Inference for One Mean: Sleepless Nights

Step 1: Research Question

For any statistical investigation, we need to start with a research question we are interested in studying. For this activity, we will consider how much do STAT 218 students sleep on a typical night? Let's make the question more clear: Is the average number of hours of sleep for any given night less than the recommended eight hours?

1. Based on the research question, what is the unit of study, the population of interest, the variable, and the parameter (and symbol)?

- Unit of Study:

- Population:

- Variable:

- Parameter:

- Parameter Symbol:

Step 2: Design a Study

We will investigate whether the mean amount of sleep last night for the population of all STAT 218 students was less than 8 hours.

Null and Alternative Hypotheses:

H_0: The population mean hours of sleep for all 218 students is 8 hours.

H_A: The population mean hours of sleep for all 218 students is less than 8 hours.

2. How would you write these hypotheses using notation instead of words?

H_0**:**

H_A**:**

To test these hypotheses, we need to collect data. Ideally a simple random sample should be collected in order to avoid bias in our sampling method. However, this would take a fair bit of work.

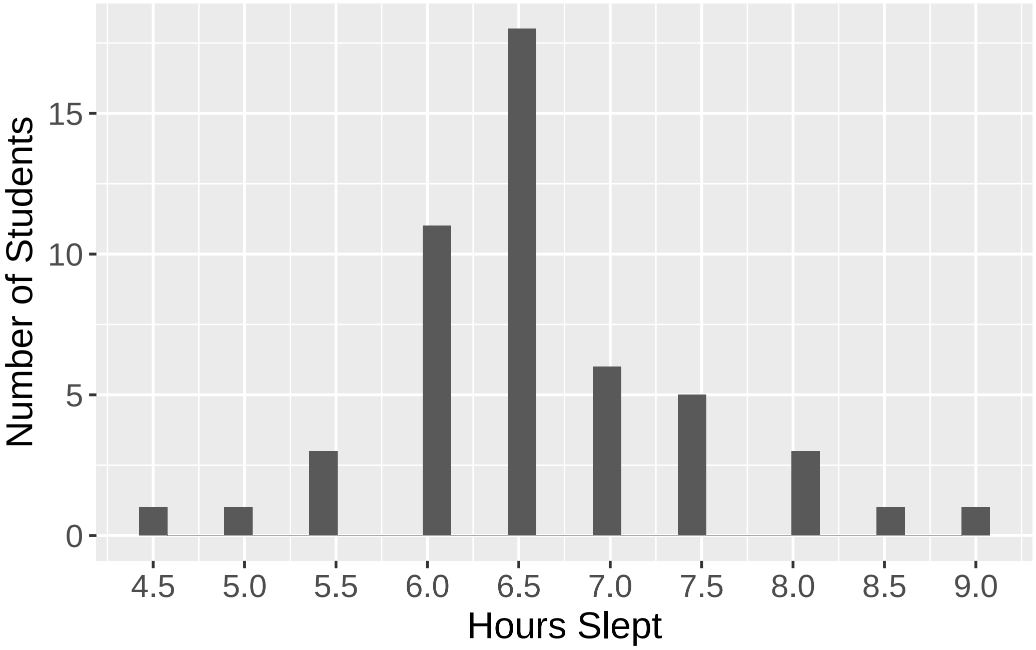
3. Two sections of STAT 218 were randomly selected out of all of the sections of STAT 218 to obtain 50 STAT 218 students. What type of sampling method could this be?

This method of sampling could be biased in some form. We will revisit the implications of using two classes as a sample of STAT 218 students later when we evaluate the study. For now, let's consider our results.

Step 3: Exploratory Data Analysis

The histogram and the statistical measures below summarize the distribution of sleep hours for the 50 sampled STAT 218 students.

```r
ggplot(data = sleep_hours,
       mapping = aes(x = hours)
       ) +
  geom_histogram() +
  labs(x = "Hours Slept",
       y = "Number of Students")
```



After your data are collected, the next step is to explore them! In the case of a quantitative variable exploring the data consists of visualizing the distribution of the responses and calculating summary statistics, to describe the shape, center, variability, and unusual observations.

4. Describe the distribution of the sample above.

Step 4: Draw inferences beyond the data

Now that we have explored the data and better understand the distribution of sleep hours for the sample of STAT 218 students, we will use both simulation and theory-based approaches to evaluate the claim that STAT 218 students get less than the recommended amount of sleep they should (8 hours).

We will use two different statistical methods to evaluate the evidence our data provide:

1. Simulation
2. Mathematical Theory

Simulation-Based Approach

To evaluate how much evidence our data provide against the null hypothesis, we need to know what means we could expect from other samples of STAT 218 students!

Statistic

The first step is to calculate the statistic of interest.

```r
favstats(~ hours, data = sleep_hours)
```

```
 min Q1 median Q3 max mean        sd  n missing
 4.5  6    6.5  7   9  6.6 0.8451543 50       0
```

5. Use the `favstats()` output above to report the statistic of interest. What is the notation for this statistic?

One Simulation

To know what other means we could expect to get from a different sample of STAT 218 students we will use **bootstrapping**. A “bootstrap” is a method of resampling from the original sample to obtain a “new” sample.

For boostrapping, the **critical** assumption we are making is that the students in our original sample are “representative” of the population of STAT 218 students. If this is true, then we can view each resample as similar to another sample that we could have gotten when sampling from the entire population.

You have been given a bag of 50 cards, where each card has the observed number of hours slept for a STAT 218 student.

6. Your team should resample (with replacement) 50 cards from the bag, writing down each number before putting the card back in. Once you have 50 values, find the mean number of hours slept in your resample.

$\bar{x} =$

7. Plot the bootstrap means other groups obtained.

Thousands of Simulations

Alright, you obtained one resample, but getting 100 resamples using our card method would take us a long time (and would be really boring). So, we will use the computer, specifically `R`, to help us get bootstrap resamples much quicker!

To obtain a boostrap resample, we have a three step process:

1. `specify` what the response variable is
2. `generate` ______ boostrap resamples
3. `calculate` the mean for each bootstrap resample

The code below does that for us and saves them in a new dataset called `bootstrap_resamples`:

```
sleep_hours %>%
  specify(response = hours) %>%
  generate(reps = 10, type = "bootstrap") %>%
  calculate(stat = "mean")
```

Here is a preview of what these bootstrap resamples look like:

```
Response: hours (numeric)
# A tibble: 10 x 2
   replicate  stat
       <int> <dbl>
 1         1  6.51
 2         2  6.62
 3         3  6.6
 4         4  6.57
 5         5  6.51
 6         6  6.74
 7         7  6.68
 8         8  6.6
 9         9  6.63
10        10  6.53
```

8. What does the `replicate` column correspond to?

9. What does the `stat` column correspond to?

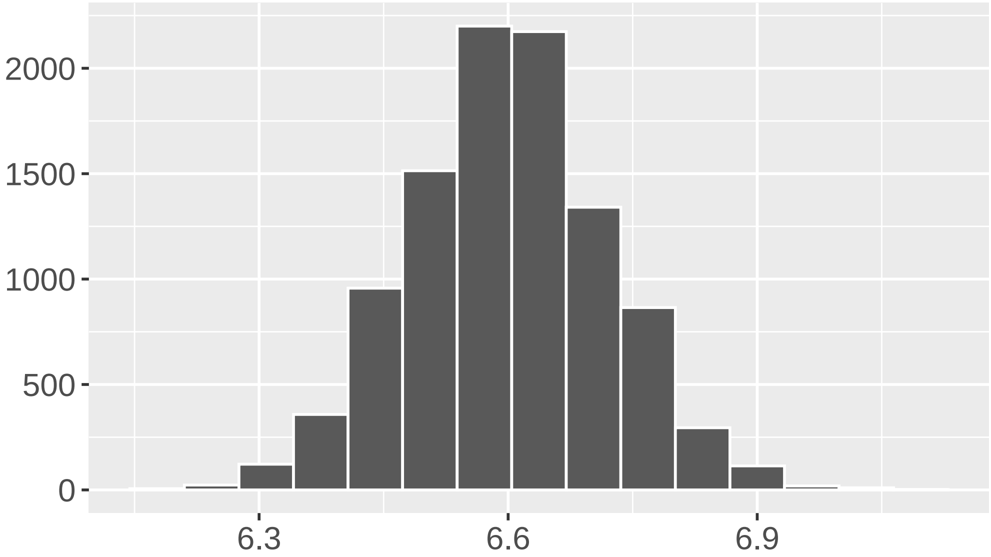
10. Of these 10 resamples, what is the smallest mean that was obtained? What was the largest mean that was obtained?

11. If you were to make a histogram or a dotplot of these 10 means, where do you believe it would be centered?

Bootstrap Distribution

Alright, to get a good idea of what means we might get from other samples, we will obtain lots of bootstrap resamples. Typically, we will use *at least* 1,000 resamples, so we get a good idea of what the distribution of the boostrap statistics looks like.

The distribution of bootstrap statistics (in this case means) has a specific name. We call this the **bootstrap distribution**. I've run the code to obtain 1,000 bootstrap resamples and plotted the results in the histogram below.



12. Fill in the x-axis and y-axis labels for the bootstrap distribution.

13. How would you describe the shape of the distribution?

14. Where is the distribution centered? Why do you believe it is centered there?

Step 5: Making Conclusions

We use a bootstrap distribution to see the variability in the statistics we might have seen from other samples from the population. These different statistics give us an idea as to where we believe the population parameter might lie.

15. What is the population parameter we are trying to estimate?

There are two ways to obtain a confidence interval, (1) using the "percentile" method and (2) using the "SE" method.

The percentile method uses percentiles to decide the endpoints of the interval. I've provided a table of different percentiles to help you create your confidence interval.

Quantile	Value
0.5%	6.40
1%	6.33
2.5%	6.37
5%	6.40
90%	6.76
95%	6.80
97.5%	6.84
99.5%	6.80

16. Suppose we are interested in constructing a 95% confidence interval. Using the table above, report the end points of this confidence interval.

17. Interpret the confidence interval in the context of this investigation.

18. Given the values of your 95% confidence interval, do you believe it is reasonable to assume that STAT 218 students get 8 hours of sleep? Why or why not?