Confidence Intervals Pt. 2

Conditions, t-distribution, Different Confidence %'s

Grinnell College

Review – Confidence Intervals

We learned how to make a "95% Confidence Interval" for estimating a population mean μ

$$\overline{x} \pm 2 \times \frac{\sigma}{\sqrt{n}}$$

- 95% came from the fact that 95% of the intervals will contain μ

Intuition: Think of the confidence interval as providing a range of 'plausible' values for μ

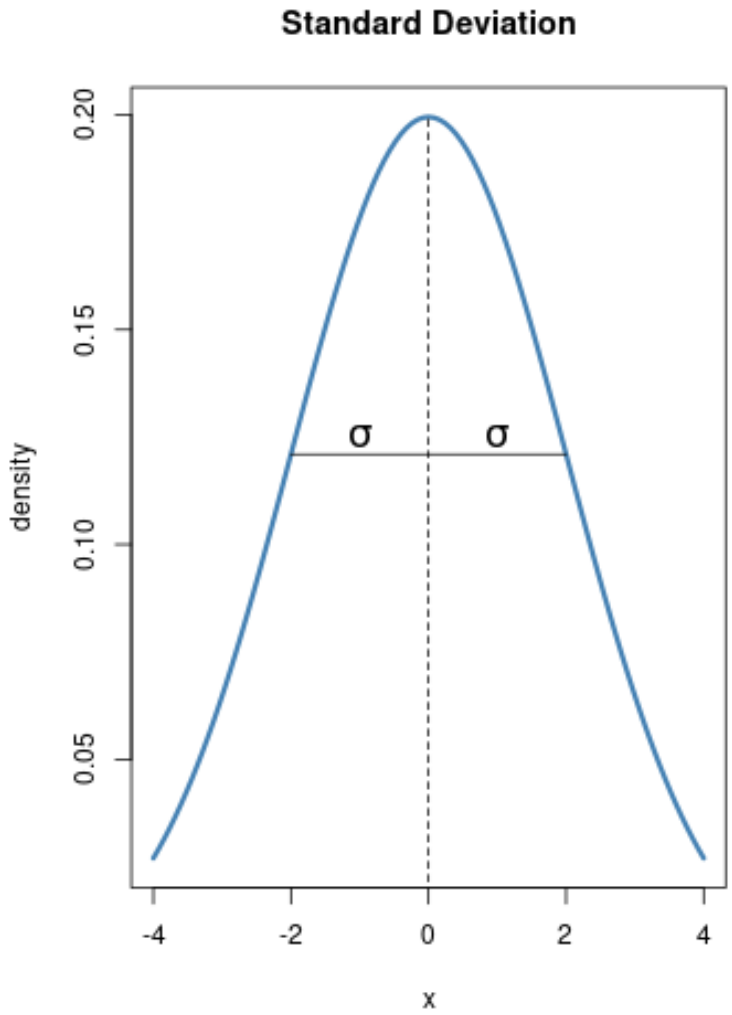
Review – Central Limit Theorem

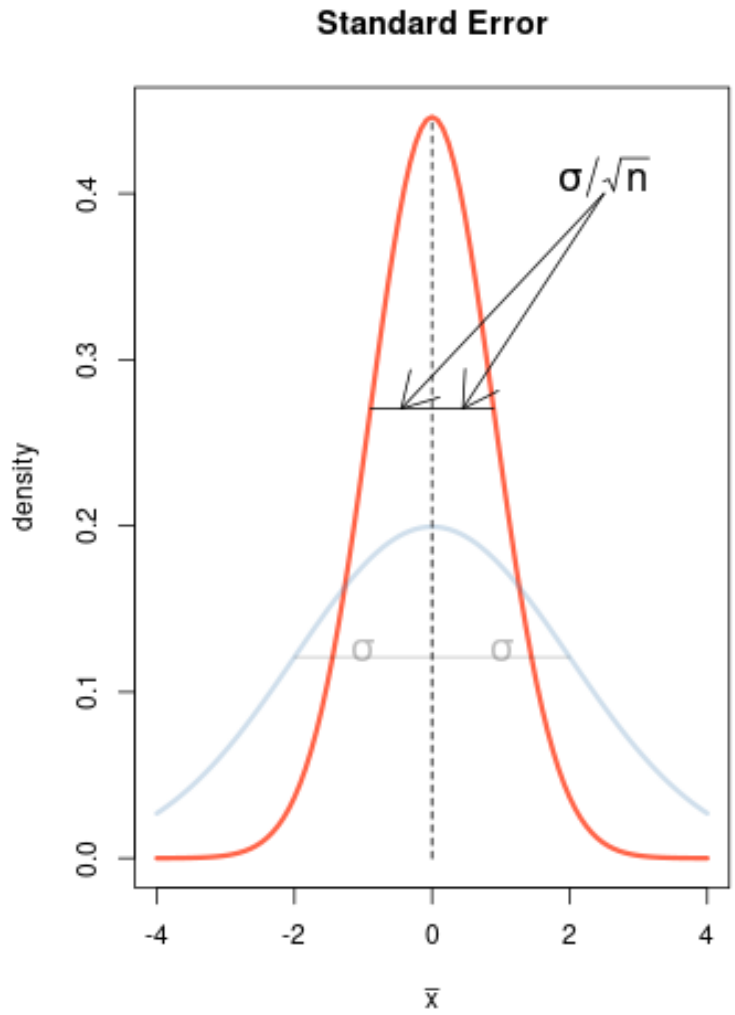
Central Limit Theorem:

1. If variable X has mean μ and std.dev. σ, and
2. If the number of observations in the sample (n) is large
3. then the sampling distribution for $\overline{X}$ (sample mean) is Normal with mean μ and standard error $\sigma/\sqrt{n}$.

$$\overline{X} \sim \mathrm{N}(\mu, \sigma/\sqrt{n})$$

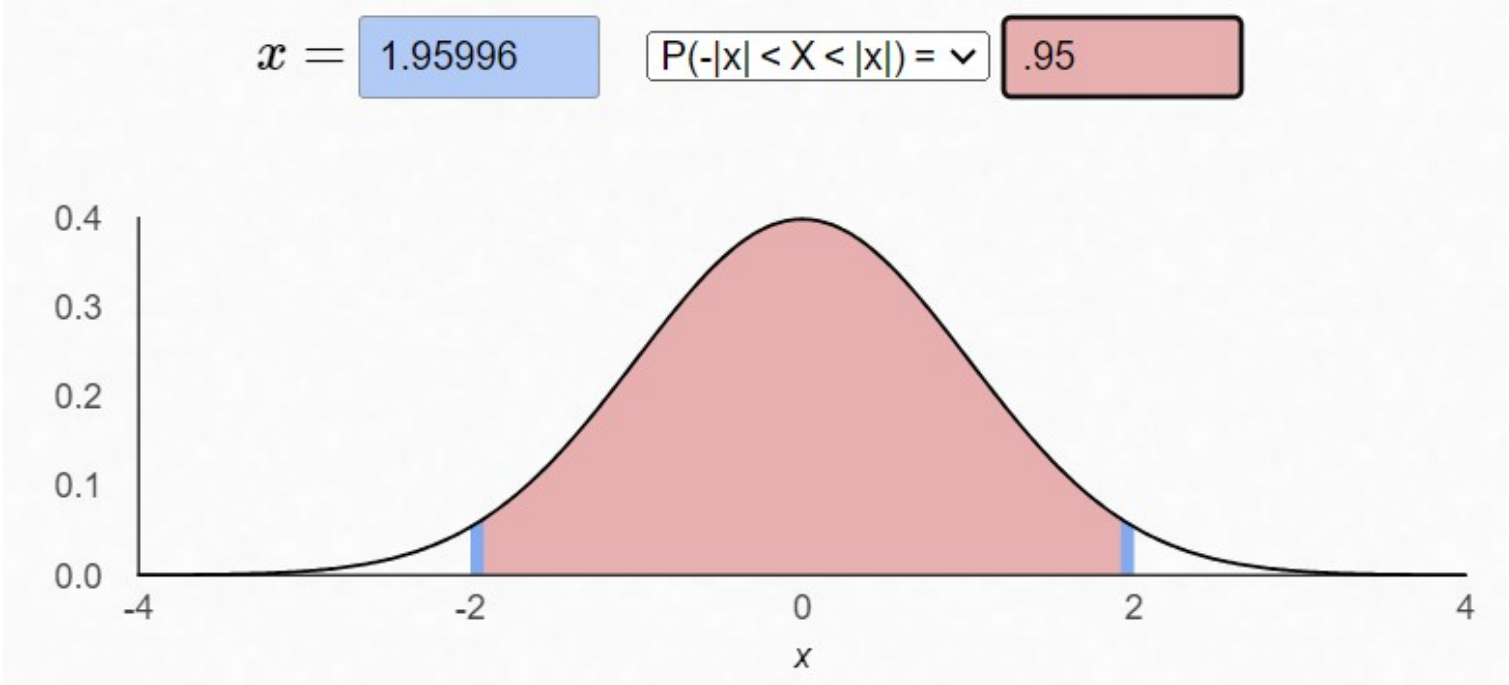
More on CLT


Standard Deviation
density
0.05
0.10
0.15
0.20
σ
σ
-4
-2
0
2
4
x


Standard Error
density
0.0
0.1
0.2
0.3
0.4
σ/√n
σ
σ
-4
-2
0
2
4
x̄

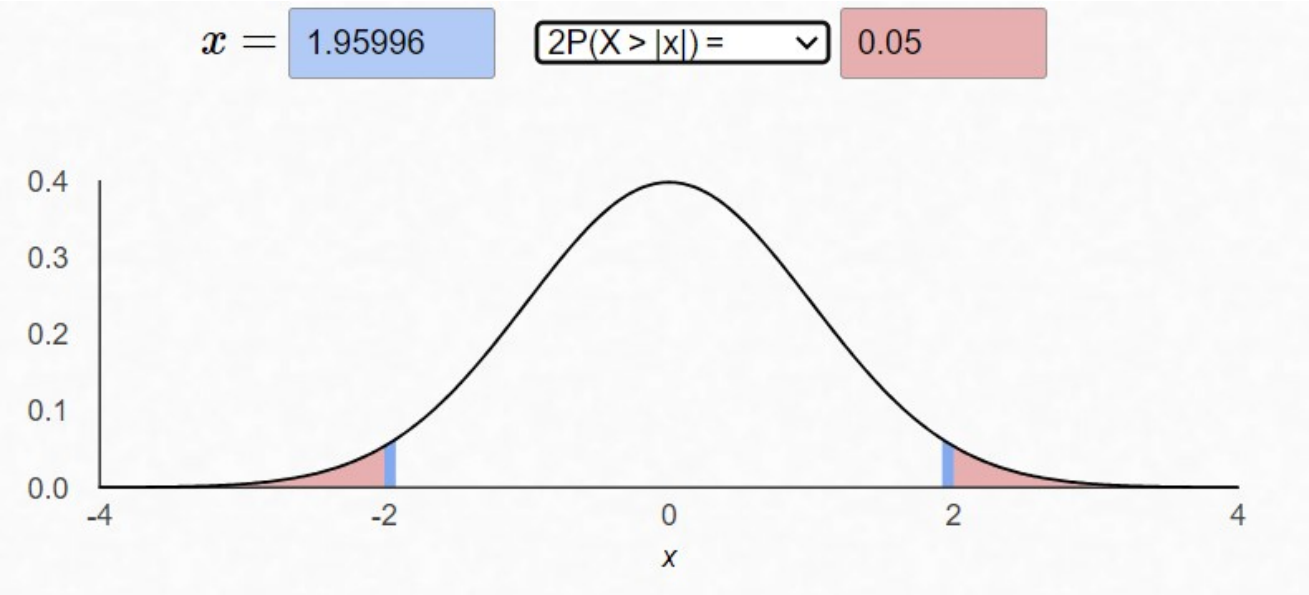
CI's using Normal Distribution

Think back to what we were doing with the normal distribution to make our 95% CI's



Note: $\pm 2 \times$SE was approximate, we actually should be using $\pm 1.96 \times$SE

CI's using Normal Distribution



We could find the value we want to use for the ME $= \pm C \times$SE using z-table

- We want 95% of values between the "cut-off" values
- equivalent we want 5% outside the cutoffs
- divide 5% in half (2.5%) to get how much probability we need on each side of the normal distribution (this gives us the cutoffs)

Conditions

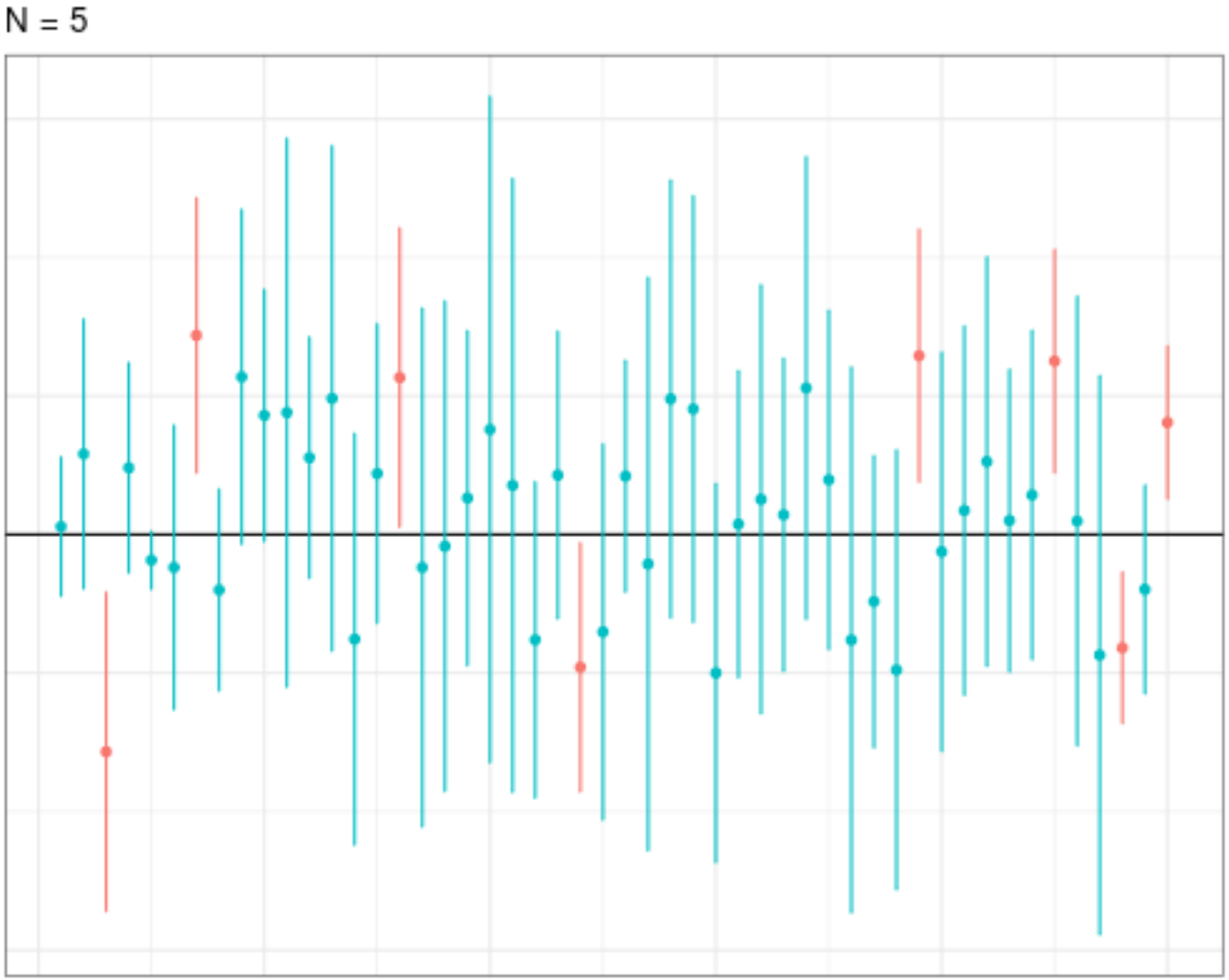
We need at least one of these things to be true in order for our 95% CI formula to work (CLT)

- The *population* must be Normal
  - we will almost always never know if this is true
- sample size large enough for CLT to make sampling dist. look Normal

Sample size rule of thumb: look at the sample distribution

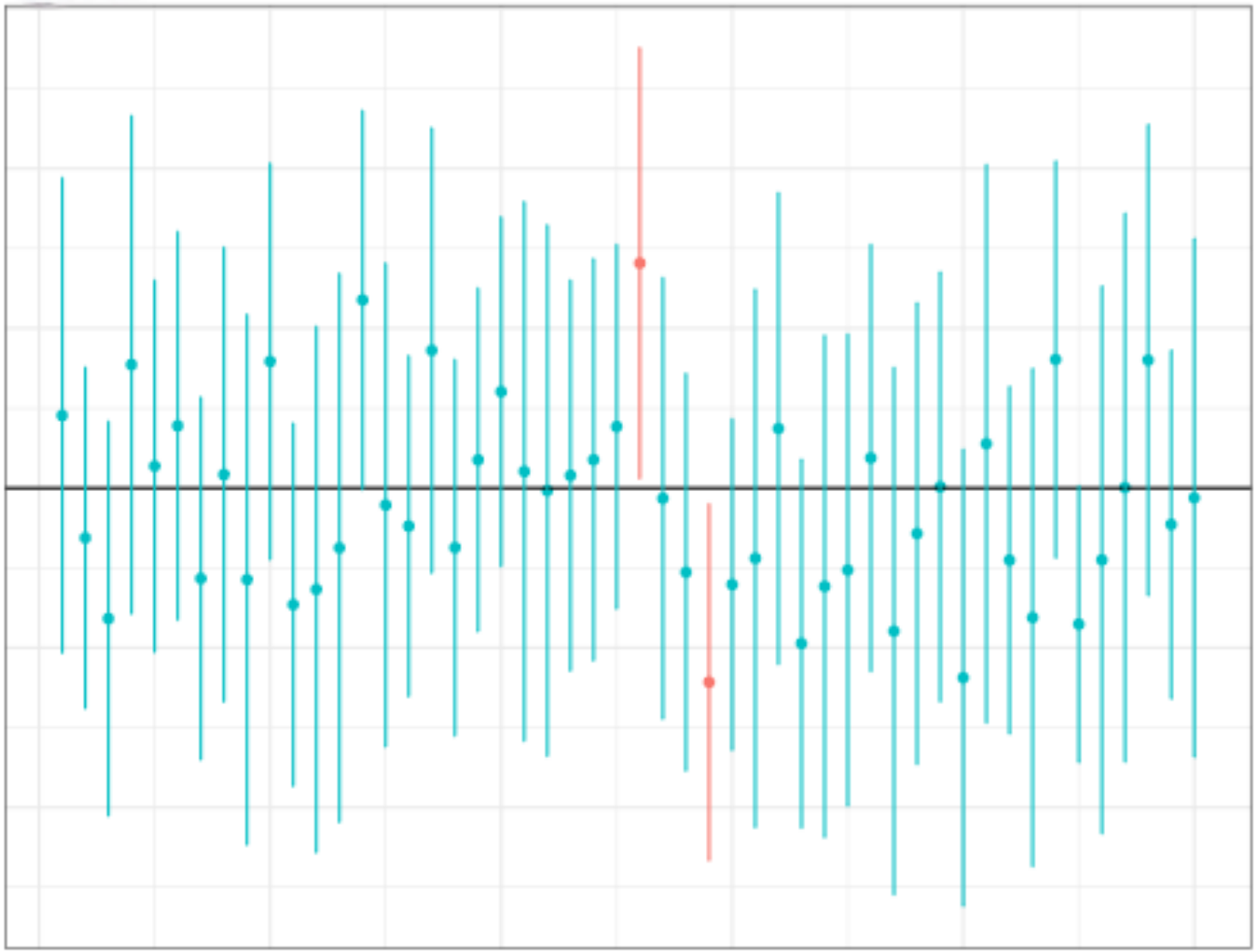
- Normal: OK
- symmetric: 15+
- moderately skewed: 30+
- extremely skewed 40+

Conditions



Conditions

$N = 50$



Conditions

We need the following to be true in order for our 95% CI formula to work (CLT)

- representative sample
- Normal pop. **OR** large enough sample size*

Why?

- Without these, the sampling distribution is not close enough to Normal (or is biased)
- Our intervals contain μ less than 95% of the time

Another Issue

95 % CI Formula (σ known):

$$\overline{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

We don't know μ, which is why we are trying to estimate it.

Do we know σ? Probably not! Let's try...

95 % CI Formula (σ unknown):

$$\overline{x} \pm 1.96 \times \frac{s}{\sqrt{n}}$$

Estimating Variability

Sampling Distribution of the sampling mean:

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

- The sampling distribution depends on σ, not **s**
- Is **s** going to be exactly equal to σ? NO!

There is uncertainty in estimating σ using **s** just like how we are using $\overline{x}$ to estimate μ

- We need a way to incorporate our uncertainty about σ into the confidence intervals we construct for $\overline{x}$

Student's t-distribution

In the 1890s, a chemist by the name of William Gosset working for Guinness Brewing became aware of the issue while investigating yields for different barley strains. (He was using CIs, substituting s for σ but the results were unreliable)

In 1906, he took a leave of absence to study under Karl Pearson where he discovered the issue to be the use of $\hat{\sigma}$ with σ interchangeably

To account for the additional uncertainty in using $\hat{\sigma}$ as a substitute, he introduced a modified distribution that has "wider tails" than the standard normal

However, because Guinness was not keen on its competitors finding out that it was hiring statisticians, he was forced to publish his new distribution under the pseudonym "student", hence "Student's t-distribution"

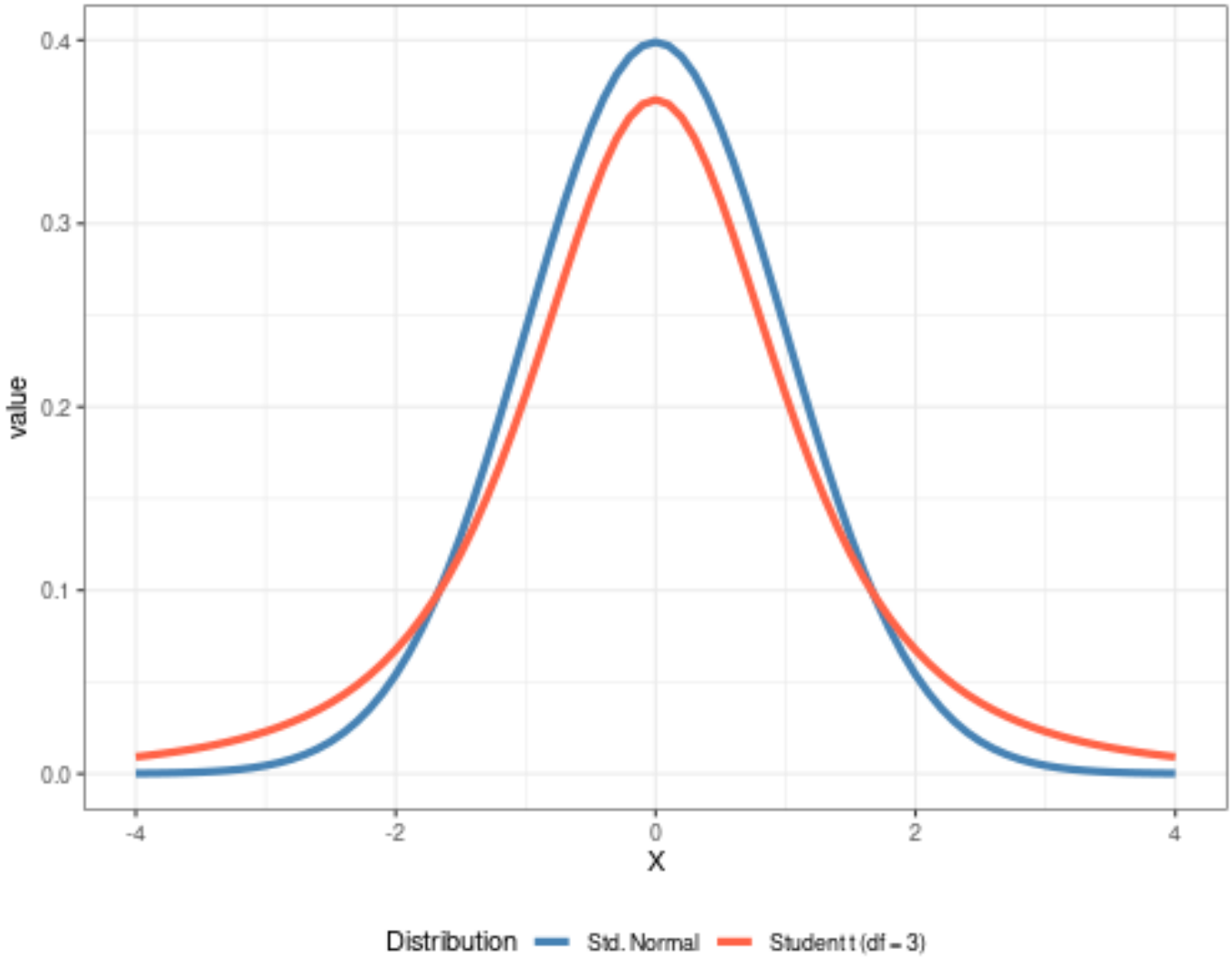
Student's t-distribution

Student's t Distribution:

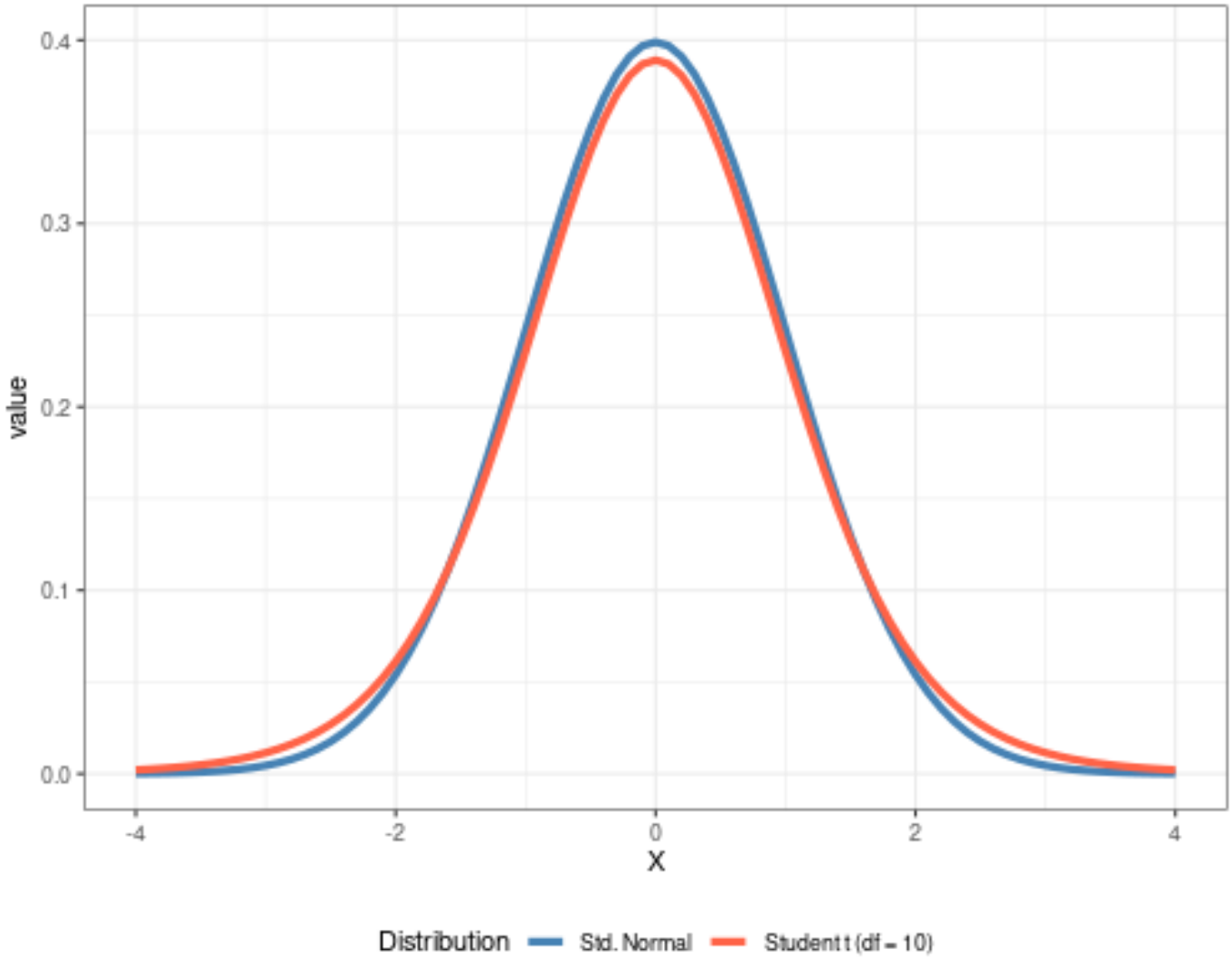
$$X \sim t_{n-1}$$

1. very similar to standard Normal
2. symmetric + unimodal (bell-shaped)
3. has more probability in the tails than the normal distribution
4. parameter called *degrees of freedom*, equal to $n - 1$, controls shape
5. The t distribution will become Normal as $n \to \infty$

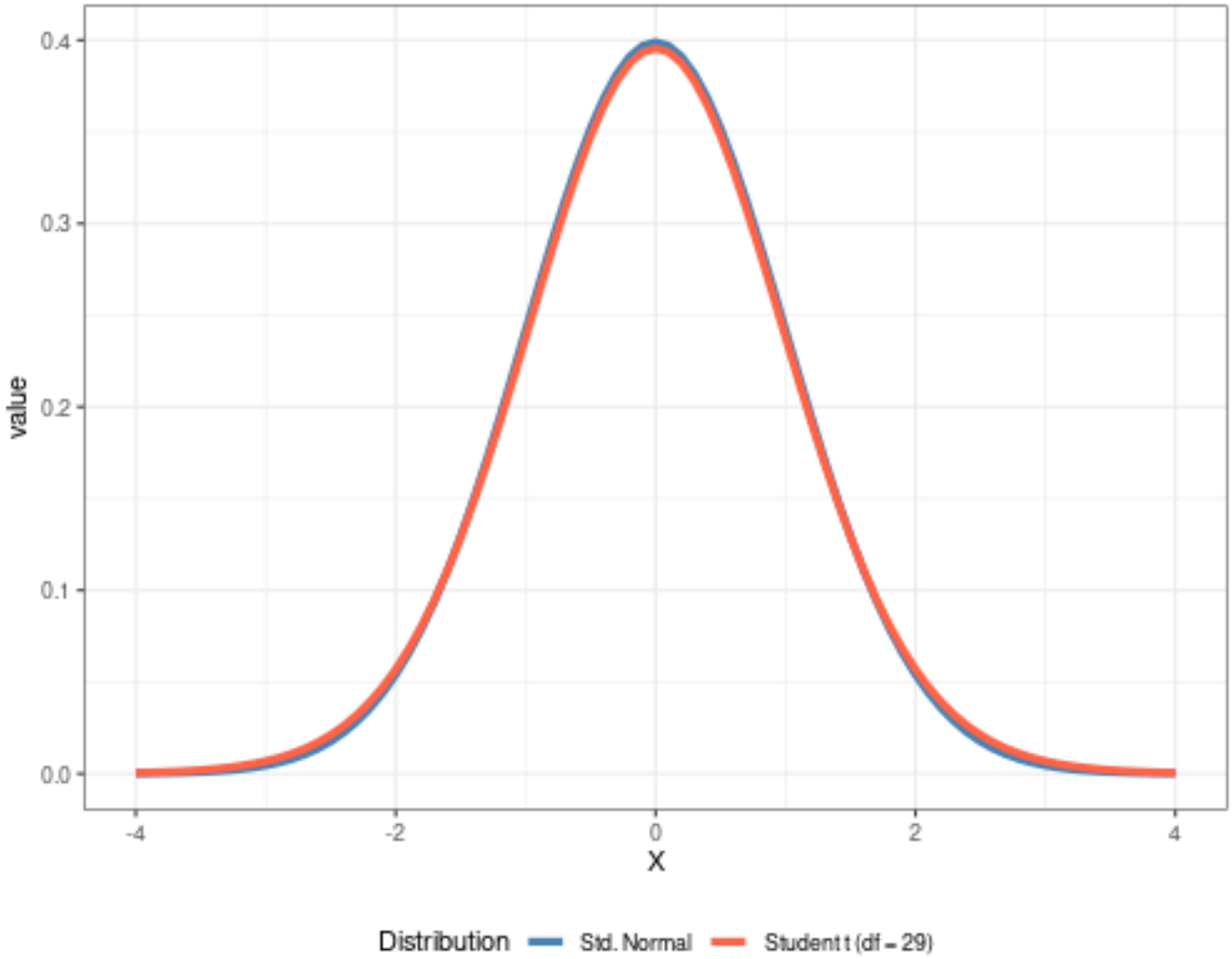
t-distribution with df=3



t-distribution with df=10

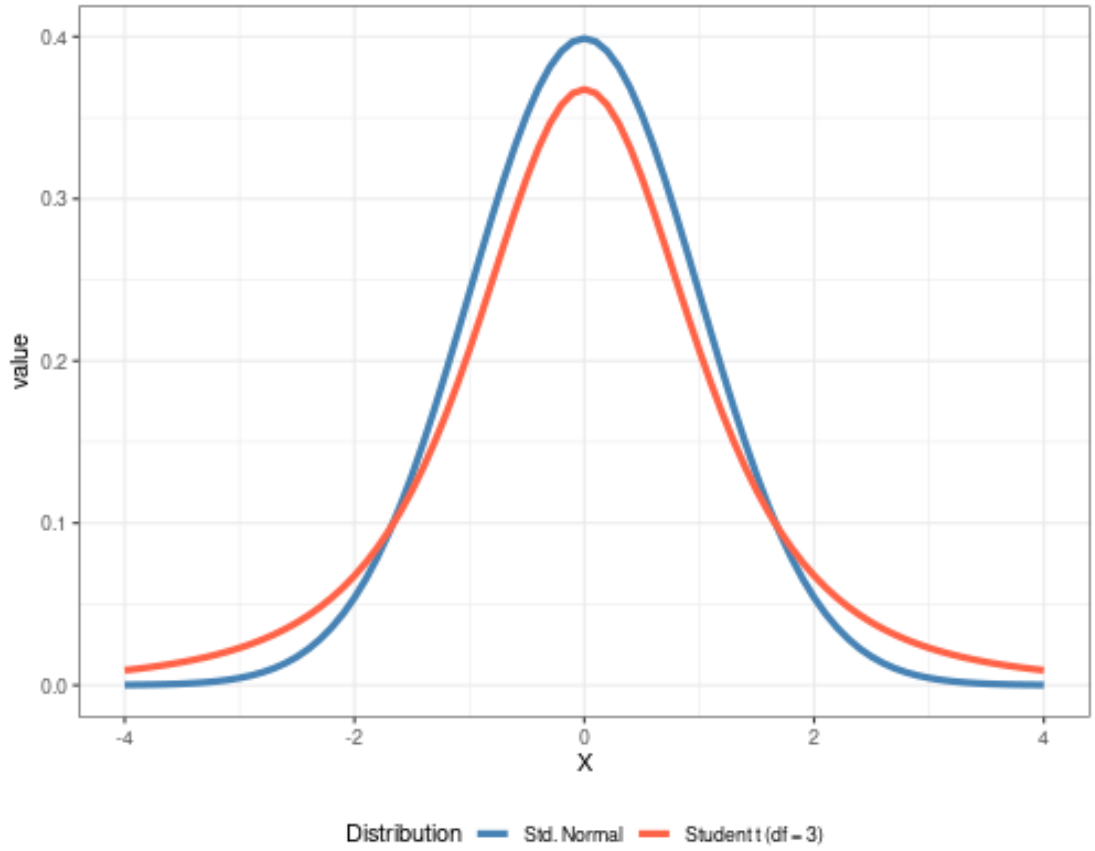


t-distribution with df=29



Implications?

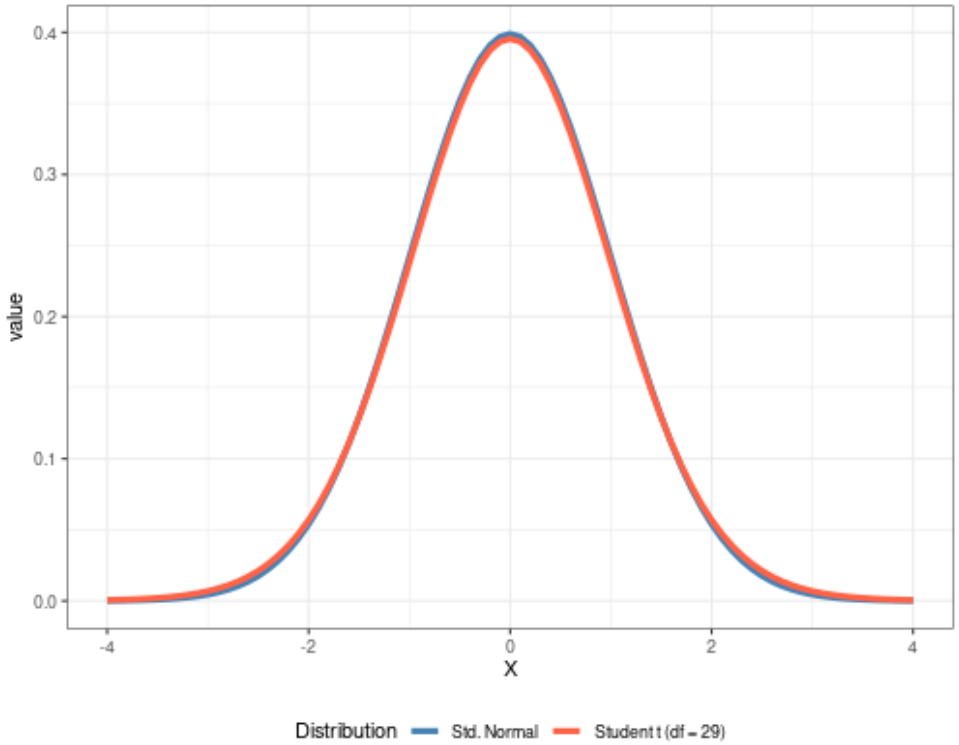
What are the implications of this for our confidence intervals?



- we can't just use $\pm 1.96 \times SE$
- we need to go out further because probability is more spread out

CI's using t-distribution

How do we go about making 95% CI's for the t-distribution?



We will use the t-table. Since we are finding the 'middle' 95% of the distribution, we use the "two tails" column with a value of 1-.95 = 0.05

- Std. Normal: $\pm 1.96 \times SE$
- t_{29} dist.: $\pm 2.04 \times SE$

CI's using t-distribution

How do we go about making 95% CI's for the t-distribution?
Instead of using:

$$\overline{x} \pm 1.96 \times \frac{s}{\sqrt{n}},$$

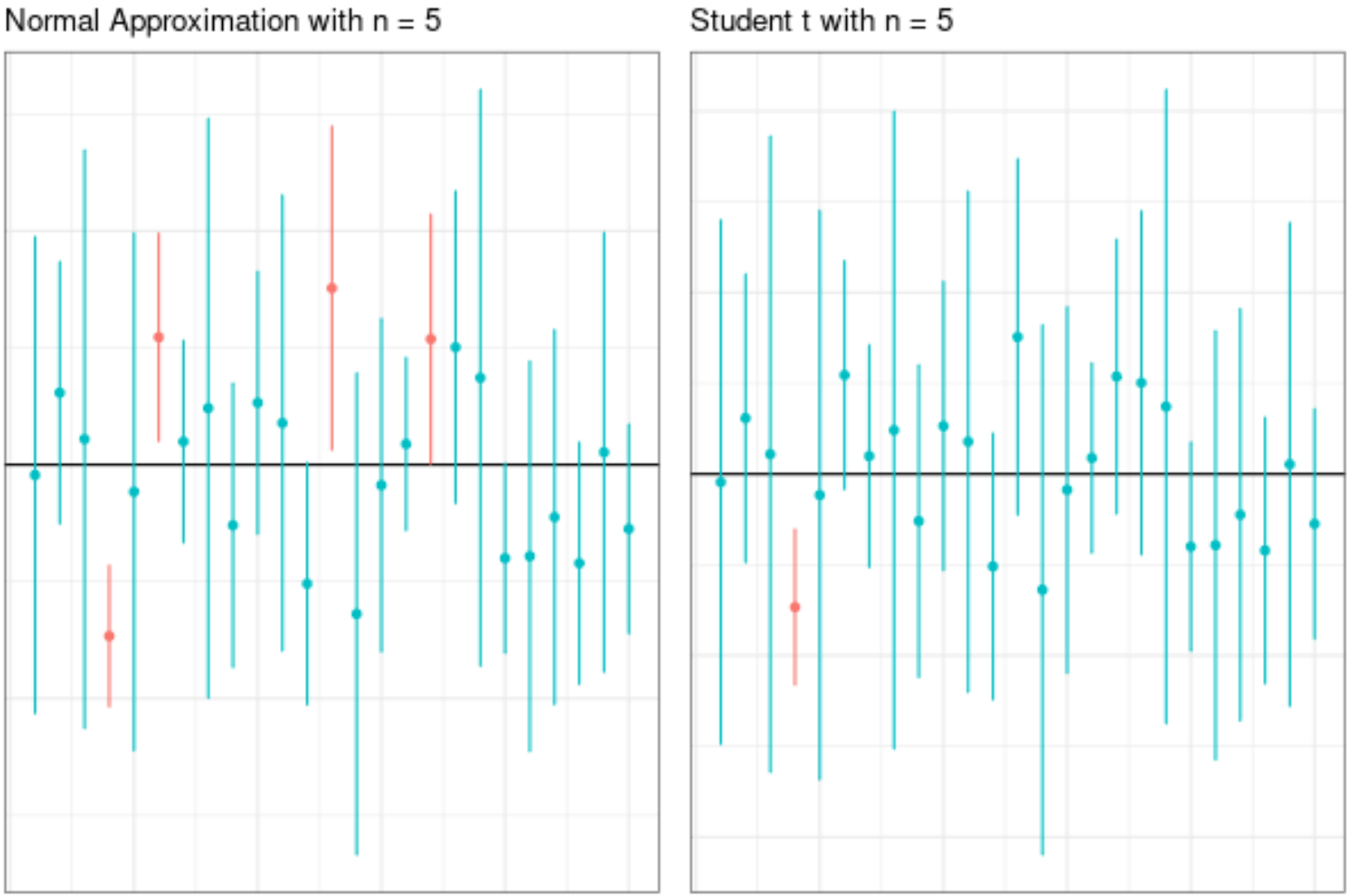
We will use:

$$\overline{x} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}},$$

- t^*_{n-1} is the value corresponding to 95% confidence and $n-1$ df
- use t-table to get this value

Comparing CI's between Normal and t-dist

Same data for both groups, CIs from t-distribution method are wider



Different % CI's

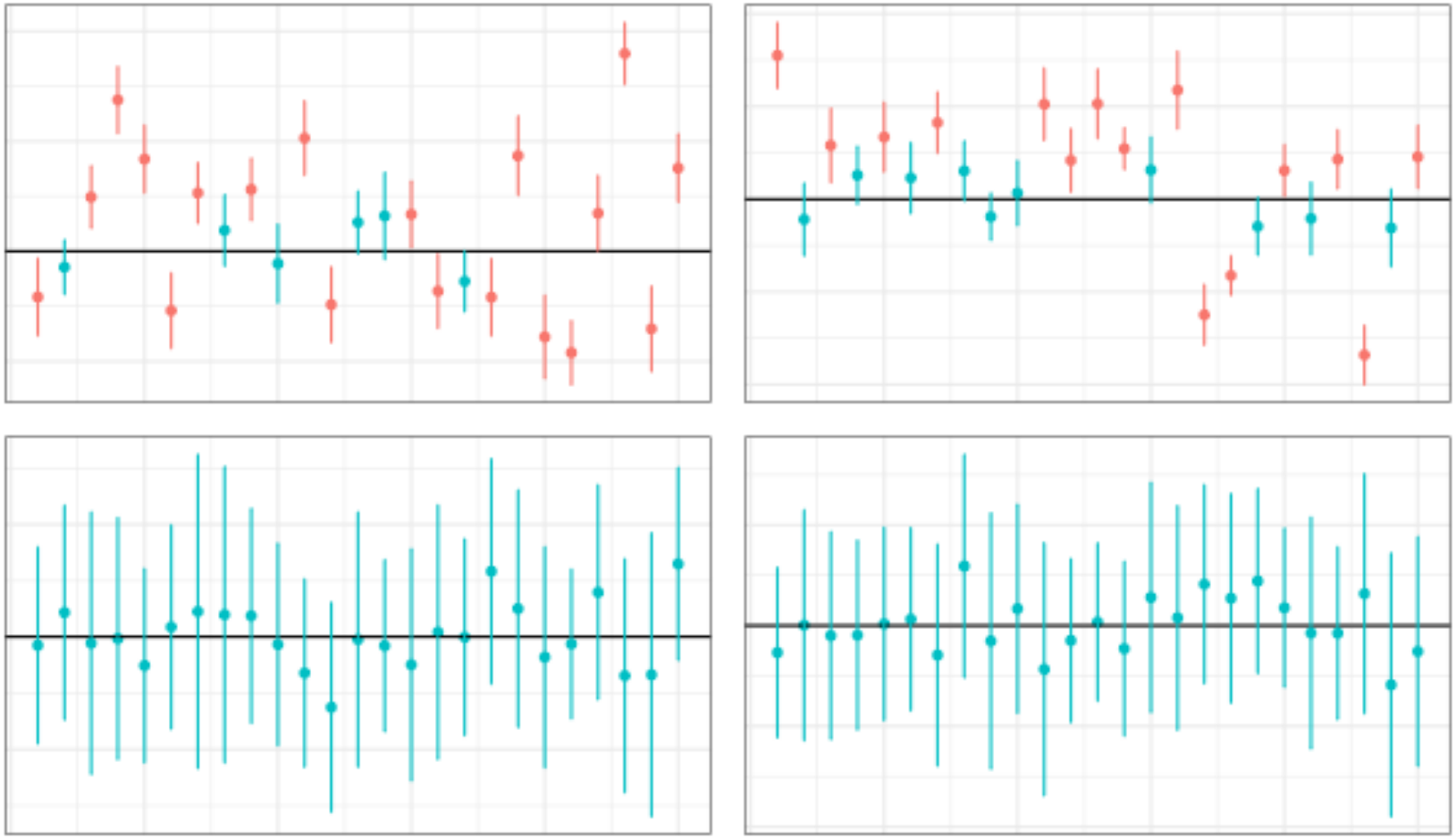
We may not always want a 95% confidence interval

- maybe we are OK with an 80% CI
- maybe we want a 99% CI
- remember that there is a trade-off between confidence % and how wide the interval is

What do we do? We just adjust the margin of error to account for the new confidence % we want.

Confidence and Width

There is a tradeoff between confidence and margin of error



Different % CIs

Let α denote the % of confidence intervals that will incorrectly estimate μ (as a decimal)

- then we can say we are trying to find a 100(1-α)% CI
- ex) for a 95% CI, we have $\alpha = .05$
- ex) for a 80% CI, we have $\alpha = .2$

To find the associated cutoffs on the t-distribution for a 100(1-α)% CI:

- calculate α value
- use "two tails" column directly with α value
- use "one tail" column with $\frac{\alpha}{2}$ value (similar to Normal distribution)

Summary of CIs for Pop. Mean

General Formula for CI (σ known):

$$\overline{x} \pm z^* \times \frac{\sigma}{\sqrt{n}},$$

where z^* corresponds to confidence level

General Formula for CI (σ unknown):

$$\overline{x} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}},$$

where t^*_{n-1} corresponds to confidence level and df