \text{(Step 1)}\\
&= \frac{1}{n^2}\mathrm{V}\left[X_1 + X_2 + \ldots + X_n\right] && \text{(Step 2)}\\
&= \frac{1}{n^2}\left(\mathrm{V}\left[X_1\right] + \mathrm{V}\left[X_2\right] + \ldots + \mathrm{V}\left[X_n\right]\right) && \text{(Step 3)}\\
&= \frac{1}{n^2}\left(\mathrm{V}\left[X\right] + \mathrm{V}\left[X\right] + \ldots + \mathrm{V}\left[X\right]\right) && \text{(Step 4)}\\
&= \frac{1}{n^2}n\mathrm{V}\left[X\right] && \text{(Step 5)}\\
&= \frac{\mathrm{V}\left[X\right]}{n} && \text{(Step 6)}
\end{aligned}
$$

Explain what property/definition/operation justifies each step in the proof.

Answer:

- Step 1: Definition of sample mean
- Step 2: Properties of variance (theorem 2.1.14 in A&M)
- Step 3: Mutual independence of $X_1, X_2, \ldots, X_n$ (part of iid)
- Step 4: $X_1, X_2, \ldots, X_n$ identically distributed (part of iid)

- Step 5: Simplifying (we have n $\mathrm{V}[X]$s)
- Step 6: Simplifying again ($n/n = 1$)

Suppose the n iid random variables $X_1, X_2, \ldots, X_n$ represent responses on a public opinion survey in a large country. Specifically, each variable is a randomly sampled citizen's response to a survey question in which the citizen was asked to give their satisfaction with government on a numerical scale where 0 is "totally dissatisfied", 100 is "totally satisfied", and 50 is "neither satisfied nor dissatisfied".

(1b) In words, what is $\overline{X}$ in this case?

Answer: It is the sample mean, or the average numerical satisfaction with government among the n survey respondents.

(1c) In words, what does $E[\overline{X}] = E[X]$ mean in this case?

Answer: It means that expected sample mean of the satisfaction scores, i.e. the average of the sample means we would get if we took many such sample means, is the same as the average satisfaction in the population.

(1d) In words, what does $\mathrm{V}[\overline{X}] = \frac{\mathrm{V}[X]}{n}$ mean in this case?

Answer: It means that sampling variance of the sample mean of the satisfaction scores, i.e. the variance of the sample means we would get if we took many such sample means, is the same as the variance in satisfaction scores in the population divided by n, the size of the sample.

Question 2: Estimating the sample mean (simulation)

By running this code, you will load two variables (`aa_2012` and `env_2012`) into `R`'s memory and set a seed so that we get the same results from simulations:

```
# make sure to run this code to get the data!
data_location <- "https://github.com/UChicago-pol-methods/IntroQSS-F24/raw/main/data/"
load(url(paste0(data_location, "CCES_variables_2012.RData")))
set.seed(60637)
```

The 2012 Cooperative Congressional Election Survey asked respondents,

> "Affirmative action programs give preference to racial minorities in employment and college admissions in order to correct for past discrimination. Do you support or oppose affirmative action?"

Response options were

- 1 Strongly support
- 2 Somewhat support
- 3 Somewhat oppose
- 4 Strongly oppose

The variable `aa_2012` contains responses to this question.

The 2012 Cooperative Congressional Election Survey also asked respondents,

> "Some people think it is important to protect the environment even if it costs some jobs or otherwise reduces our standard of living. Other people think that protecting the environment is not as important as maintaining jobs and our standard of living. Which is closer to the way you feel, or haven't you thought much about this?"

Response options were

- 1 Much more important to protect environment even if lose jobs
- 2 Environment somewhat more important
- 3 About the same
- 4 Economy somewhat more important
- 5 Much more important to protect jobs, even if environment worse

The variable `env_2012` contains responses to this question.

To start with, assume that the data contains the whole population of interest.

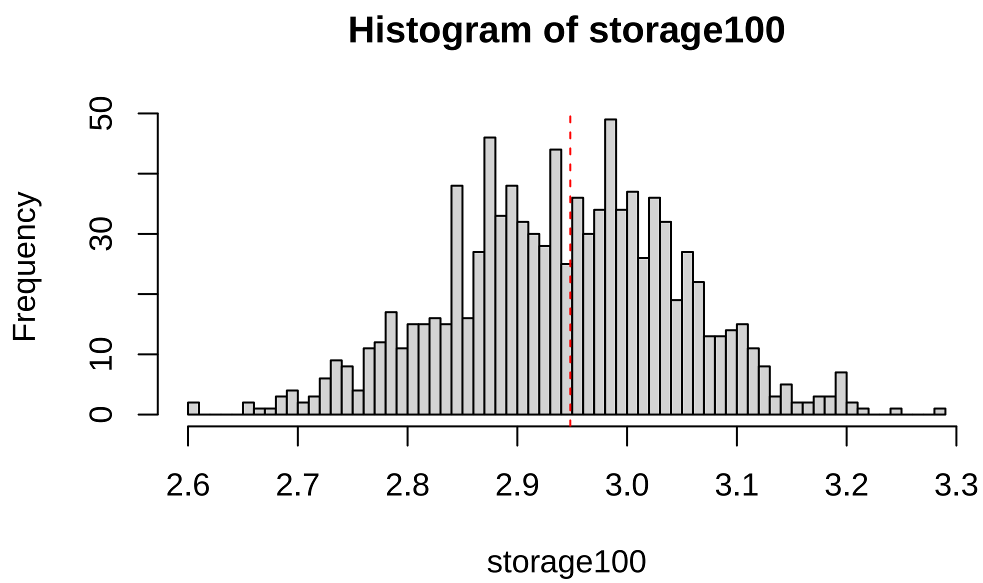
(2a) Write code to draw a single sample of 100 responses to `aa_2012` (without replacement) and compute average opposition to affirmative action (on the 1-4 scale in the raw data) in this sample.

```
# your code here
mean(sample(aa_2012, size = 100))
```

```
[1] 3.14
```

(2b) Use a for-loop to do the same thing 1000 times and store all of the results. That is, compute the sample mean 1000 times, each time drawing a different sample of 100 respondents (without replacement) and computing the mean, and store each of these sample means. Use `hist()` to make a histogram of the results, and use `abline()` to add a vertical line at the population mean.

```r
m <- 1000
storage100 <- rep(NA, m)
for(i in 1:m){
  storage100[i] <- mean(sample(aa_2012, size = 100))
}
hist(storage100, breaks = 80)
abline(v = mean(aa_2012), col = "red", lty = 2)
```



(2c) Use `var()` to compute the variance of your sample means. Compare this to the theoretical value given by Theorem 3.2.4 (which you explicated in question 1).

Answer:

Using `var()`, the variance of my sample means is:

```r
var(storage100)
```

```
[1] 0.01122031
```

The theoretical value can be computed using `R` as

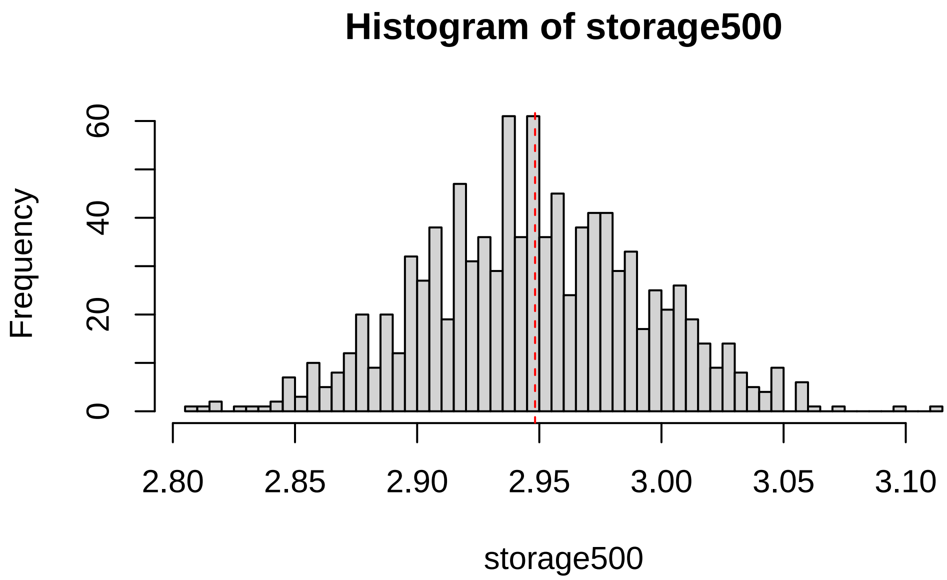
```r
var(aa_2012)/100
```

```
[1] 0.01055306
```

(2d) Repeat (2b), but now each of your samples should be size 500.

Answer:

```
m <- 1000
storage500 <- rep(NA, m)
for(i in 1:m){
  storage500[i] <- mean(sample(aa_2012, size = 500))
}
hist(storage500, breaks = 80)
abline(v = mean(aa_2012), col = "red", lty = 2)
```



(2e) Repeat (2c) for your samples of size 500. That is, use `var()` to compute the variance of your sample means (with samples of size 500). Compare this to the theoretical value given by Theorem 3.2.4 (which you explicated in question 1).

Answer:

Using `var()`, the variance of my sample means is:

```
var(storage500)
```

```
[1] 0.002119886
```

The theoretical value can be computed using `R` as

```
var(aa_2012)/500
```

```
[1] 0.002110613
```

(2f) The 9452 people who answered these two questions on the CCES do not constitute the entire population of interest (the US voting-age population); they are a sample from that population. Using all of the observations as your sample, and supposing this is an iid sample, what is our best guess of average opposition to affirmative action in the population? What is the (estimated) standard error of your estimate?

Answer:

The sample mean, using the entire sample, is

```
mean(aa_2012)
```

```
[1] 2.948265
```

The standard error of the sample mean is the square root of the variance of the sample mean. The variance of the sample mean is $\mathrm{V}[X]/n$, where $\mathrm{V}[X]$ is the variance of the random variable and n is the sample size. Because we don't have the population, we don't know $V[X]$, but as an estimate $\hat{\mathrm{V}}[X]$ we can use the variance of the sample, as in the code below:

```
sqrt(var(aa_2012)/length(aa_2012))
```

```
[1] 0.01056641
```

For confirmation, we can use `lm()`, which we will see a lot more of – we use it for linear regression. The code below runs a linear regression of `aa_2012` on a constant and gets a `summary()` table of the results. Note the `Std. Error` of the `(Intercept)` matches our answer above:

```
summary(lm(aa_2012 ~ 1))
```

```

Call:
lm(formula = aa_2012 ~ 1)

Residuals:
```

```
     Min       1Q   Median       3Q      Max 
-1.94826 -0.94826  0.05174  1.05174  1.05174 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.94826    0.01057     279   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.027 on 9451 degrees of freedom
```

Question 3: plug-in sample variance and covariance

(3a) Suppose that the first 100 responses to `env_2012` is your sample. Compute the sample variance both using the plug-in sample variance estimator (Definition 3.2.18 in Aronow & Miller) and using the `var()` function in `R` (which is the *unbiased sample variance*). Confirm that the difference between them matches theory.

```r
samp <- env_2012[1:100]
# plug-in sample variance
mean(samp^2) - mean(samp)^2
```

```
[1] 1.6275
```

```r
# unbiased sample variance
var(samp)
```

```
[1] 1.643939
```

```r
# the first should be the second times (n-1)/n
var(samp)*(99/100)
```

```
[1] 1.6275
```

(3b) Using the plug-in sample variance as a guide, write down the formula for plug-in sample covariance between two random variables X and Y.

Answer:

Because covariance of X and Y can be written $\text{Cov}[X, Y] = \text{E}[XY] - \text{E}[X]\text{E}[Y]$, the plug-in sample covariance is $\overline{XY} - \overline{X}\,\overline{Y}$.

(3c) Using `R`, compute the plug-in sample covariance between the first 100 observations of `aa_2012` and the first 100 observations of `env_2012`. Compare it to the covariance computed using `cov()`. Does the sign of the sample covariance make sense?

Answer:

```
y <- aa_2012[1:100]
x <- env_2012[1:100]
# plug-in sample covariance
mean(x*y) - mean(x)*mean(y)
```

```
[1] 0.8275
```

```
# unbiased sample covariance
cov(x, y)
```

```
[1] 0.8358586
```

As with sample variance, the plug-in version is smaller than the unbiased version (`cov()`).

It makes sense that the covariance between `aa_2012` and `env_2012` is positive, because in American politics attitudes toward affirmative action are related to attitudes toward taking action to protect the environment: people who oppose one tend to oppose the other, and people who support one tend to support the other.

(3d) Suppose we want to summarize the relationship between `aa_2012` and `env_2012` in the US population using the whole sample (all 9452 observations). We focus on the **best linear predictor** (BLP) of `aa_2012` using `env_2012`. Compute the plug-in estimate of the BLP's slope coefficient (β). Compare it to the slope coefficient you get using `lm(aa_2012 ~ env_2012)`.

Answer

The slope coefficient of the BLP of Y using X is $\frac{\text{Cov}[X,Y]}{\text{V}[X,Y]}$.

The plug-in estimator of this slope coefficient is therefore

$$\frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2}.$$

In R, we have:

```
y <- aa_2012
x <- env_2012
(mean(x*y) - mean(x)*mean(y))/(mean(x^2) - mean(x)^2)
```

```
[1] 0.3678295
```

This is exactly the same as the regression slope:

```
reg_result <- lm(y ~ x)
reg_result
```

```

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     1.7729       0.3678
```

```
coef(reg_result)
```

```
(Intercept)           x
  1.7728994   0.3678295
```