Lab 4: Simulation, Sampling, and the Central Limit Theorem

Statistics 12

Submission Instructions

1. Save your final answers in PDF format.
2. Upload your PDF to the CCLE submission link.
3. Submit the assignment as final.

Objectives

1. Reinforce understanding of simulating and random sampling
2. Understand using confidence intervals to estimate population proportions
3. Demonstrate the Central Limit Theorem's application to proportions

Collaboration Policy

In Lab you are encouraged to work in pairs or small groups to discuss the concepts on the assignments. However, DO NOT copy each other's work as this constitutes cheating. The work you submit must be entirely your own. If you have a question in lab, feel free to reach out to other groups or talk to me if you get stuck.

Simple Random Sampling

In lab 3, you learned about how to simulate the roll of a die by sampling from the numbers 1 to 6 using the sample() function. We can also use the sample() function to conduct simple random sampling from a population by sampling from the row numbers of a data frame. This is done in two steps. We use NCbirths as an example:

(1) We will use sample() to randomly select *n* numbers between 1 and 1992 (the number of babies in the NCbirths data frame). This represents choosing the babies based on ID numbers. Note that we typically use the default argument replace = FALSE to ensure we get *n* unique ID's
(2) We then use the selected numbers from Step (1) as an index for the rows of observations in the NCbirths data frame that we want to extract as our sample.

As an example, try out the code on the following page which takes a simple random sample of size 5 from the NCbirths data frame.

```
# Set the seed for reproduceability
set.seed(123)
# Select 5 numbers from 1 to 1992.
sample.index <- sample(1992, size = 5)
# Display the indices we sampled.
sample.index
## [1]  573 1570  814 1757 1870
# Extract the rows in NCbirths that correspond to sample.index.
NCbirths[sample.index,]
```

Exercise 1

The Sweetums candy factory in Pawnee, Indiana, is under investigation for violating EPA regulations. Factory workers have improperly disposed of arsenic and sulfur waste from the candy-making process, and the contamination has reached the local water supply! We have data for arsenic and sulfur levels from the water in all houses within a 2-mile radius of the factory. Download the "pawnee.csv" file from CCLE, then read it into RStudio with the following line:

```
pawnee <- read.csv ("pawnee.csv", header=TRUE)
```

Some important variables include:

- Arsenic: arsenic levels for each home in ppm
- Sulfur: sulfur levels for each home in ppm
- New_hlth_issue: Indicates "Y" if someone living at the home has experienced a major health issue after the date of contamination, "N" if no new health complications.

a. Use the head() function to print out the first few rows of this data. Then, use the dim() function to print out the number of rows and columns of this data frame.

b. Set the seed to 1337 and take a simple random sample of size 30 from the entire pawnee data frame. Save the random sample as a separate R object, and print the first few lines to make sure you saved it correctly.

c. Report the mean arsenic level from the sample you took in b. Also report the proportion of households experiencing a major health issue from your sample.

d. What symbol from lecture would we use for the mean arsenic level in the sample? What symbol would we use for the proportion of health issues in the sample?

e. Now, let's generate confidence intervals for the population proportion using the sample results. Produce 90%, 95%, and 99% confidence intervals for the true population proportion. Consult your lecture slides if you are unsure how to do this. You can use R and/or a calculator for this question, but please include code or calculations to show your work.

f. What would be the bounds of a 100% confidence interval for the population proportion?

g. Report the proportion of all households which experienced a new major health issue.

h. Create a plot that visualizes the distribution of arsenic levels for the houses in Pawnee. *Hint: we can consider the arsenic levels continuous data*

Simulations of samples and sample distributions using a for loop

We want to illustrate the sampling variability (also called sample-to-sample variability) of the sample mean and the sample proportion. That is, when we take diff erent random samples, how does the sample mean of the arsenic levels vary from sample to sample? How does the sample proportion of households who experienced a new health issue vary from sample to sample?

We can simulate sampling many (1000) random samples of size 30 from the population of households in the Pawnee data using a for loop. For each random sample, we can compute the mean arsenic levels and the proportion of the sample who experience a new health issue.

Here the code is a for loop to simulate the sample proportions from 1000 random samples of size 30. Please carefully read the comments for each line to understand the code. We will also discuss the code in lab section.

```
# We first create objects for common quantities we will use for this exercise.
n <- 30 # The sample size
N <- 541 # The population size
M <- 1000 # Number of samples/repetitions
# Create vectors to store the simulated proportions from each repetition.
phats <- c() # for sample proportions
# Set the seed for reproduceability
set.seed(123)
# Always set the seed OUTSIDE the for loop.
# Now we start the loop. Let i cycle over the numbers 1 to 1000 (i.e., iterate 1000 times).
for(i in 1:M){
        # The i-th iteration of the for loop represents a single repetition.
        # Take a simple random sample of size n from the population of size N.
        index <- sample(N, size = n)
        # Save the random sample in the sample_i vector.
        sample_i <- pawnee[index,]
        # Compute the proportion of the i-th sample of households with a new health issue.
        phats[i] <- mean(sample_i$New_hlth_issue == "Y")
}
```

Note that the do() function from last lab could have been used here, but for loops are much more versatile and can be used in a wider variety of settings.

Exercise 2 - Proportions

a. Run the entire chunk of code in the previous page to run a for loop that creates a vector of sample proportions. Using the results, create a relative frequency histogram of the sampling distribution of sample proportions. Superimpose a density curve by adding the argument: density = TRUE.

b. What is the mean and standard deviation of the simulated sample proportions?

c. Do you think the simulated distribution of sample proportions is approximately normal? Explain why or why not.

d. Using the theory-based method (i.e., normal approximation by invoking the Central Limit Theorem), what would you predict the mean and standard deviation of the sampling distribution of sample proportions to be? How close are these predictions to your answers from Part B?

Exercise 3 - Means

a. Create a new for loop to create a vector of sample means of the arsenic levels. Use n = 30, N = 514, and M =1000 just like before, and set the seed to 123.

b. Create a relative frequency histogram of the sampling distribution of sample means for arsenic. Superimpose a density curve by adding the argument: density = TRUE.

c. Do you think the simulated distribution of sample arsenic means is approximately normal? Explain why or why not. If your answer was different from your answer to Exercise 2(c), why do you think this is the case?