Student's t

Grinnell College

October 28, 2024

Starting Questions

1. Suppose you have a sample with a mean value of 0. What would happen to the mean if you added 10 to all of the observations. What would happen to the standard deviation?
2. Suppose you had a sample with a mean value of 0. What would happen to the mean if you multiplied all of the values by 10. What would happen to the standard deviation
3. Think of the 95% confidence interval for a normal distribution,

$$\overline{x} \pm 2 \times \frac{\sigma}{\sqrt{n}} = \left(\overline{x} - 2 \times \frac{\sigma}{\sqrt{n}}, \ \overline{x} + 2 \times \frac{\sigma}{\sqrt{n}} \right)$$

what quantiles of the distribution would each endpoint represent?

Starting Questions 2

1. Do quantiles exist for all data, regardless of their distribution? If so, how does the 0.05 quantile for a normal distribution compare to a 0.05 quantile of a skewed distribution?
2. Explain why for an 80% confidence interval we need to use the quantiles 0.1 and 0.9 as the critical values
3. Suppose that I have two datasets, each with 100 observations
   - Dataset A has a mean of $\mu_A = 50$ and $\sigma_A = 5$
   - Dataset B has a mean of $\mu_B = 25$ and $\sigma_B = 10$
     1. How many observations should fall between the 0.1 and 0.9 quantiles for A?
     2. How many observations should fall between the 0.1 and 0.9 quantiles for B?
     3. If I find the sample mean for each dataset, $\overline{x}_A$ and $\overline{x}_B$, which sample mean will have a larger 95% confdience interval?
     4. If I find the sample mean for each dataset, $\overline{x}_A$ and $\overline{x}_B$, which confidence interval will be centered further to the right?

Review

A **sampling distribution** refers to the distribution of a sample statistic (i.e., $\overline{x}$) if we were to repeatedly sample from a population and recompute the statistic

- What values would they take?
- How frequently would they appear?

The **Law of Large Numbers** guarantees that, as the number of observations n in my sample increases, my estimate of the parameter will converge to the true value

The **Central Limit Theorem** states that is my statistic is an average or a proportion, then the sampling distribution of my statistic will be approximately normal, with

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

Big Goals for Today

- Standard Normal
- Confidence intervals and quantiles
- t-distribution

Standardization

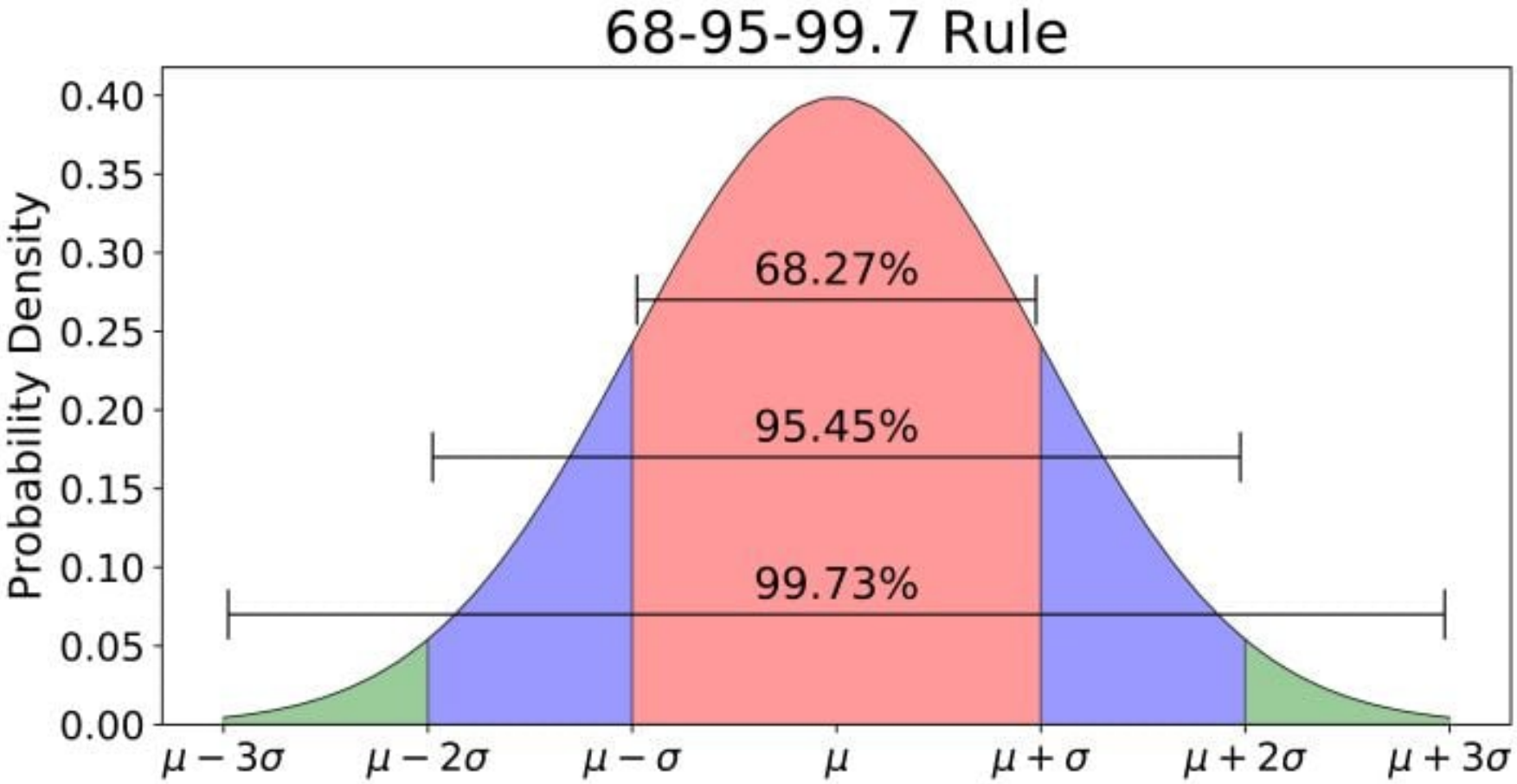
Previously we saw that

$$Z = \frac{X - \mu}{\sigma}$$

would create a standardized variable with mean 0 and sd 1. From the CLT, what would we need to do to $\overline{x}$ to standardize it?

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

We call this a **standard normal** distribution


68-95-99.7 Rule
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
68.27%
95.45%
99.73%
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Quantiles

What would be the quantiles for the 95% confidence interval?

Consider an expression written like:

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

This C term, called a **critical value** gives me my relationship between percentiles and standard deviation for normal.

For example, based on the empirical rule on the previous slide, we know that setting $C = 2$ will cover 95.45% of our distribution

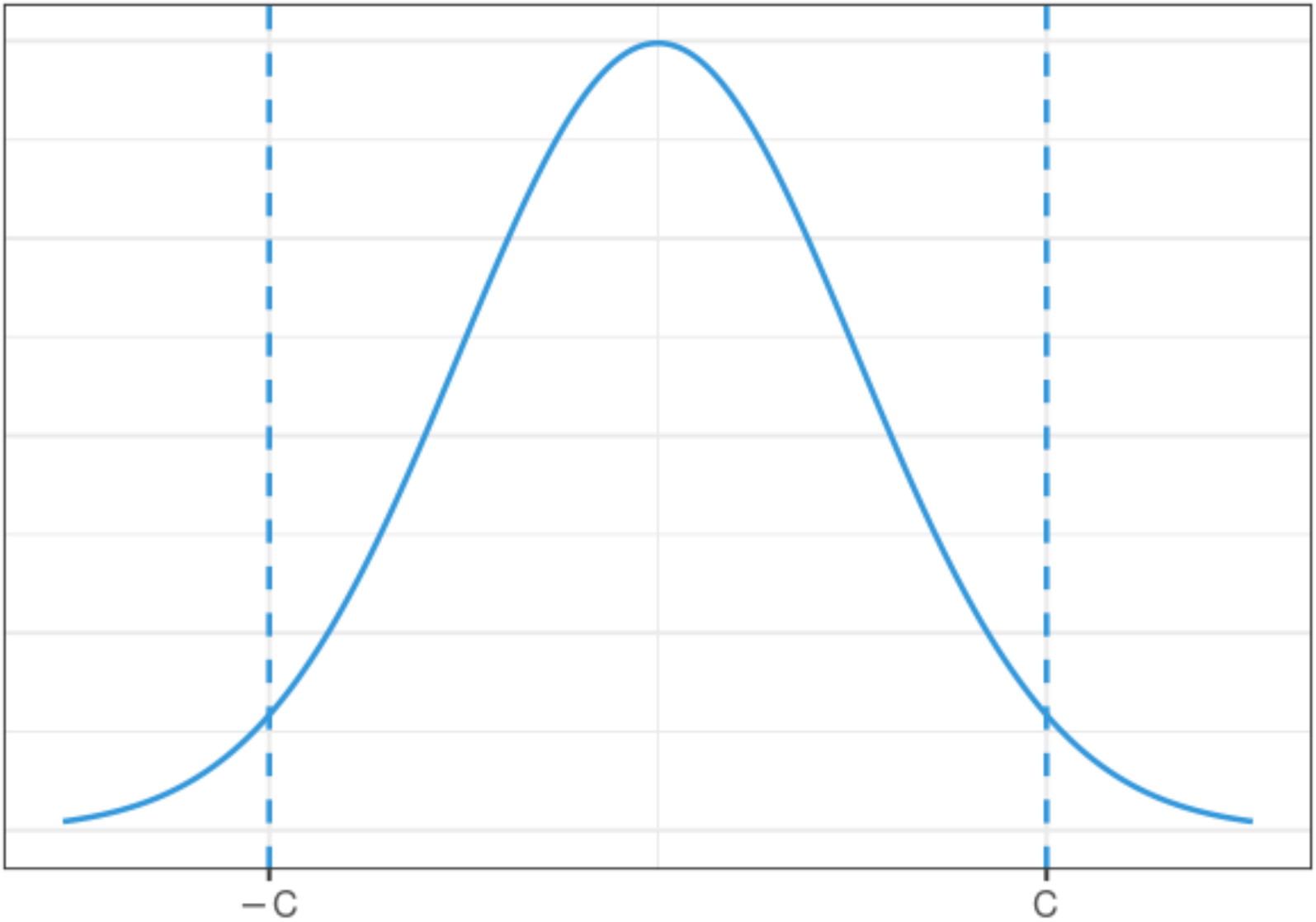
We can find the values for C by considering a standard normal distribution where $\mu = 0$ and $\sigma = 1$

$$\begin{aligned} \mu \pm C \times \sigma &= 0 \pm C \times 1 \\ &= \pm C \\ &= (-C, C) \end{aligned}$$

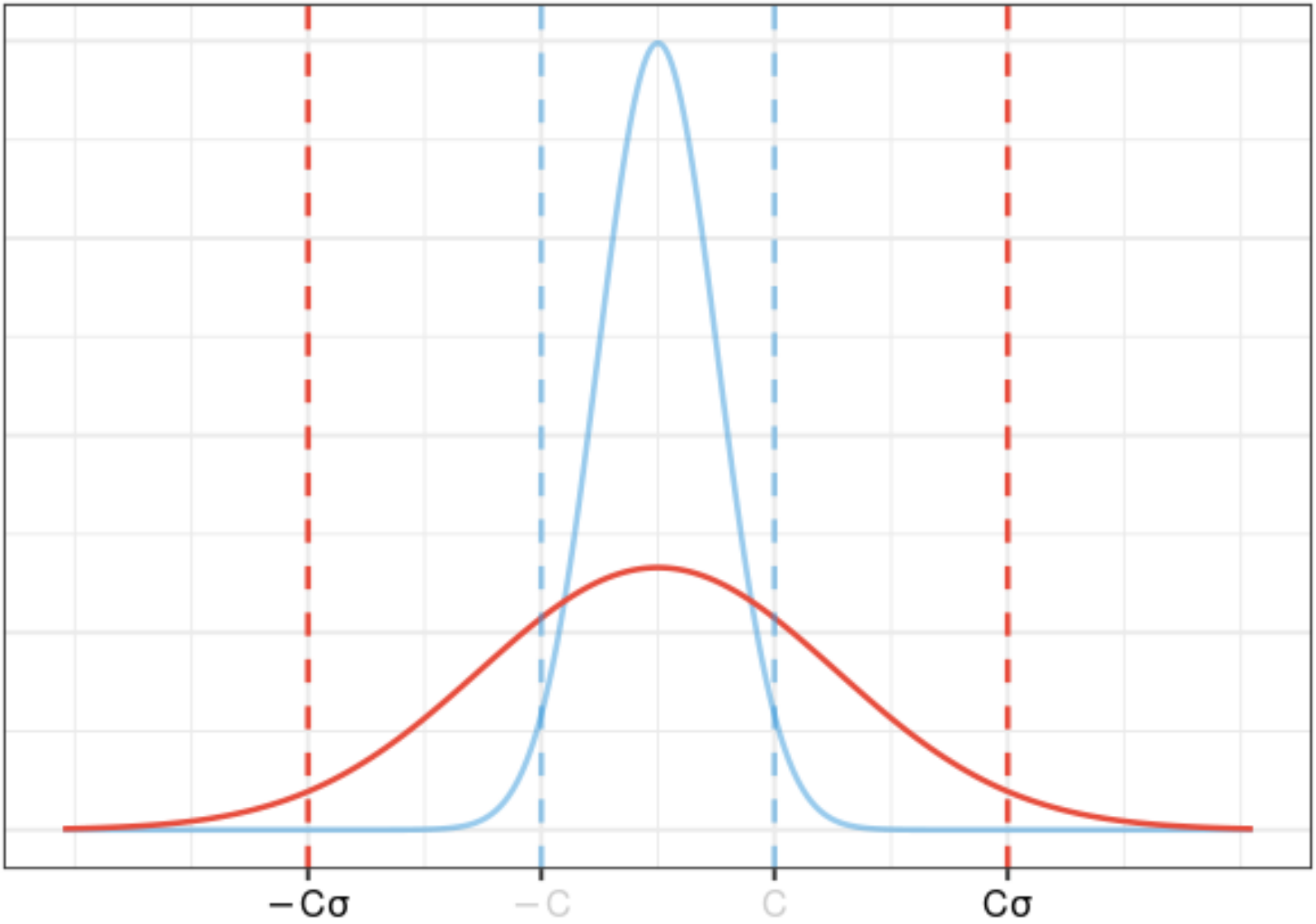
If we want an $m\%$ confidence interval, then, we must choose the values of C such that the interval $(-C, C)$ covers the middle $m\%$ of a standard normal distribution

For a 95% confidence interval, then, that means we need the 0.025 and 0.975 quantiles (i.e., this is the same 2.5th and 97.5th percentiles)

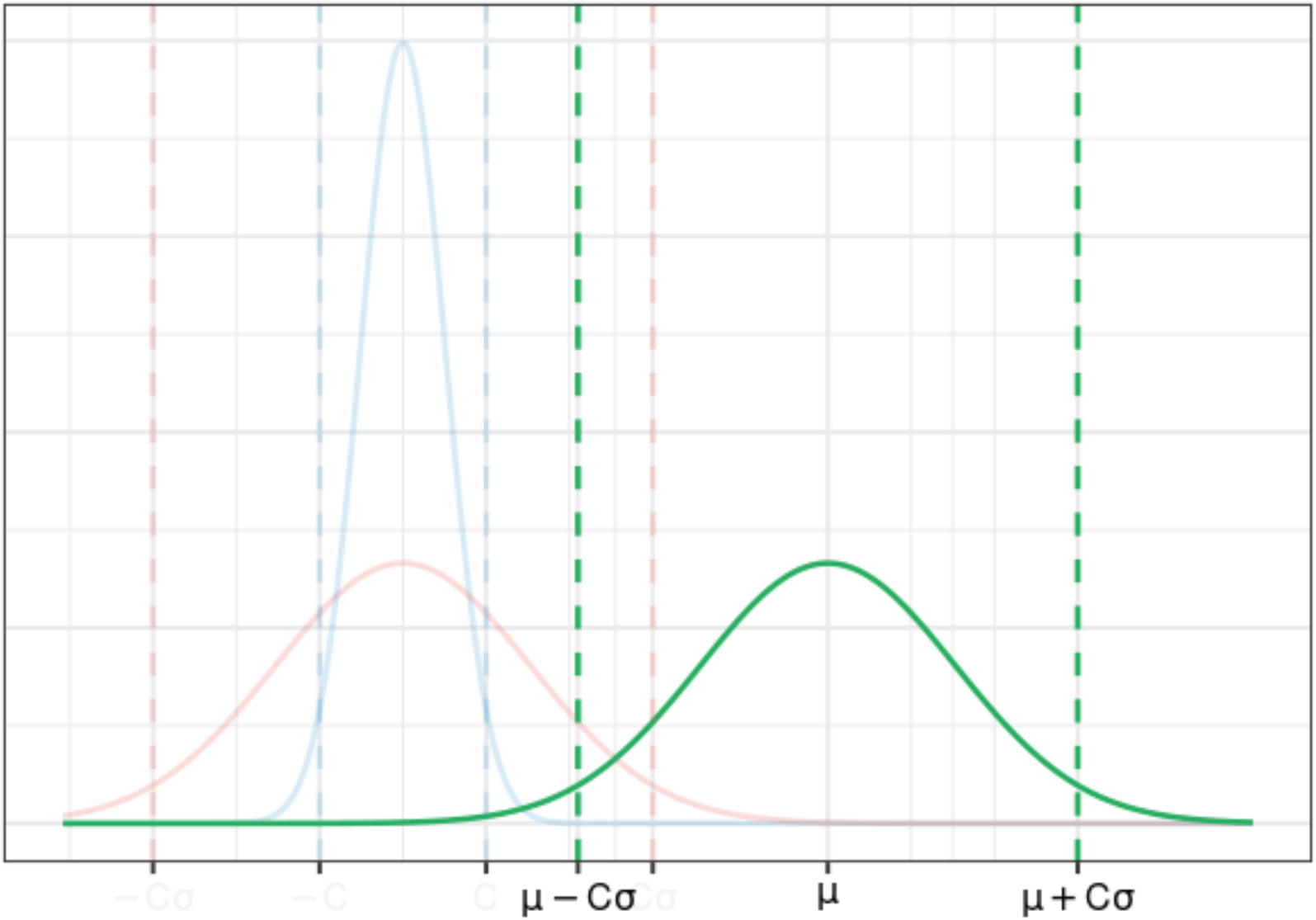
$$X \sim N(0, 1)$$



$$X \sim N(0, \sigma)$$



$$X \sim N(\mu, \sigma)$$



Quantiles

We can find the quantiles of a normal distribution with the R function `qnorm` (for **q**uantile of **norm**al) which takes as arguments the quantiles we want, as well as the mean and the standard deviation of the distribution:

```
> quants <- c(0.025, 0.975)

> qnorm(quants, mean = 0, sd = 1)
[1] -1.96  1.96
```

This means that for a *true* 95% confidence interval, we should be using

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

Example

To illustrate, consider our penguin dataset, where we consider the flipper length (mm) of male gentoo penguins. Our summary statistics give us the following:

$$\overline{x} = 199.94, \quad \hat{\sigma} = 5.9766, \quad n = 34$$

To find a 95% confidence interval, we could use our formula, $\overline{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}}$ *or* we could use the `qnorm` function, passing in the mean and standard error from the CLT:

```
## Using qnorm function
> qnorm(c(0.025, 0.975), mean = 199.91, sd = 5.9766 / sqrt(34))
[1] 197.90 201.92

## Using our formula
> 199.91 + c(-1.96, 1.96)*(5.9766 / sqrt(34))
[1] 197.90 201.92
```

Note

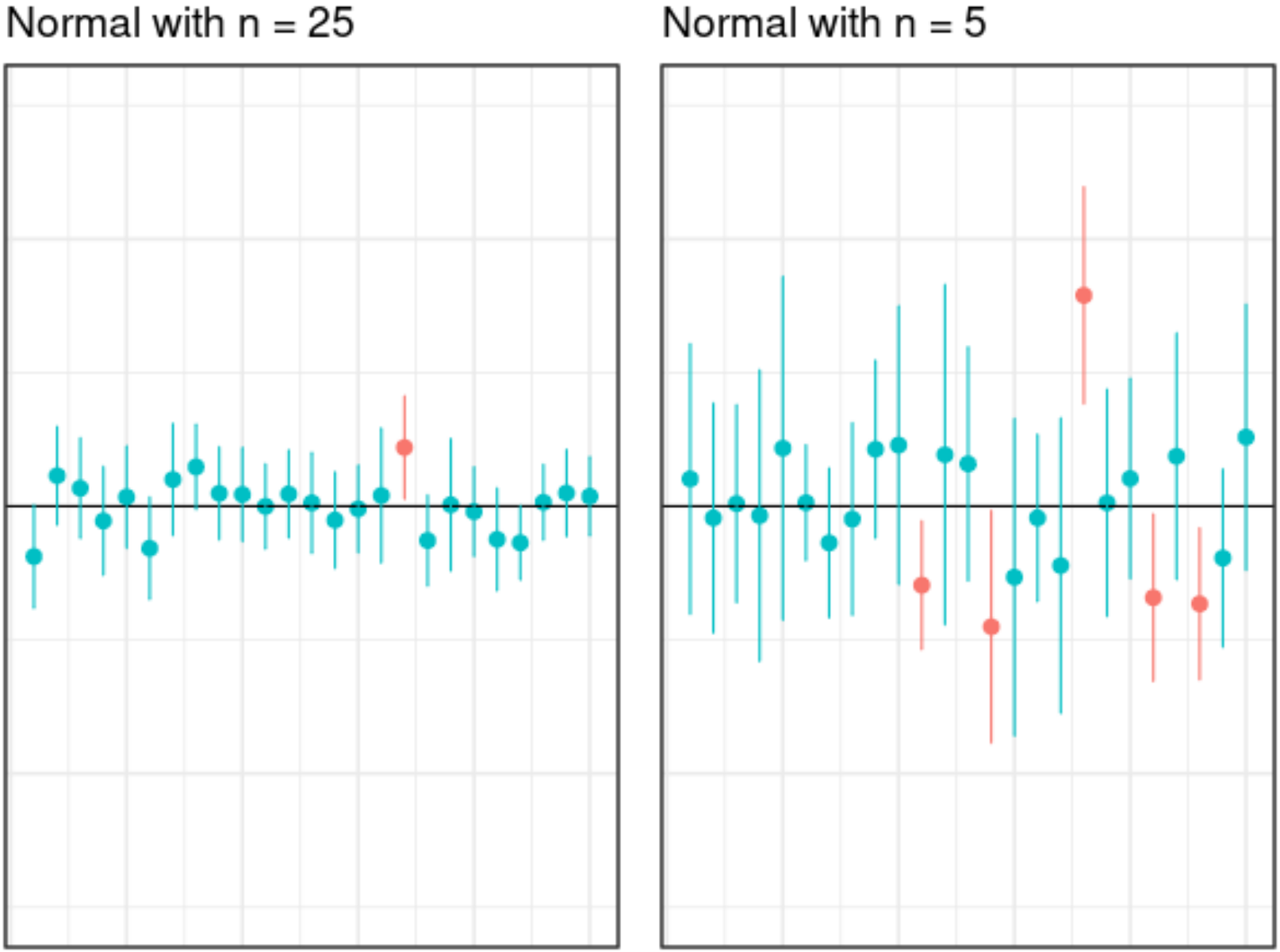
It is worth noting – the values for our critical value for the normal distribution will be the same *regardless* of the sample mean and sample standard deviation

This is because they are based on the theoretical standard normal distribution

Last last week, we examined how the CLT is an *approximation* that gets better as n increases

Especially when the population is skewed, larger values of n are necessary for our approximations to be useful

However, even when the population looks approximately normal, there are other issues that come about when our value for n is small


Normal with n = 25
Normal with n = 5

Estimating Variance

The problem we have lies in our estimation of σ

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

- If we knew σ precisely, the standard deviation of our *population*, we would have no issue in computing confidence intervals
- If we had enough observations in our sample to estimate σ, we would likewise run into few problems

$$\overline{X} \pm C \times \left(\frac{\sigma}{\sqrt{n}}\right) \qquad \text{vs} \qquad \overline{X} \pm C \times \left(\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

Estimating Variance

The problem we have lies in our estimation of σ

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

- If we knew σ precisely, the standard deviation of our *population*, we would have no issue in computing confidence intervals
- If we had enough observations in our sample to estimate σ, we would likewise run into few problems

What we need, then, is a way to incorporate our uncertainty about σ into the confidence intervals we construct around $\overline{x}$

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

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 "fatter 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"

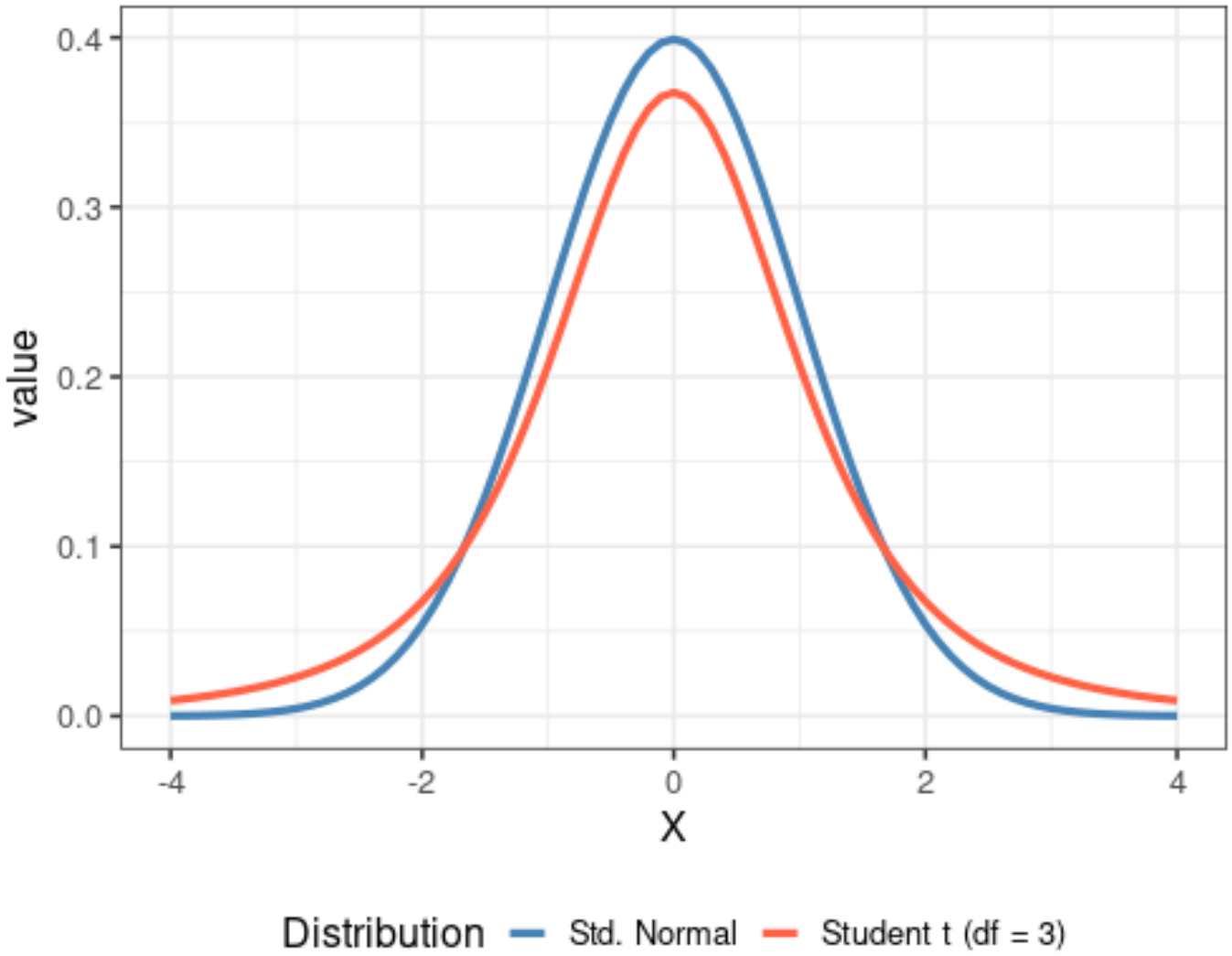
Student's t-distribution

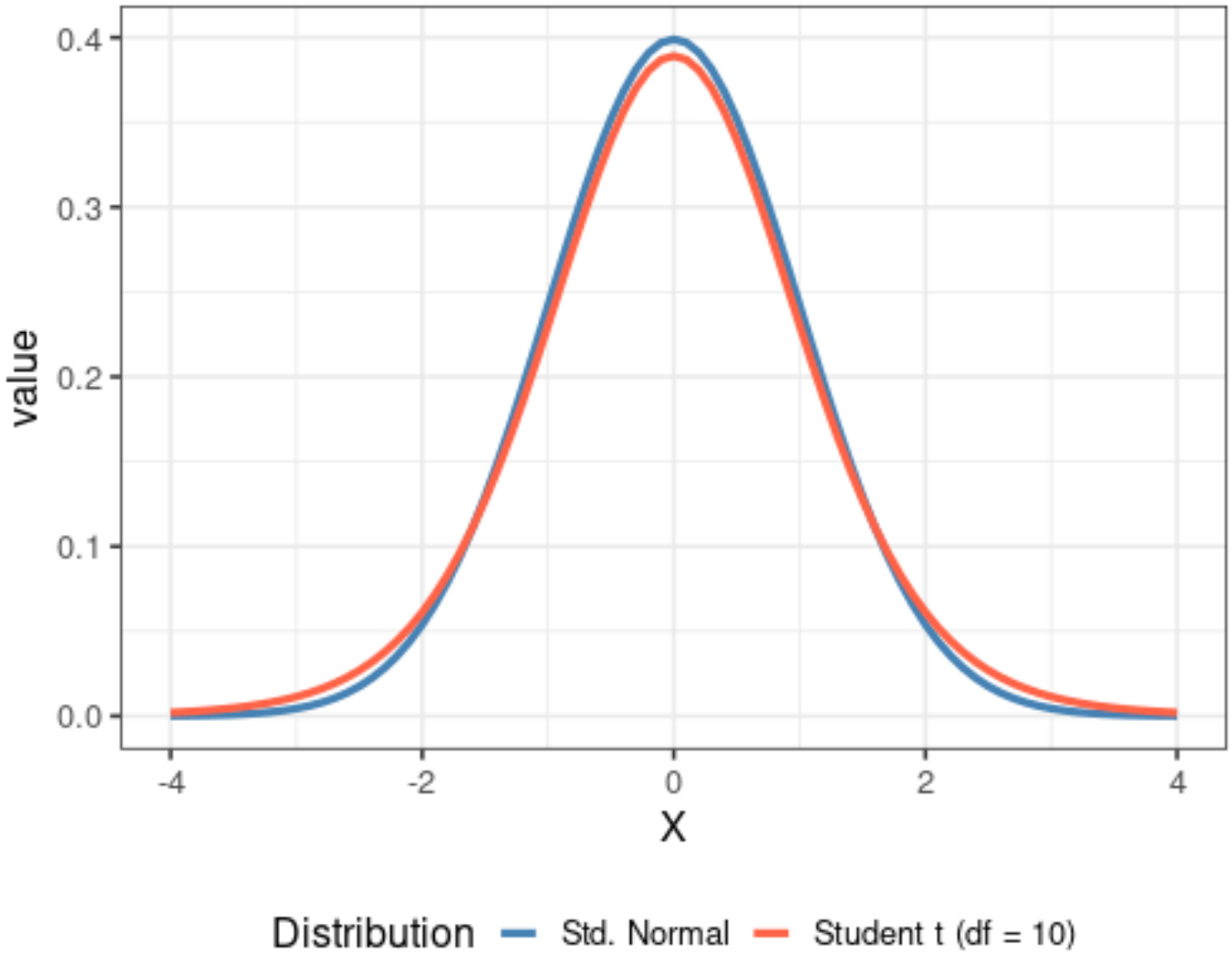
Student's t Distribution:

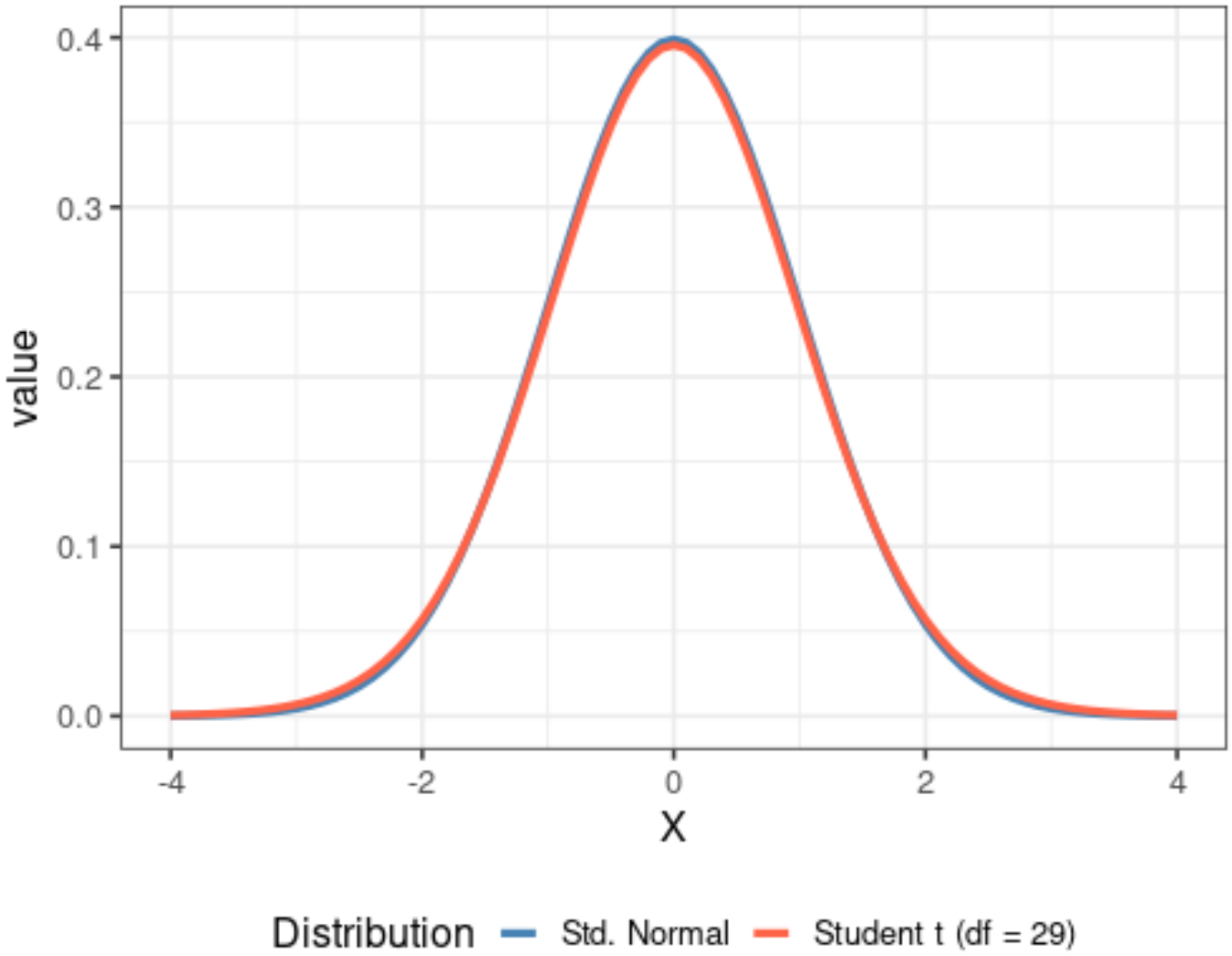
$$t = \frac{\overline{x} - \mu}{\hat{\sigma}/n}$$

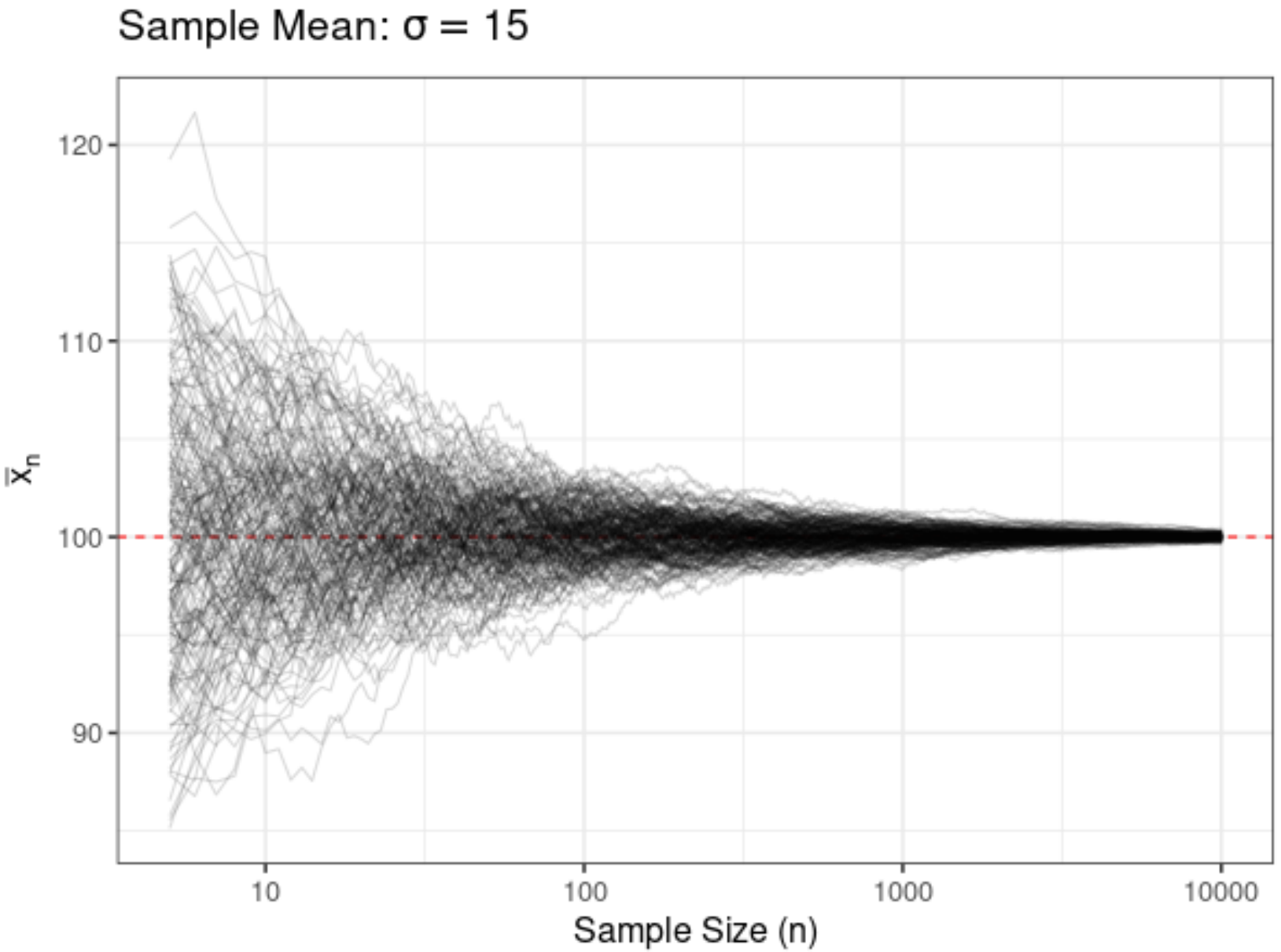
$$t \sim t(n-1)$$

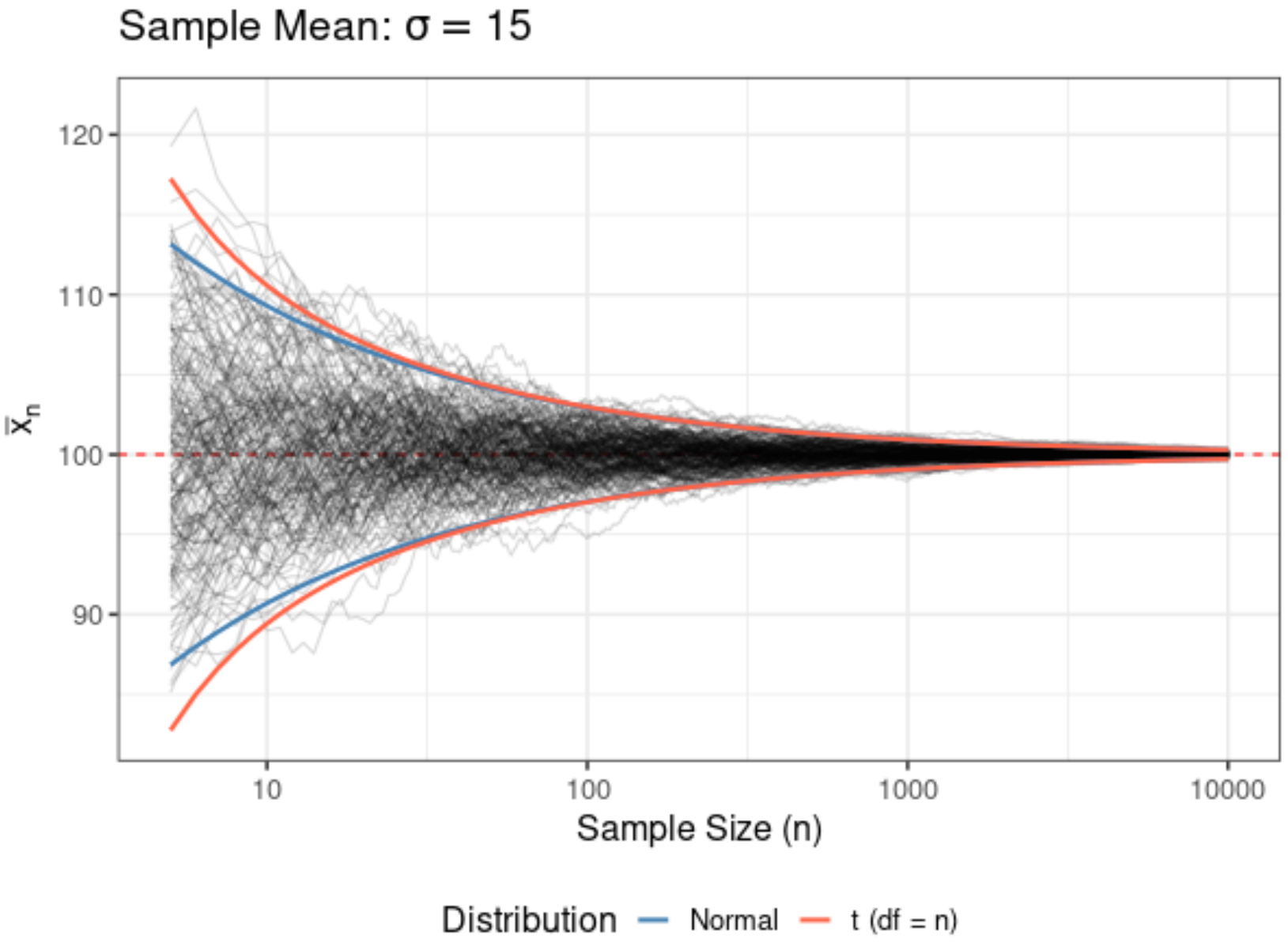
1. The t distribution has only one *distributional parameter* called the *degrees of freedom*, equal to $n - 1$
2. The t distribution has "fatter tails" than the normal distribution, allowing for the possibility of larger values
3. The t distribution will become standard normal as $n \to \infty$


0.4
0.3
0.2
0.1
0.0
value
-4
-2
0
2
4
X
Distribution
Std. Normal
Student t (df = 3)


0.4
0.3
0.2
0.1
0.0
value
-4
-2
0
2
4
X
Distribution
Std. Normal
Student t (df = 10)


0.4
0.3
0.2
0.1
0.0
value
-4
-2
0
2
4
X
Distribution
Std. Normal
Student t (df = 29)


Sample Mean: σ = 15
120
110
100
90
$\bar{x}_n$
10
100
1000
10000
Sample Size (n)


Sample Mean: σ = 15
120
110
100
90
$\bar{x}_n$
10
100
1000
10000
Sample Size (n)
Distribution
Normal
t (df = n)

Just as with normal, we can use a quantile function to find the quantiles of the t-distribution (this is true of every distribution, btw)

```
> quants <- c(0.025, 0.975)
> qt(quants, df = 5)
[1] -2.5706  2.5706

> qt(quants, df = 25)
[1] -2.0595  2.0595

> qt(quants, df = 100)
[1] -1.984  1.984

> qnorm(quants)
[1] -1.96  1.96
```

What happens as the degrees of freedom increases?

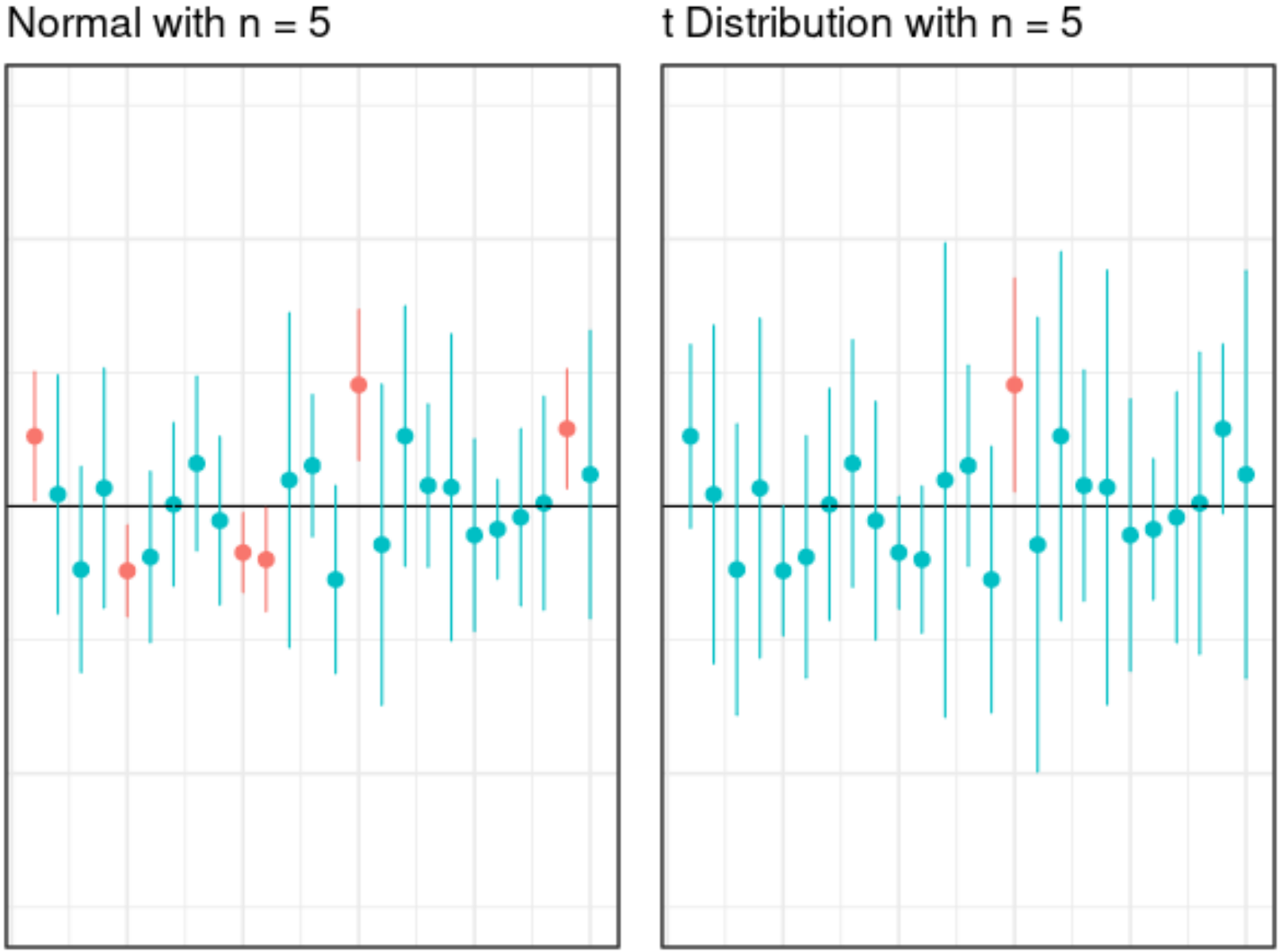
```
> quants <- c(0.025, 0.975)
> qt(quants, df = 5)
[1] -2.5706  2.5706

> qnorm(quants)
[1] -1.96  1.96
```

If, for example, I wanted to find a 95% confidence interval of a t distribution with $n - 1 = 5$ degrees of freedom, I would need

$$\overline{x} \pm 2.5706 \times \frac{\hat{\sigma}}{\sqrt{6}}$$

As opposed to

$$\overline{x} \pm 1.96 \times \frac{\hat{\sigma}}{\sqrt{6}}$$


Normal with n = 5
t Distribution with n = 5

Example

We can return again to our penguin example, this time considering using quantiles from the t-distribution

$$\overline{x} = 199.94, \quad \hat{\sigma} = 5.9766, \quad n = 34$$

We can find the critical values for a 95% CI using the `qt()` function:

```
> quants <- c(0.025, 0.975)
> qt(quants, df = 34-1)
[1] -2.0345  2.0345
```

From this, we can construct

$$\begin{aligned} \overline{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}} &= 199.94 \pm 2.0345 \times (5.9766/\sqrt{34}) \\ &= (197.82, 202.00) \end{aligned}$$

Compare this with our estimate using the normal distribution:

$$\overline{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}} = (197.90, 201.92)$$

Example

Consider the `trees` data in R, investigating the height of $n = 31$ felled cherry trees.

We were able to find the following summary statistics:

$$\overline{x} = 13.25, \quad \hat{\sigma} = 3.14$$

We can find the critical values for a 95% CI using the `qt()` function:

```
> quants <- c(0.025, 0.975)
> qt(quants, df = 31 - 1)
[1] -2.0423  2.0423
```

From this, we can construct

$$\begin{aligned} \overline{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}} &= 13.25 \pm 2.0423 \times (3.14/\sqrt{31}) \\ &= (10.913, 15.587) \end{aligned}$$

Review

Big day today:

- We can standardize sampling distribution to derive the **standard normal distribution**
- By assuming a distribution, we can use quantiles to determine our **critical values** for constructing confidence intervals
- Our estimation of $\hat{\sigma}$ necessitates accomodating extra uncertainty in our estimates of $\overline{x}$
- The **t-distribution** is such a distribution; it is centered at zero and has **degrees of freedom** as its only distributional parameter
- The `qnorm()` and `qt()` functions in R allow us to derive quantiles from these distributions