Normal-based Inference for Population Means

The Central Limit Theorem

The Central Limit Theorem (CLT)

For a "sufficiently large" random sample, n, the sample mean is well approximated by a normal model:

$$\overline{x} \sim \mathcal{N}\left(\mu,\ \frac{\sigma}{\sqrt{n}}\right)$$

Note: If sampling with replacement, then the sample should be less than 10% of the population

How large is large enough?

It depends on the population distribution...

- **If the population distribution is normal, then any sample size works**
- **If the population distribution is not severely skewed, then $n \geq 30$ is a good rule of thumb**
- **If the population distribution is severely skewed, then a larger sample size is needed**

Example

Assessment records indicate that the value of homes in Appleton is right skewed, with a mean of \$140,000 and standard deviation of \$60,000. Consider a random sample of 100 homes in Appleton. Describe the sampling distribution of $\overline{x}$.

Inference for the **population mean**

Estimating a population mean

As we have seen, confidence intervals have the form

$$\text{statistic} \pm (\text{critical value}) \times \text{SE}$$

If the parameter of interest is the population mean this becomes

Finding the SE

Problem: In practice, we usually don't know σ!

Solution: Plug in an estimate for σ

Finding the critical value

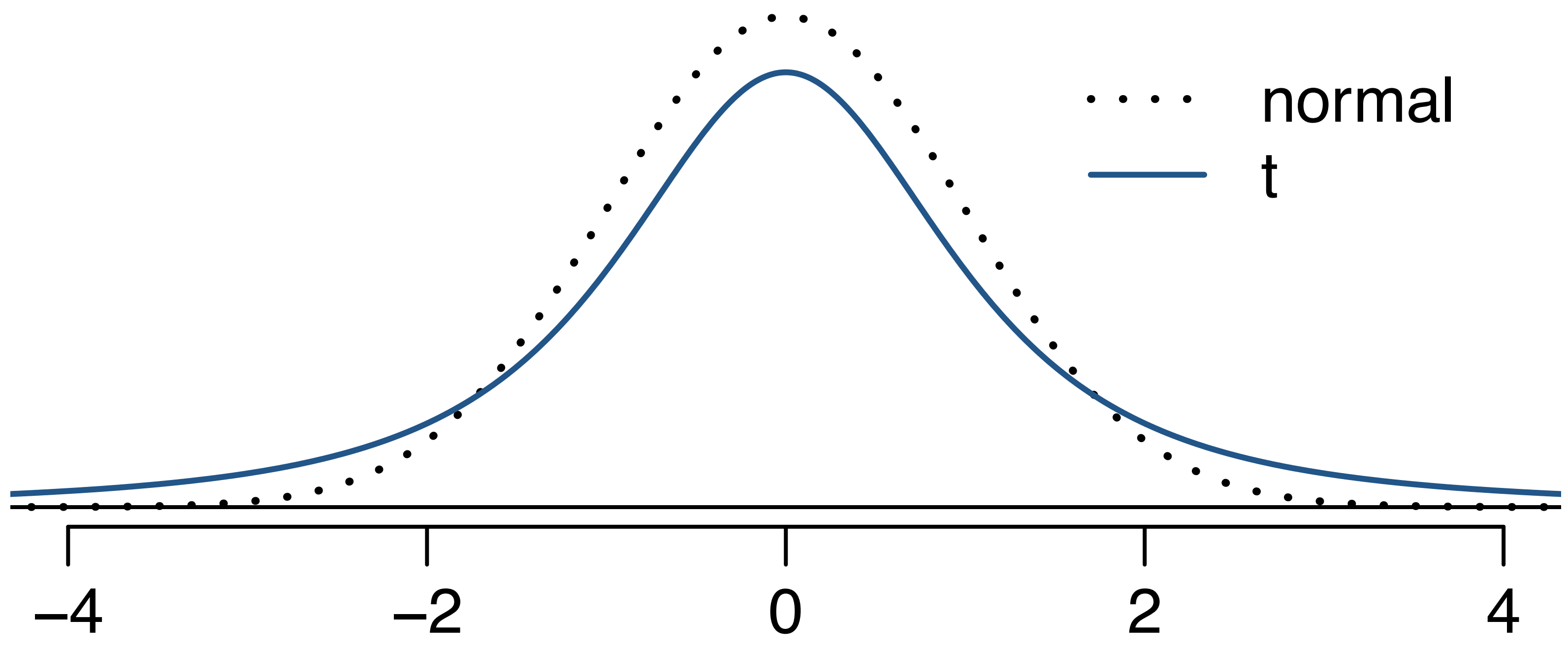
Problem: Plugging in an estimate introduces additional uncertainty.

Solution: Use a more "conservative" distribution than the normal distribution.

t-distribution

Observations more likely to fall beyond two SDs from the mean than under the normal distribution

Extra thick tails help mitigate effect of a less reliable estimate for the SE

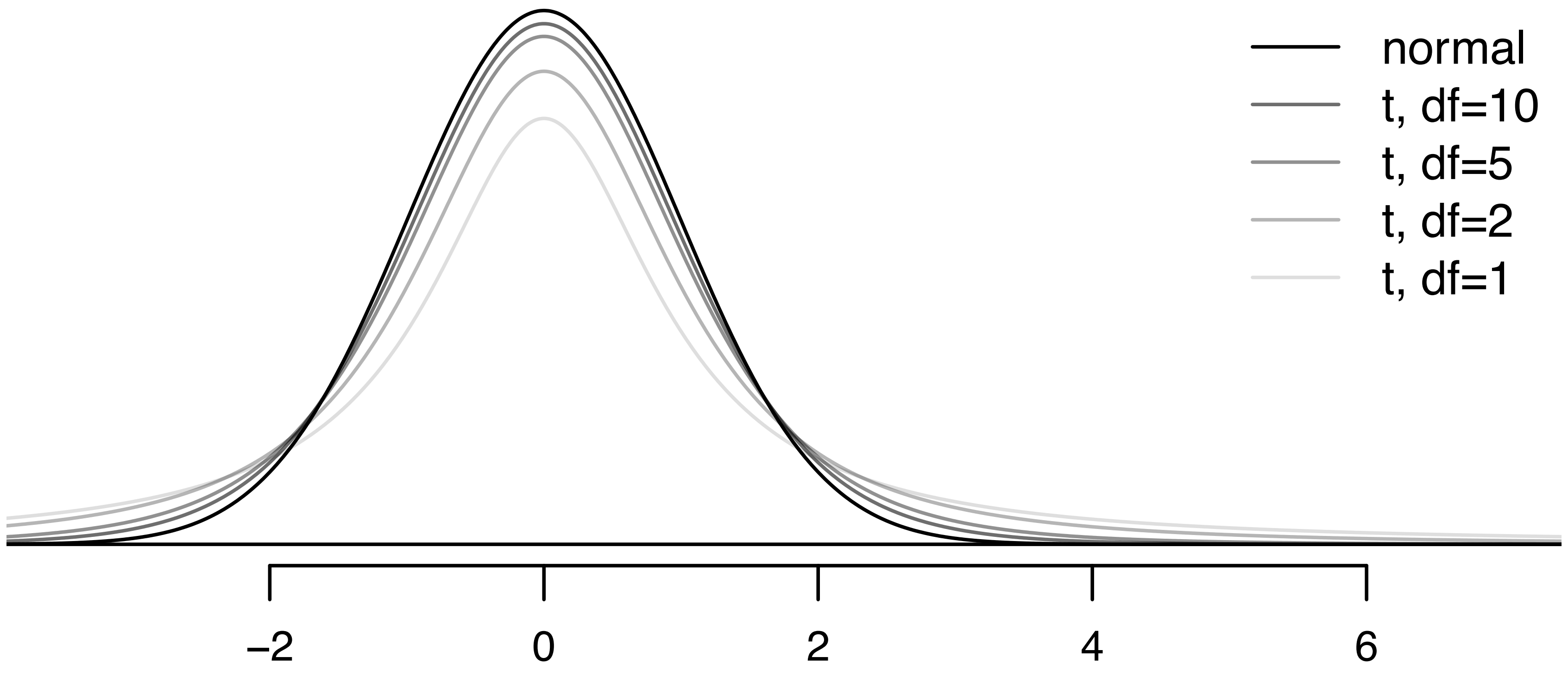


Properties of the t-distribution

Always centered at 0

Requires a single parameter to be specified, the degrees of freedom

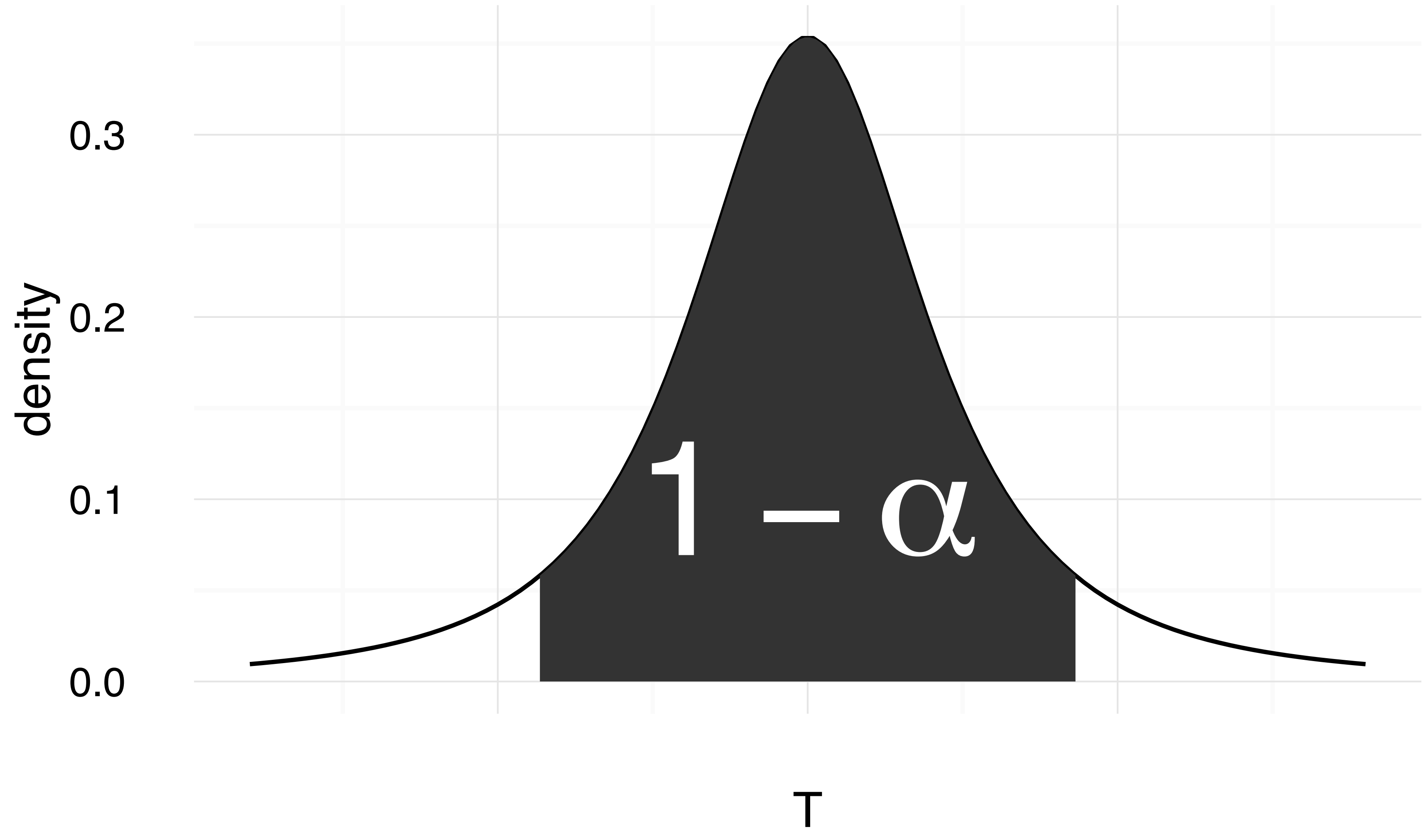
for one sample: $df = n - 1$



Confidence intervals for μ

Formula:

Finding critical values



In R use `qt(p, df)`

The critical value, t*, for a 90% confidence interval is the...

A. 90th percentile

B. 92.5th percentile

C. 95th percentile

D. 97.5th percentile

The critical value, t*, for a 97% confidence interval is the...

A. 97^{th} percentile

B. 97.5^{th} percentile

C. 98^{th} percentile

D. 98.5^{th} percentile

E. 99^{th} percentile

Example: Commuting in Atlanta, GA

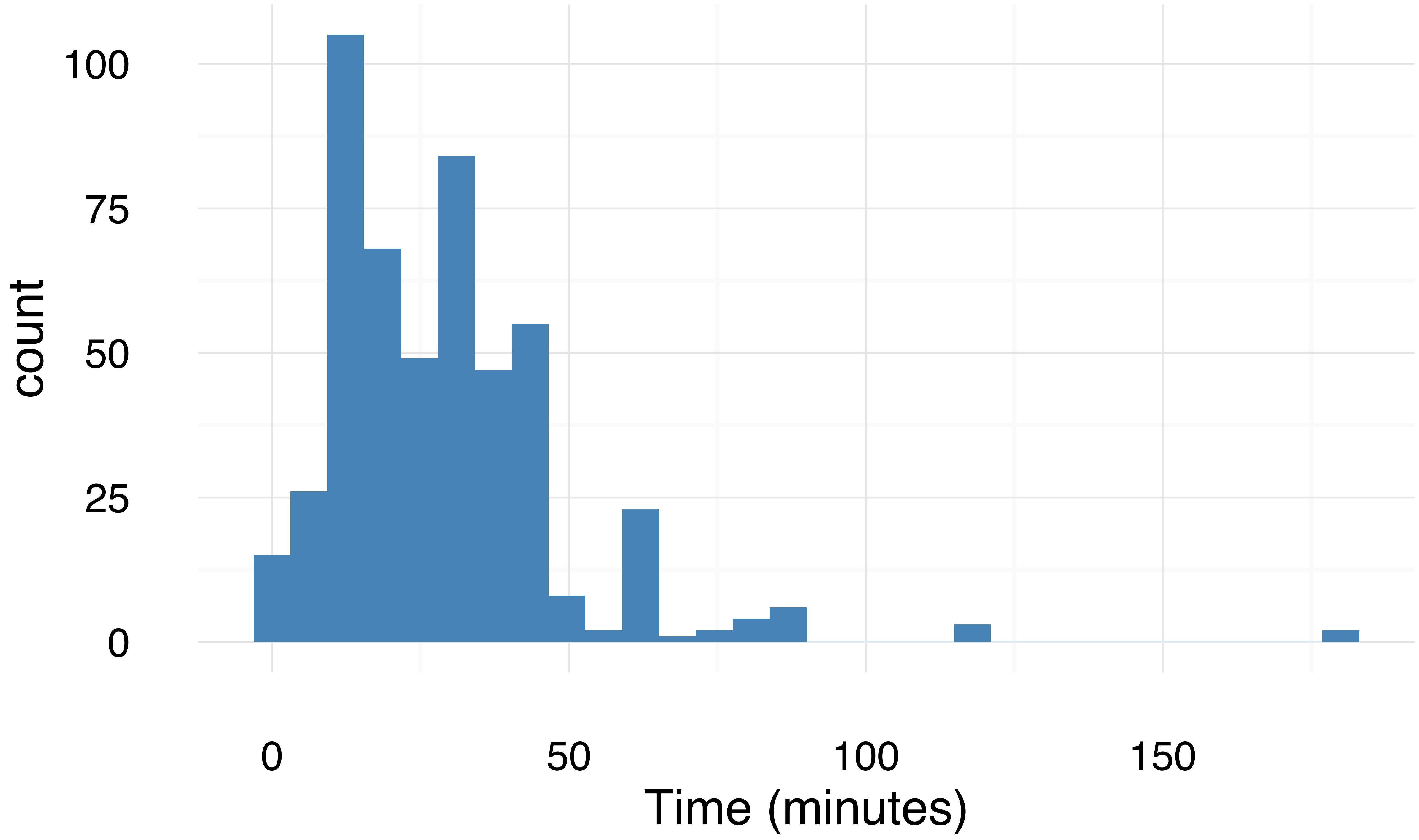
Source: American housing survey by US Census Bureau

Commute times (in min.) for random sample of 500 commuters

Sample mean is 29.11 min.

Sample standard deviation is 20.7 min.

Goal: Find and interpret a 95% confidence interval for average commute time for someone living in Atlanta, GA



100
75
50
25
0
count
0
50
100
150
Time (minutes)

Calculating CIs in R

```
# Load the mosaic package
library(mosaic)

# Load the data
commute <- read.csv("data/CommuteAtlanta.csv")

# Construct a 95% confidence interval
confint(t.test(~Time, data = commute, conf.level = 0.95))
```

Testing a population mean

Step 1. Set the hypotheses

H_0: μ = *null value*

H_a: μ < or > or ≠ *null value*

Step 2. Check the necessary assumptions

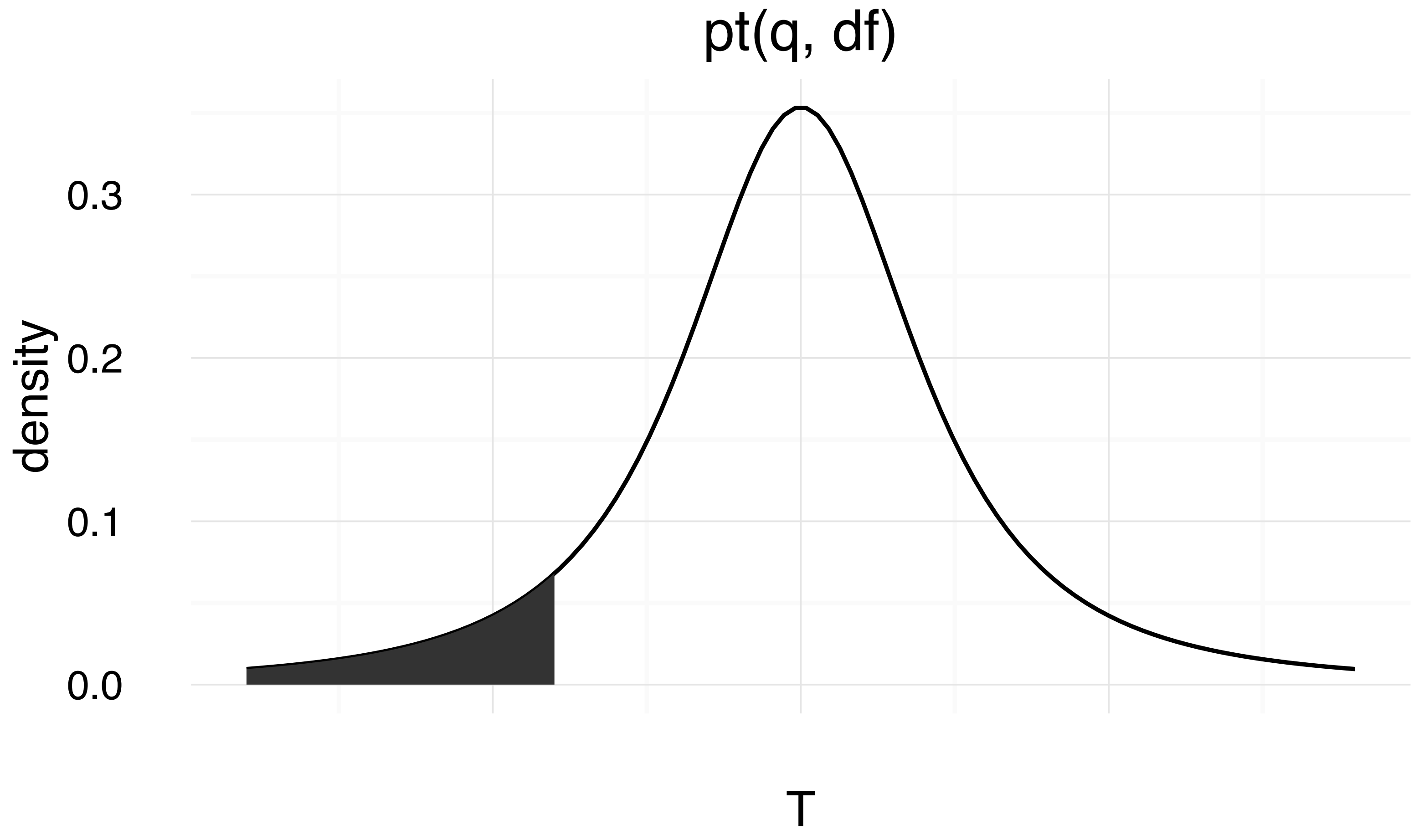
- Independence: Do we have a random sample?
- Sample size/skew: n ≥ 30 (or larger if sample is skewed)
  or
  If n < 30, nearly normal sample
- 10% condition: Is the sample size < 10% of the population?

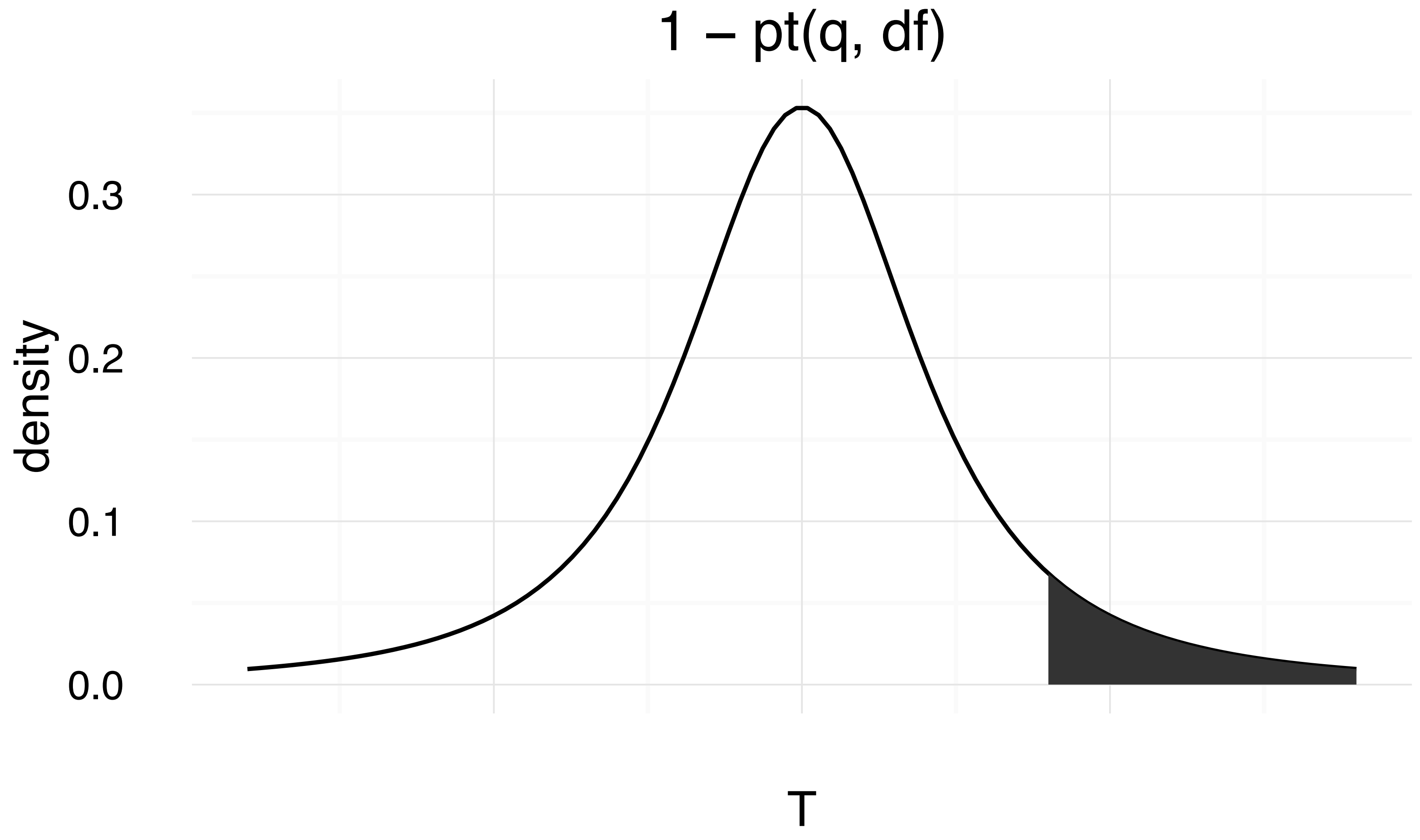
Testing a population mean

Step 3. Calculate a test statistic and a p-value

$$T = \frac{\overline{x} - \text{null value}}{SE}$$ **, where** $SE = \frac{s}{\sqrt{n}}$

In R, we use `pt` to find the p-value





Testing a population mean

Step 4. Make a decision and interpret it in context of the research question

- Use a p-value as strength of evidence against H_0
- If you are working with a set significance level, use the rejection rule

Example: College student sleep habits

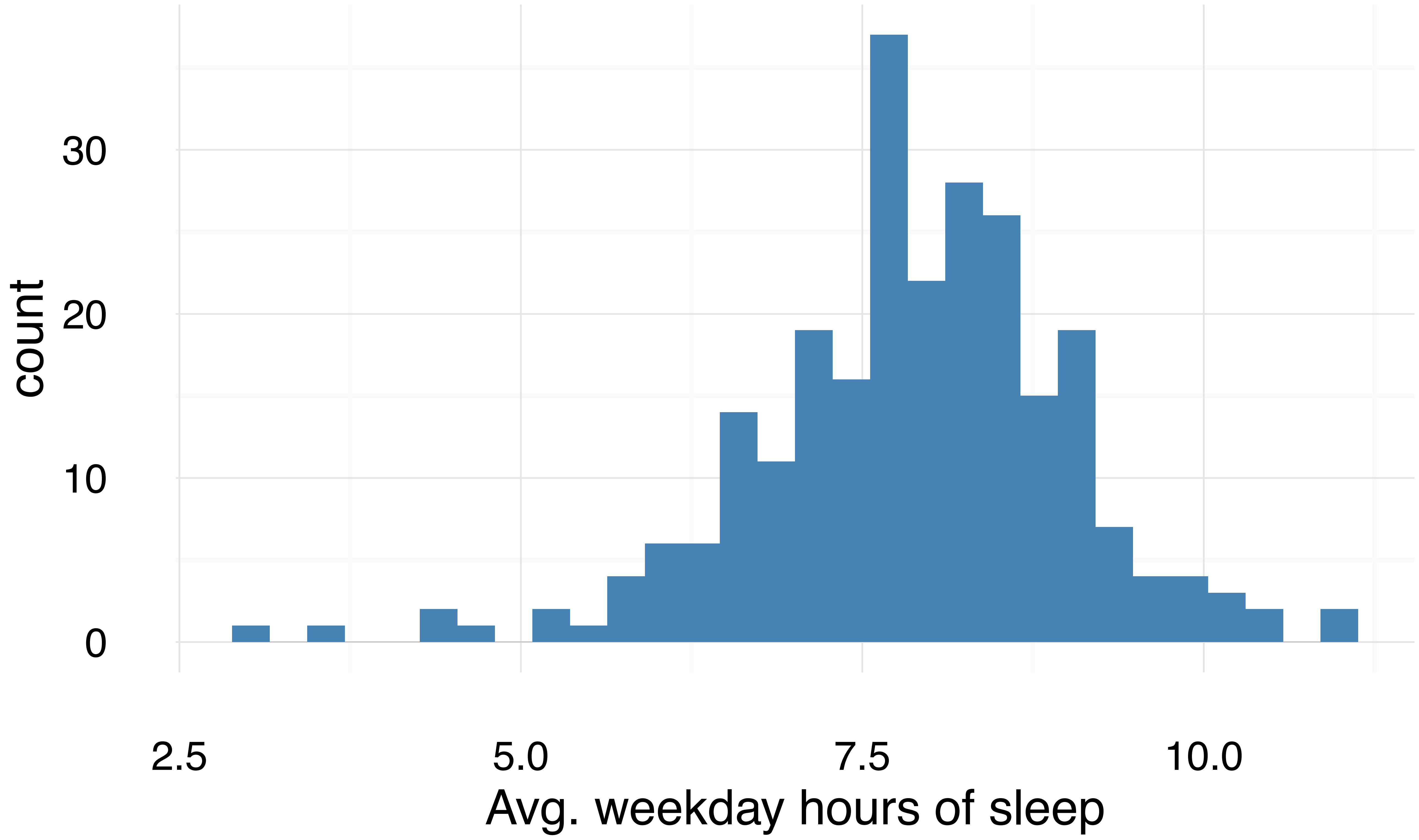
Source: random sample of 253 college students

Average hours of sleep on weekdays collected

Sample mean is 7.9 hours

Sample standard deviation is 1.1 hours

Question: Do college students sleep less than 8 hours per night?


30
20
10
0
count
2.5
5.0
7.5
10.0
Avg. weekday hours of sleep

Running hypothesis tests in R

```
# Load the mosaic package
library(mosaic)

# Load the data
sleep <- read.csv("data/SleepStudy.csv")

# Running a hypothesis test
t.test(~WeekdaySleep, data = SleepStudy, mu = 8, alternative = "less")
```

Reading `t.test` output

```
	One Sample t-test

data:  SleepStudy$WeekdaySleep
t = 107.14, df = 252, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 7.721416 8.010599
sample estimates:
mean of x
 7.866008
```