BIOSTAT 702: Exercise 2

Estimation

August 22, 2025

Contents

Learning Objectives

1. Use simulation to understand sampling distributions and confidence intervals

How to Do This Exercise

We recommend that you read this entire document prior to answering any of the questions. If anything is unclear please ask for help from the instructors or TAs before getting started. You are also allowed to ask for help from the instructors or TAs while you are working on the assignment. You may collaborate with your classmates on this assignment—in fact, we encourage this–and use any technology resources available to you, including Internet searches, generative AI tools, etc. However, if you collaborate with others on this assignment please be aware that *you must submit answers to the questions written in your own words. This means that you should not quote phrases from other sources, including AI tools, even with proper attribution.* Although quoting with proper attribution is good scholarly practice, it will be considered failure to follow the instructions for this assignment and you will be asked to revise and resubmit your answer. In this eventuality, points may be deducted in accordance with the grading rubric for this assignment as described below. Finally, you do not need to cite sources that you used to answer the questions for this assignment.

Grading Rubric

The assignment is worth 20 points (4 points per question). The points for each question are awarded as follows: 3 points for answering all parts of the question and following directions, and 1 point for a correct answer. Partial credit may be awarded at the instructor's discretion.

Question 1

Consider our research question from class: What is the average height of Duke students (in inches)? Suppose we are all-knowing, and know the true mean is 67 inches (5'7''). The true standard deviation is 3.5 inches. The distribution is normal.

Simulate 100 samples with n = 100 from this distribution. The answer to this question will simply be your code.

```
mu = 67
sd = 3.5

set.seed(10935)
height_samples = sapply(1:100, function(x)
  rnorm(100, mu, sd))
```

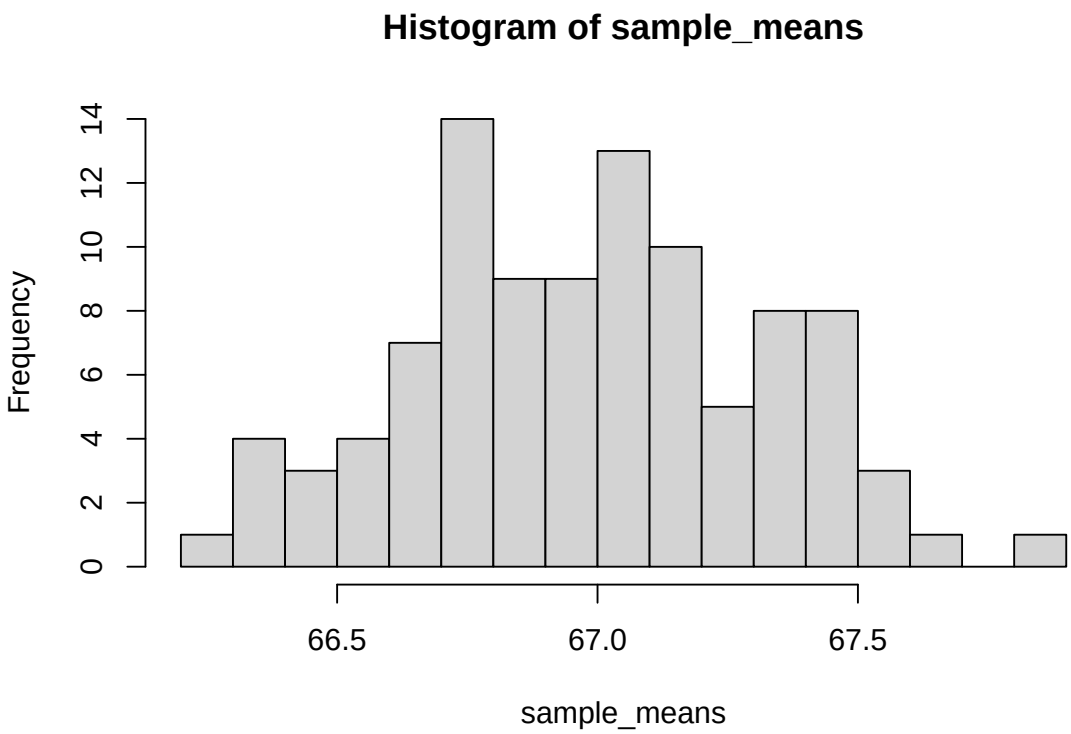
Question 2

1. Calculate the sample means for each of your 100 samples. Plot these in a histogram. What does it look like?

   The histogram looks pretty symmetric around 67 inches, with a slight right skew.

   ```
   sample_means = height_samples %>% colMeans()

   hist(sample_means, breaks = 20)
   ```



2. What is the mean of your sample means? What did you expect it to be?

   The mean is 66.98. I expected it to be 67.

```
mean(sample_means)
```

```
## [1] 66.98327
```

3. What is the standard deviation of your sample means? What did you expect it to be?

   The standard deviation is 0.337. I expected it to be the true SD divided by the square root of n, which would be 0.35. It is pretty close to this.

```
sd(sample_means)
```

```
## [1] 0.3366694
```

Question 3

1. Calculate the 95% Confidence Interval for each sample. The answer to this question is just your code.

```
sample_se = apply(height_samples, 2, function(x) sd(x)/10)
lower_ci = sample_means - 1.96*sample_se
upper_ci = sample_means + 1.96*sample_se
```

2. What percentage of your samples contain the true mean? What did you expect?

   96 contained the true mean. I expected 95% of them to contain 67.

```
height_stats = data.frame(sample_means = sample_means,
                          sample_se = sample_se,
                          lower_ci = lower_ci,
                          upper_ci = upper_ci) %>%
  mutate(includes_mean = ifelse(lower_ci <= 67 & upper_ci >= 67, 1, 0))

sum(height_stats$includes_mean)
```

```
## [1] 96
```

Question 4

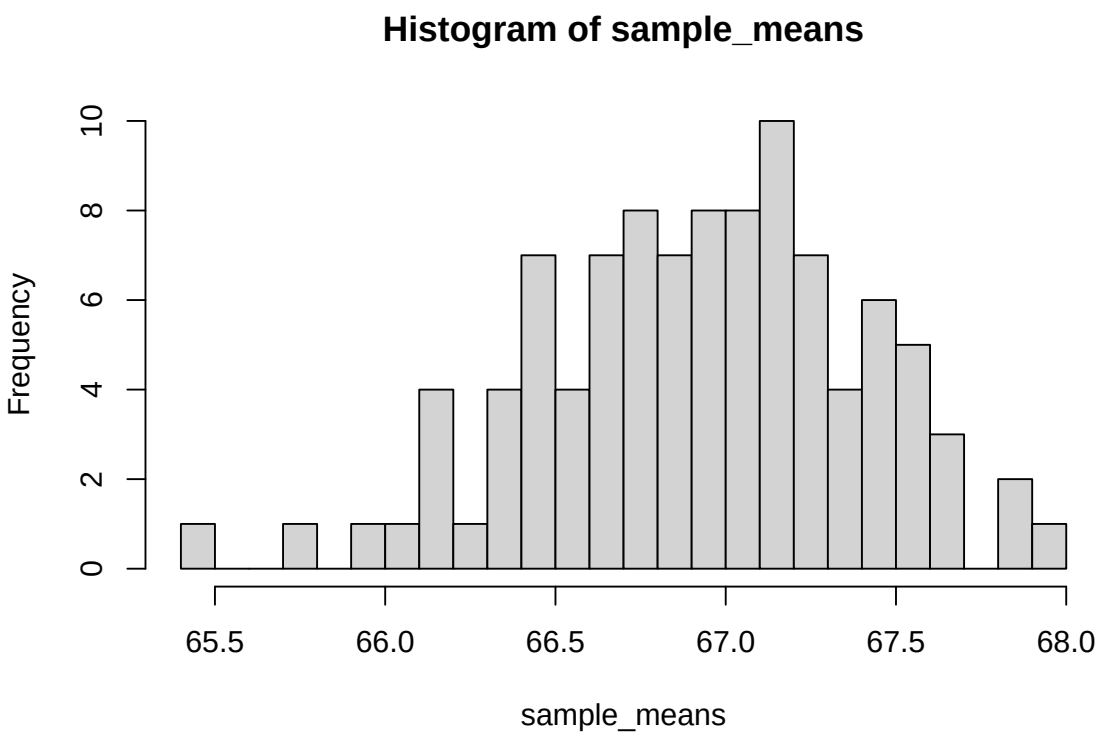
We are going to repeat everything with n = 50. Simulate 100 samples with n = 50 from the same distribution as Question 1.

```
set.seed(10935)
height_samples = sapply(1:100, function(x)
  rnorm(50, mu, sd))
```

1. Calculate the sample means for each of your 100 samples. Plot these in a histogram.

```
sample_means = height_samples %>% colMeans()

hist(sample_means, breaks = 20)
```



2. What is the mean of your sample means? What is the standard deviation of your sample means?

```
mean(sample_means)
```

```
## [1] 66.92124
```

```
sd(sample_means)
```

```
## [1] 0.4828906
```

3. Calculate the 95% Confidence Interval for each sample. What percentage of your samples contain the true mean?

```
sample_se = apply(height_samples, 2, function(x) sd(x)/sqrt(50))
lower_ci = sample_means - qt(0.975, 49)*sample_se
upper_ci = sample_means + qt(0.975, 49)*sample_se

height_stats = data.frame(sample_means = sample_means,
                          sample_se = sample_se,
                          lower_ci = lower_ci,
                          upper_ci = upper_ci) %>%
  mutate(includes_mean = ifelse(lower_ci <= 67 & upper_ci >= 67, 1, 0))

sum(height_stats$includes_mean)
```

```
## [1] 95
```

Question 5

1. How much did the results differ from using n = 100. Why do you think this is?

   The histogram looks slightly skewed the opposite direction, but still centered on 67. The range of values for the sample mean is larger, as well as the standard deviation of the sample means. The mean of the sample means is still close to 67, but slightly further away. The reason the standard deviation is larger is because the standard error for the sample mean is calculated by dividing by the square root of a smaller sample size; it is still close to what we expect it to be (3.5/sqrt(50) = 0.49). The coverage probability for the confidence intervals is lower at 92% if we use the z critical value, but is 95% using the t critical value.

2. What do you think would happen if you increased the number of samples to 1000?

   The mean and standard deviation of the sampling distribution should be even closer to what was expected, as well as the coverage probability for the confidence intervals.

3. What do you think would happen if you did this again with a different 100 samples? Or, if you compared with a friend, would you get the exact same results?

   Different samples will yield slightly different results due to sampling variability. If we increased the number of samples, we would get results closer to each other and closer to what is expected from what we know about the population.