Week 3 Day 2: Sleepless Nights

Last class, we explored utilizing a bootstrap distribution to obtain a confidence interval for the population mean. There is another approach we could have used instead, which focuses on mathematical formulas and not simulation.

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

> **Obtain 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.

Last class, 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

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

2. 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).

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.

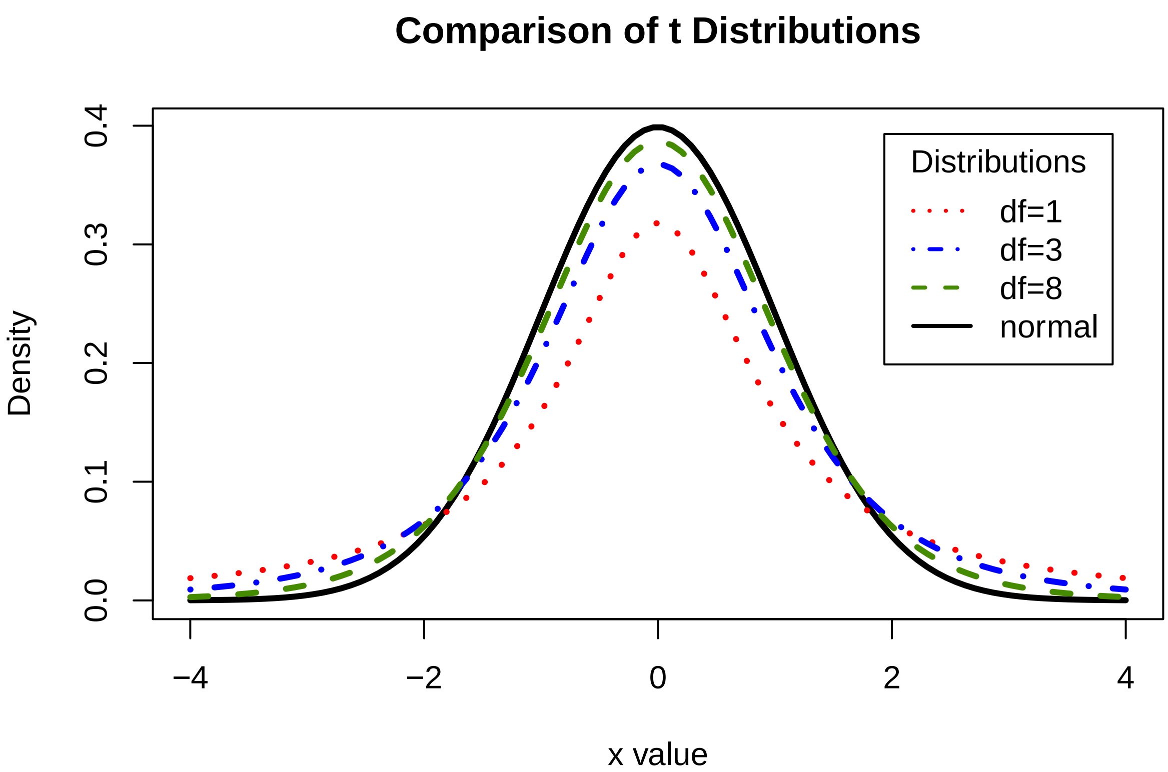


Figure 1: Comparison of the standard Normal vs t-distribution with various degrees of freedom

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.

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

4. 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
##    5 6.625    7.5  8   9 7.34 0.9971388 50       0
```

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

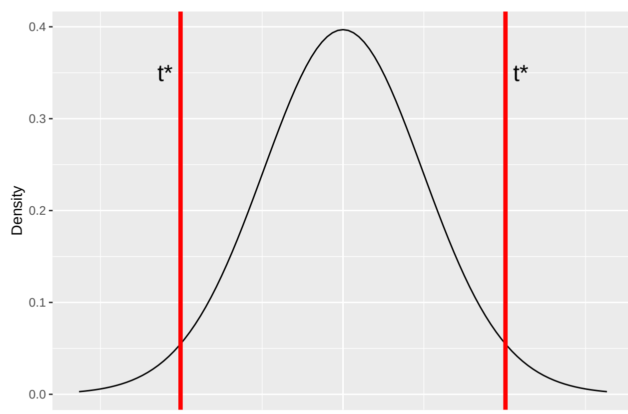
Did you notice that $\frac{\sigma}{\sqrt{n}}$ did not equal s? This is because the variability between **individuals'** number of hours slept is **VERY** different from the variability between the **average** number of hours slept across samples of people.

> **Key Idea: There will be less sample-to-sample variability than in person-to-person variability!**

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.

6. 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:

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

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

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

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

10. How do you interpret the interval you found?

Exploring Confidence Intervals

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

12. 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

13. 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

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

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

16. 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

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

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

Generalizability

19. 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!