Activity 3: Sleepless Nights

Name: _____

October 4, 2022

Part One: Simulation-Based Methods

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)?

- Observational unit:
- Population of interest:
- Variable of interest:

- Parameter (in words and assign a symbol):

- Parameter Symbol:

Step 2: Design a Study & Collect Data

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.

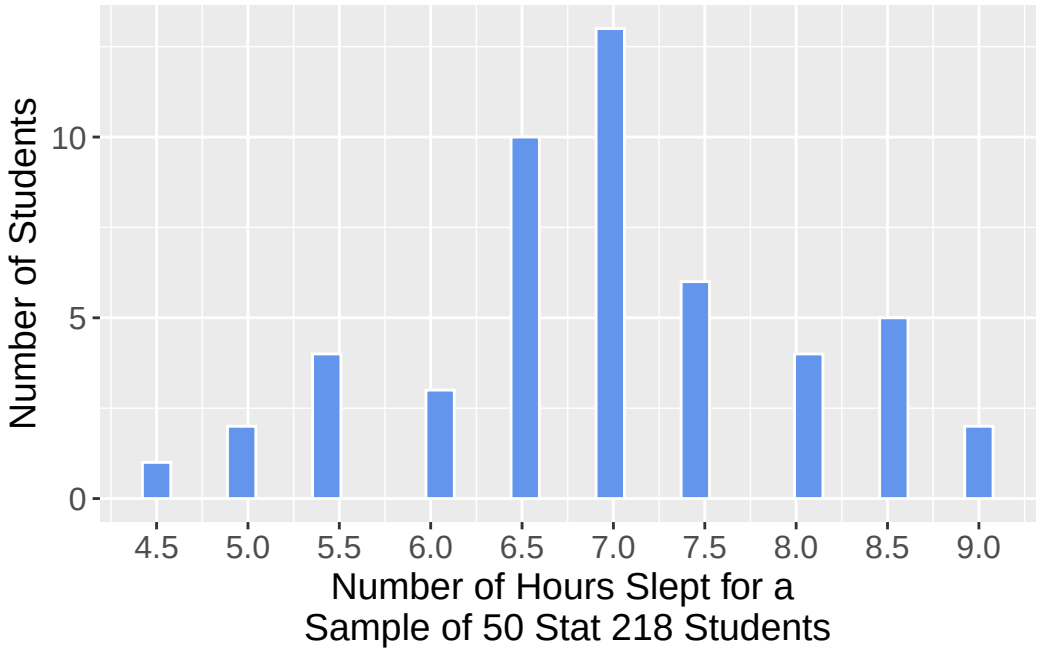
3. Two sections of STAT 218 were randomly selected out of all of the sections of STAT 218 to obtain 50 STAT 218 students. What type of sampling method could this be?

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 (EDA)

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

```
ggplot(data = sleep_hours,
       mapping = aes(x = hours)
       ) +
  geom_histogram() +
  labs(x = "Number of Hours Slept for a Sample of 50 Stat 218 Students",
       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.

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

```
 min  Q1 median  Q3 max mean       sd  n missing
 4.5 6.5      7 7.5   9 6.96 1.044128 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  7.1
 2         2  7.04
 3         3  7.05
 4         4  6.86
 5         5  6.66
 6         6  7.06
 7         7  6.98
 8         8  7.03
 9         9  6.71
10        10  7.23
```

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

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

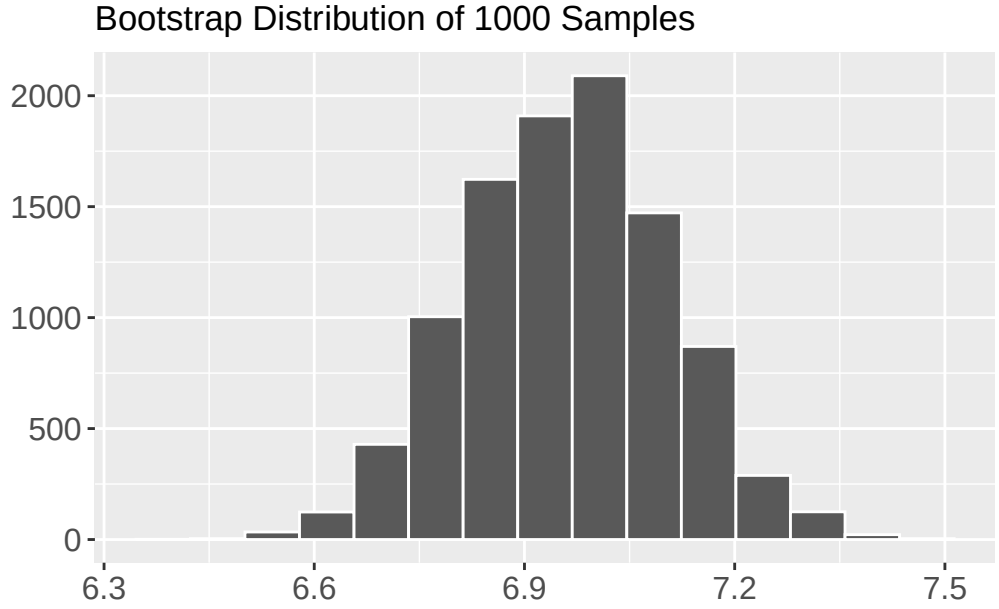
10. Of these 10 resamples, what is the smallest mean that was obtained? What was the largest mean that was obtained?

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%	6.72
1%	6.63
2.5%	6.68
5%	6.72
90%	7.15
95%	7.19
97.5%	7.24
99.5%	7.19

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?

Part Two – Using Theory-Based Methods (AKA Mathematical SE Method)

Thus far, we explored utilizing a bootstrap distribution to obtain a confidence interval for the population mean. But there is another approach we could have used instead! Now, we will focus on methods wihc use mathematical formulas instead of computer simulation.

These "theory-based" mathematical formulas have a similar idea:

> **Approximate a distribution of statistics we might have expected from other samples.**

This distribution has a special name, it is called a **sampling distribution**. A sampling distribution is a *distribution of statistics* calculated for different samples. This week, we are focusing on the mean. So, our sampling distribution will visualize the variability in sample means we would expect from other samples.

During Part 1, we created a sampling distribution using bootstrapping. We resampled from our original data to create "new" samples we could have expected to obtain from other samples of STAT 218 students. This class, we will instead use mathematical theory to obtain our sampling distribution.

Central Limit Theorem (CLT)

This key theorem in Statistics says that when we collect a "sufficiently large" sample of n **independent** observations from a population with mean μ and standard deviation σ, we know the **sampling distribution** of $\bar{x}$ will be nearly Normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

In order for us to feel confident that we can use the CLT with our data, we need to check two conditions:

- Independence of Observations
- Normality

19. Do you believe that the 50 observations collected in this sample are independent? Why or why not?

20. Based on the histogram from yesterday's activity, do you believe it is safe to say that the distribution of hours slept is approximately Normal? Why or why not?

The t-distribution

The t-distribution became well known in 1908, in a paper in *Biometrika* published by William Sealey Gosset. Gosset published the paper under the pseudonym "Student," which is why you

sometimes hear the distribution called "Student's t." Gosset worked at the Guinness Brewery in Dublin, Ireland, and was interested in the problems of small samples (Wikipedia article).

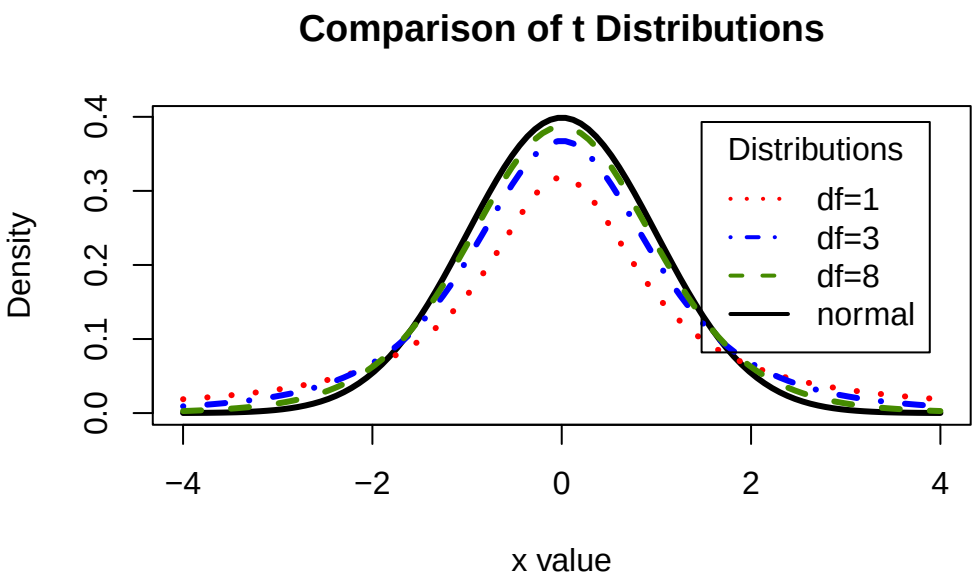


Figure 1: Comparison of the standard Normal vs t-distribution with various degrees of freedom

If we believe the CLT can work for our data, mathematically we will use the t-distribution as an **approximation** for the sampling distribution. The t-distribution is always centered at zero and has a single parameter: **degrees of freedom**. The degrees of freedom describe exactly what the shape of the t-distribution looks like.

We will use a t-distribution with $n-1$ degrees of freedom to model the sample mean. When we have more observations, the degrees of freedom will be larger and the t-distribution will look more like the Normal distribution.

21. How many degrees of freedom will we use for our t-distribution?

22. Compared to a t-distribution with 20 degrees of freedom, will your distribution have *more* or *less* area in the tails?

The CLT says if we have a "large" sample of independent observations and don't have any outliers, then we know the sampling distribution has a standard deviation of $\frac{\sigma}{\sqrt{n}}$. But, we don't usually know the value of σ, since it is the **population** standard deviation. So, instead we substitute in s, the sample standard deviation: $\frac{s}{\sqrt{n}}$.

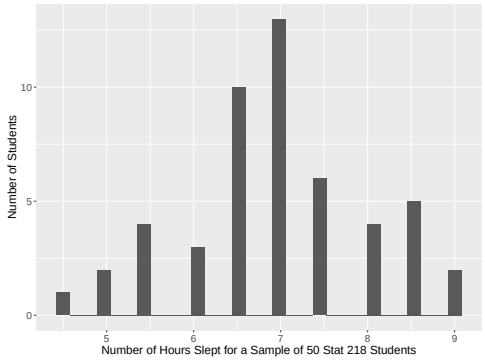
```
favstats(~ hours, data = sleep_hours)
```

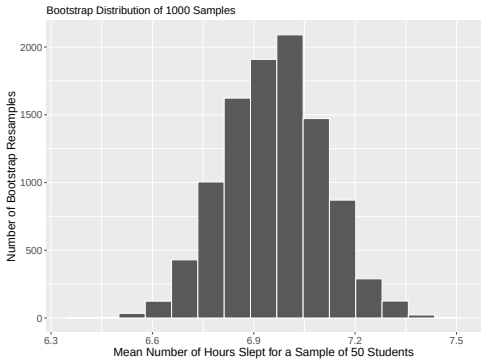
```
 min  Q1 median  Q3 max mean       sd  n missing
 4.5 6.5      7 7.5   9 6.96 1.044128 50       0
```

23. Given the summary statistics above, calculate the estimated standard deviation of the sampling distribution (standard error).

Did you notice that $\frac{\sigma}{\sqrt{n}}$ (the standard deviation of our bootstrap sampling distribution) did not equal s (the standard deviation of our observed sample)? This is because the variability between **individuals'** number of hours slept is **VERY** different from the variability between the **average** number of hours slept across samples of people.

> **Key Idea: There will be less sample-to-sample variability than in person-to-person variability!** *See the observed sample distribution and the bootstrap sampling distribution with the same x scales below to compare.*

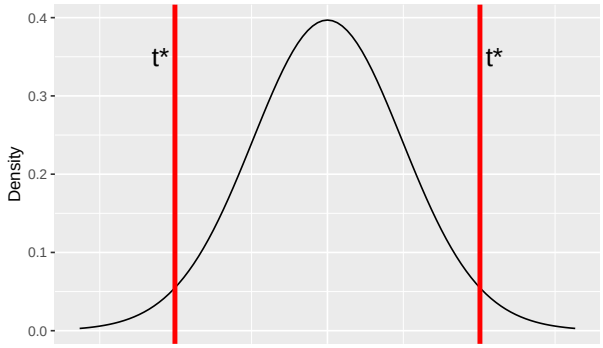




Using the t-distribution to create a confidence interval

Previously, we found our confidence interval by finding different percentiles on our bootstrap distribution. For example, we used the 2.5th and 97.5th percentile to obtain a 95% confidence interval.

When we are using a t-distribution to obtain our confidence interval, the process has similar ideas, but a slightly different approach. Since the t-distribution is centered at 0 and symmetric, the number associated with the 2.5th percentile and the 97.5th percentile **is the same**. Well, one is positive and one is negative, but they have the same numbers. So, we only need to find **one** number to make our confidence interval!



The number we are finding is called the **multiplier**. The multiplier for a confidence interval depends on two things, (1) the degrees of freedom and (2) the side of confidence interval you want. In our case we know we should use a t-distribution with 49 degrees of freedom.

24. We are interested in making a 95% confidence interval. Using the table below, circle the correct multiplier we should use to make our interval.

`R` code	Value
`qt(0.90, df = 49)`	1.299069
`qt(0.95, df = 49)`	1.676551
`qt(0.975, df = 49)`	2.009575
`qt(0.995, df = 49)`	2.679952

Now that we have the multiplier, we can put all of the pieces together! The "formula" for a t-based confidence interval is:

$$\text{point estimate} \pm t^*_{df} \times SE$$

25. Using your answers to questions 5 and 6, create a 95% confidence interval for the mean hours slept for all STAT 218 students.

26. What do we hope is contained in this interval?

27. Do we know if the interval contains this value?

28. How do you interpret the interval you found?

Exploring Confidence Intervals

29. Do you think a 90% confidence interval be wider or narrower than your 95% confidence interval? Explain.

30. When you change from a 90% to a 95% confidence interval, which part of the confidence interval is changing? (circle the correct answer)

- Statistic (midpoint)
- Multiplier
- Standard error

31. How does the multiplier change from the 95% to the 90% confidence interval? (circle the correct answer)

- Multiplier is larger
- Multiplier is smaller
- Multiplier stays the same

32. How would the center change for a 99% confidence interval compared to the 90% interval?

33. How would the standard error change for a 99% confidence interval compared to the 90% interval? Explain.

34. How would the 95% confidence interval change if you surveyed a much smaller number of students? Assume that the sample mean would still be 6.6.

Comparison with Previous Results

35. What confidence interval did you obtain yesterday? Is it similar to or different from the interval you obtained today?

36. Why do you think your intervals were different / similar?

Generalizability

37. Think again about how the sample was selected from the population. Do you feel comfortable generalizing the results of your analysis to the population of all STAT 218 students at your school? Explain.

Conclusions

It's important to keep in mind that these conditions are rough guidelines and not a guarantee! All theory-based methods are approximations which work best when the distributions are symmetric, when sample sizes are large, and when there are no large outliers. When in doubt, use a simulation-based method as a cross-check! If the two methods give very different results you should consult a statistician!