Student's t

Grinnell College

March 28, 2025

Warm-up

On Wednesday, we found that the flipper length of male gentoo penguins had the following statistics from a sample of size $n = 34$:

$$\overline{x} = 199.94, \quad \hat{\sigma} = 5.97$$

- What is the standard error for the sampling distribution of $\overline{x}$?
- What are the critical values for a 95% CI based on this sample?
- Suppose that somebody claims that the true average length of the male gentoo penguin flipper is $\mu_0 = 202$:
  - Find the standardized value of this proposed mean using $\overline{x}$ as our best guess for the true mean
  - Compare the standardized value to the critical value found above. What does this say about the plausibility of the proposed mean?

Review

percentiles and standard errors from normal
use qnorm to find these from standard normal. critical value comes from std normal when sigma $= 1$
specifically, if i see a standardized value of, say, 2.12, i know this is outside range of middle 95

Big Goals for Today

- t-distribution

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:

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

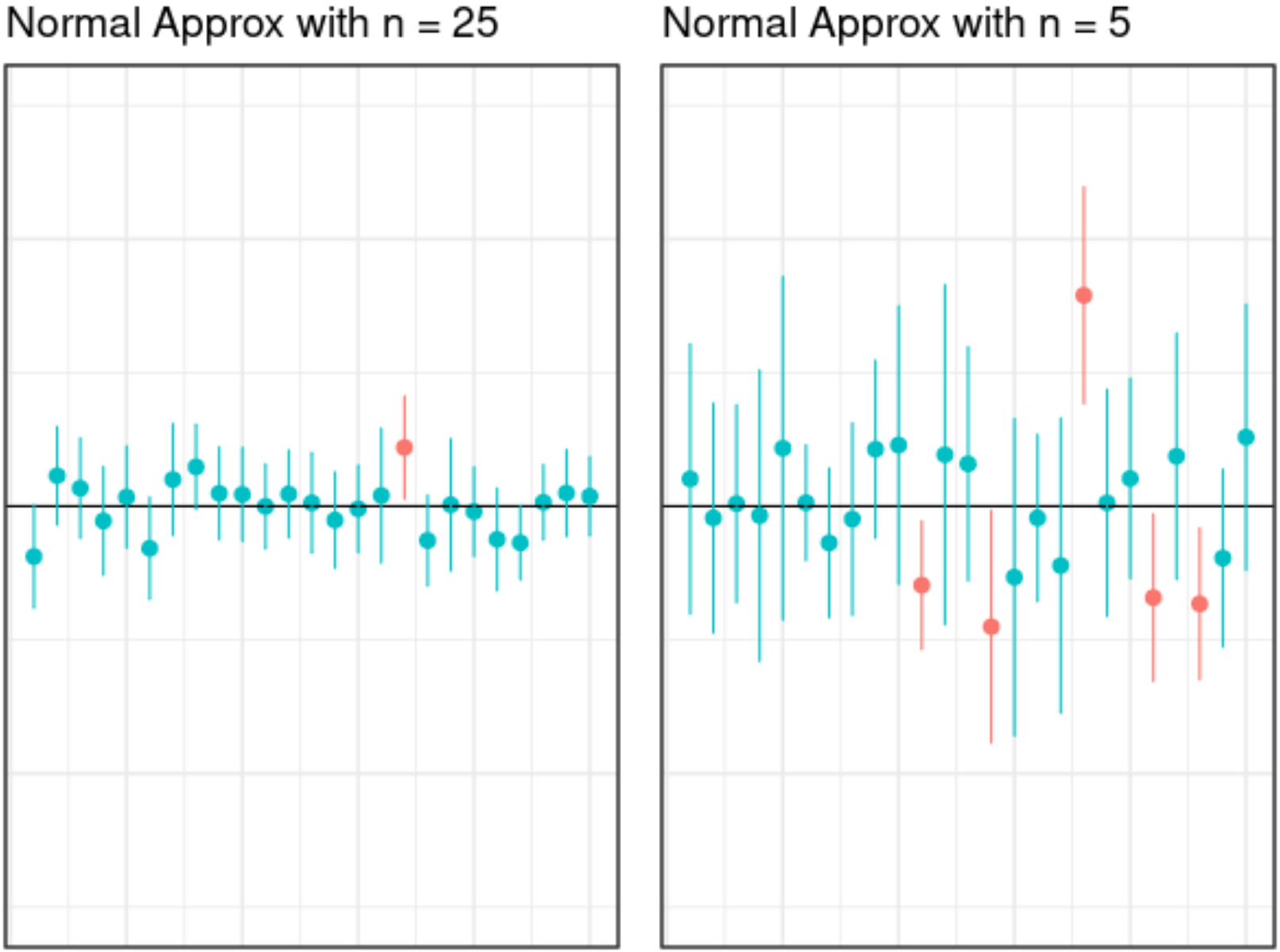
To find a 95% confidence interval, we could use our formula, $\overline{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}}$ *or* we could use the `qnorm` function, passing in the mean and standard error from the CLT:

```
## 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
```

In our funsheet, 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 Approx with n = 25
Normal Approx with n = 5

Important to Note

Notice that:

- The points are closer to the line when $n = 25$
- Thelength of the bands are larger when $n = 5$
- The critical value for each of these is the same

It is absolutely critical that we understand that the differences in these first two points are *exclusively* the consequence of sample size, which directly impacts the standard error for the sample mean

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 with $\hat{\sigma}$, we would likewise run into few problems
- When our sample size is smaller, we *overestimate* how certain we are about our estimation of σ

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

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"

t Statistic

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

The **t statistic** arises when we standardize our sample mean using $\hat{\sigma}$, our estimate of the population standard deviation, rather than the true (usually unknown) value, σ

The sampling distribution of the t-statistic is known as the t-**distribution**

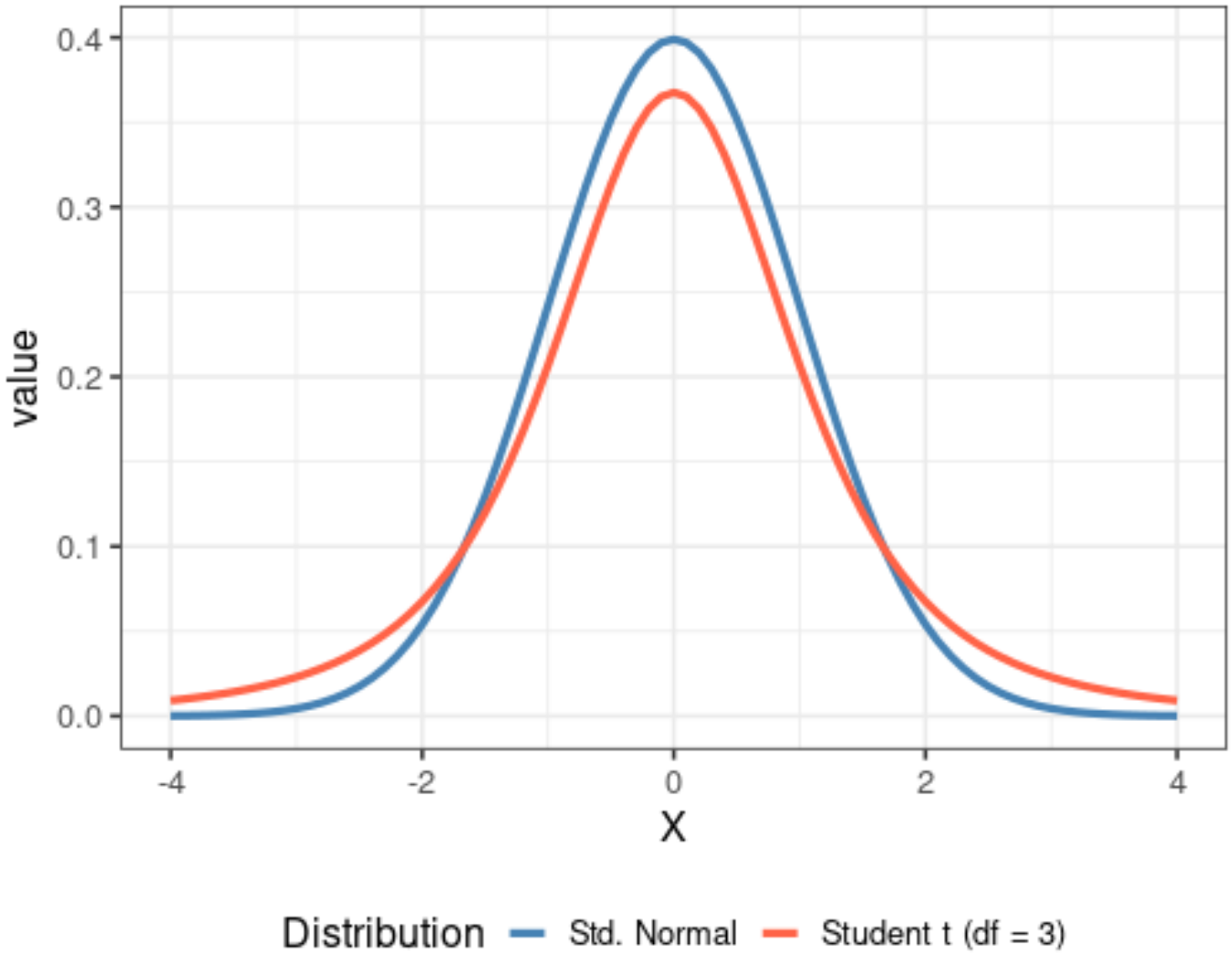
Student's t-distribution

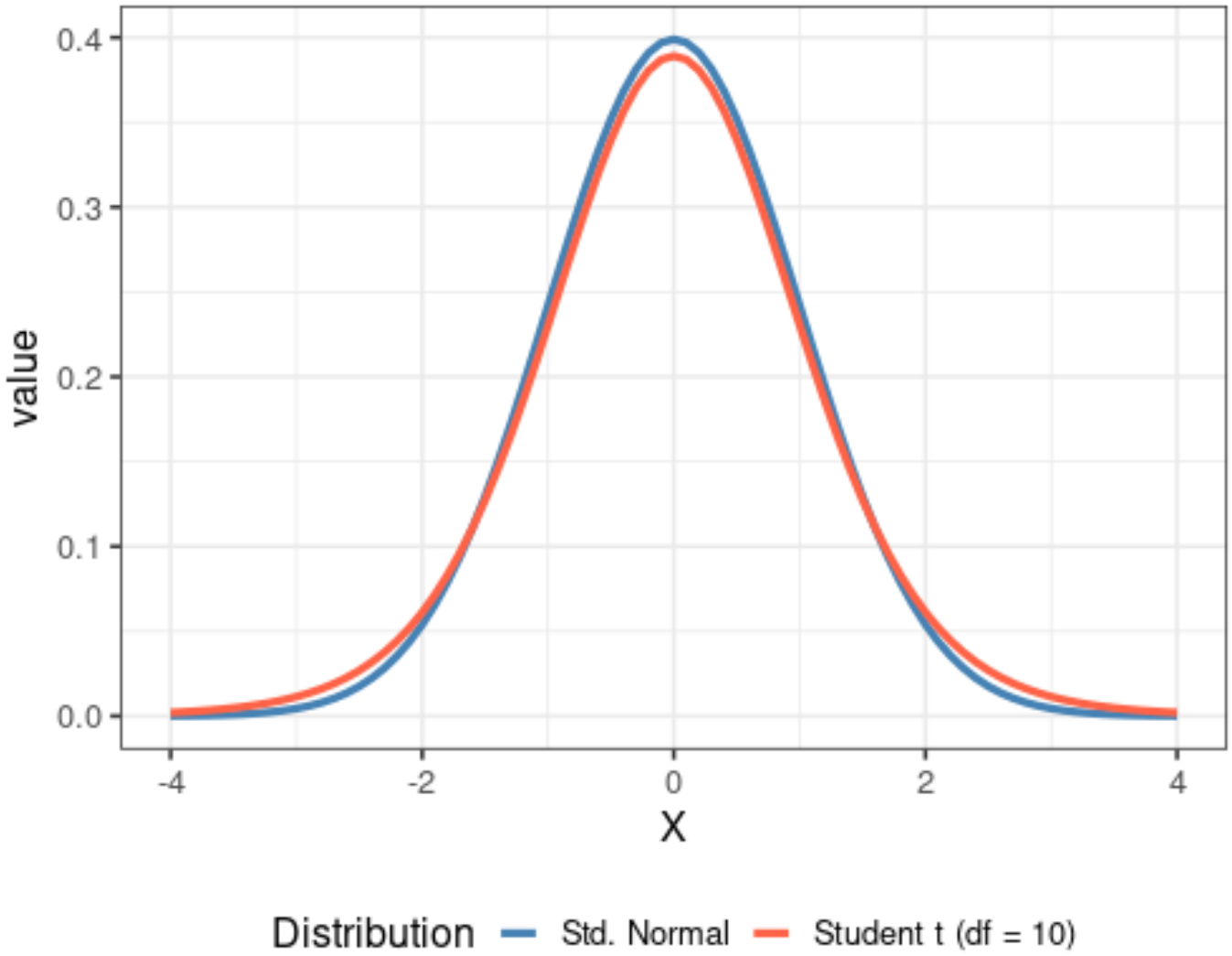
Student's t Distribution:

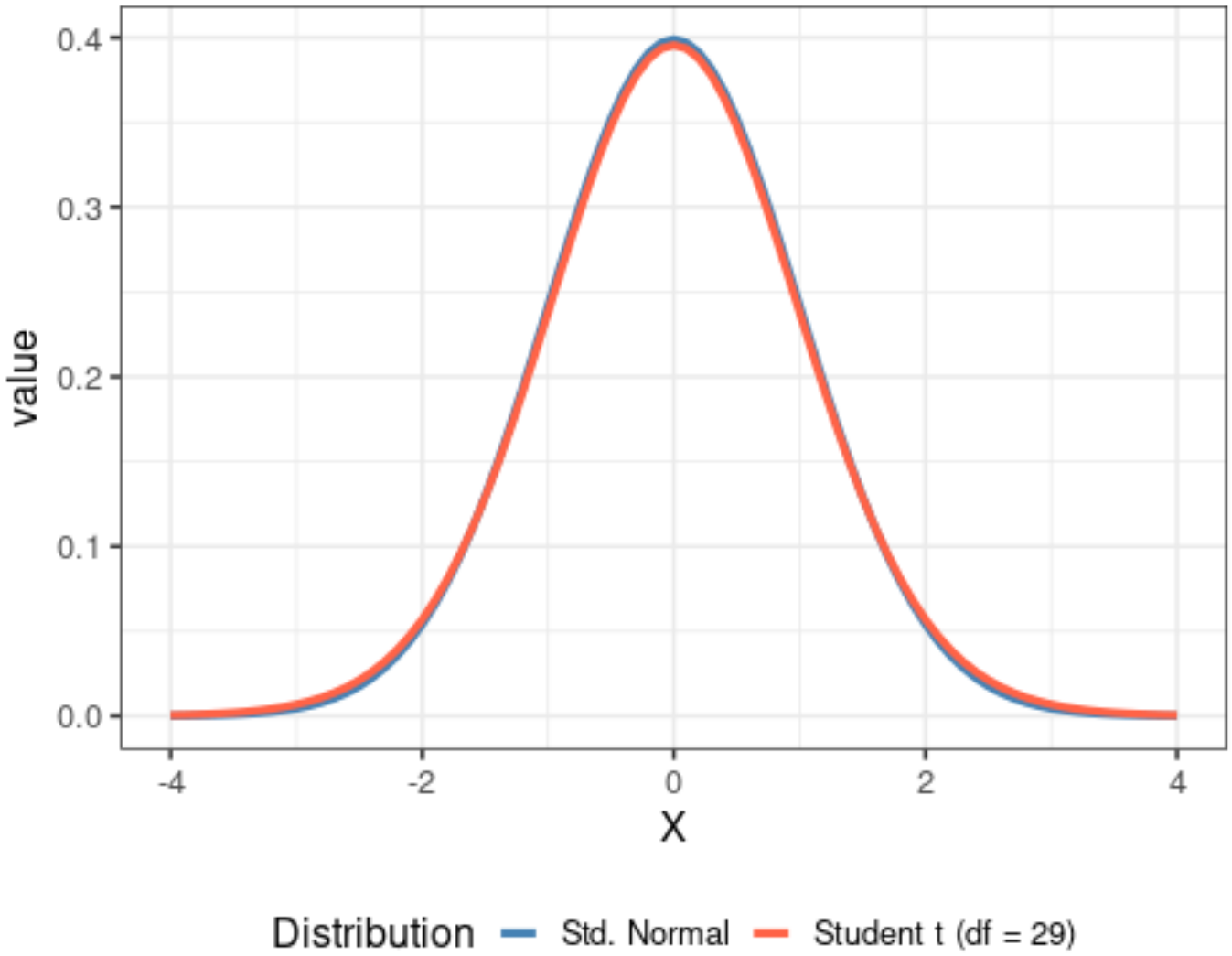
$$t = \frac{\overline{x} - \mu}{\hat{\sigma}/\sqrt{n}}$$

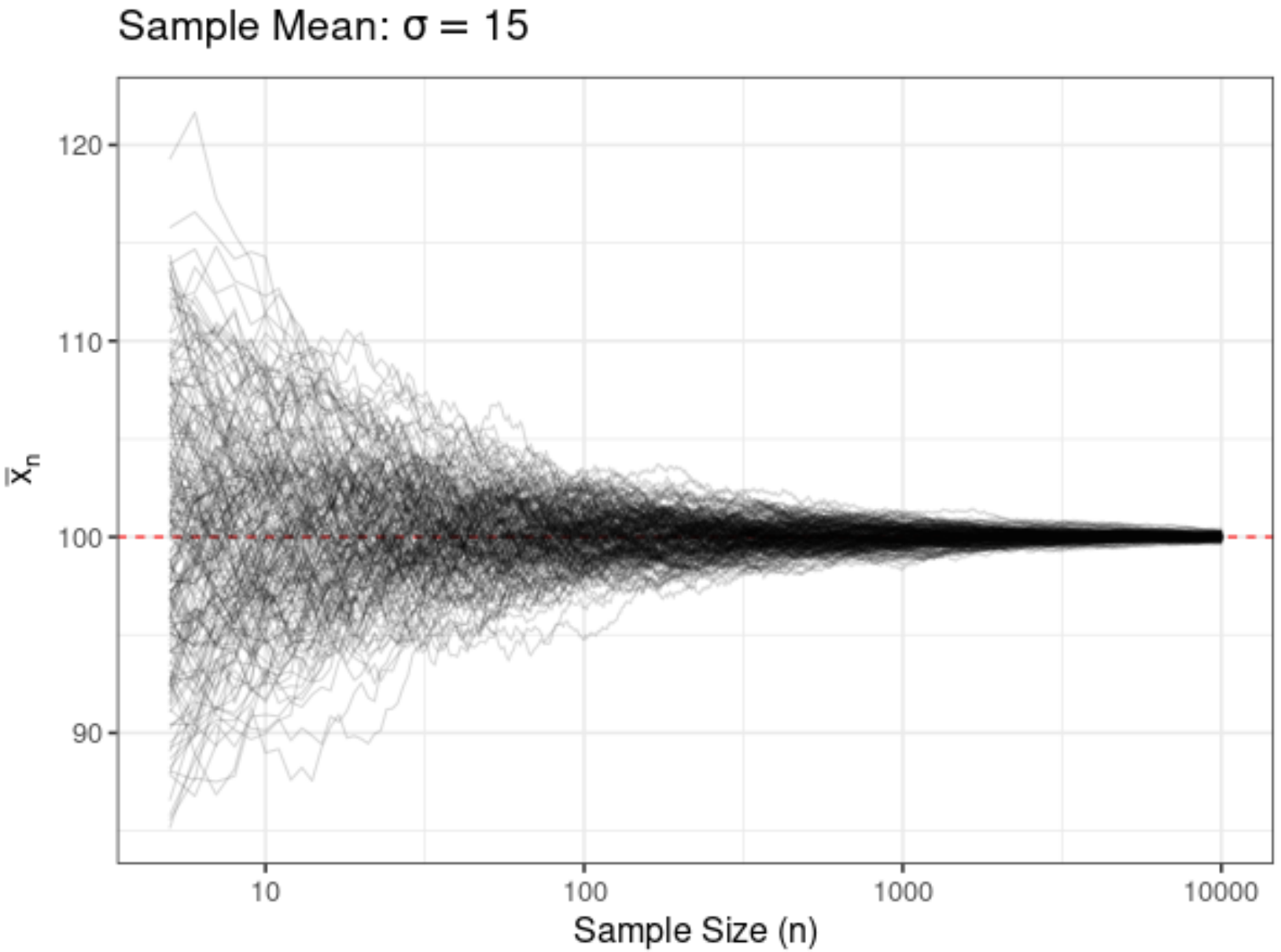
$$t \sim t(n-1)$$

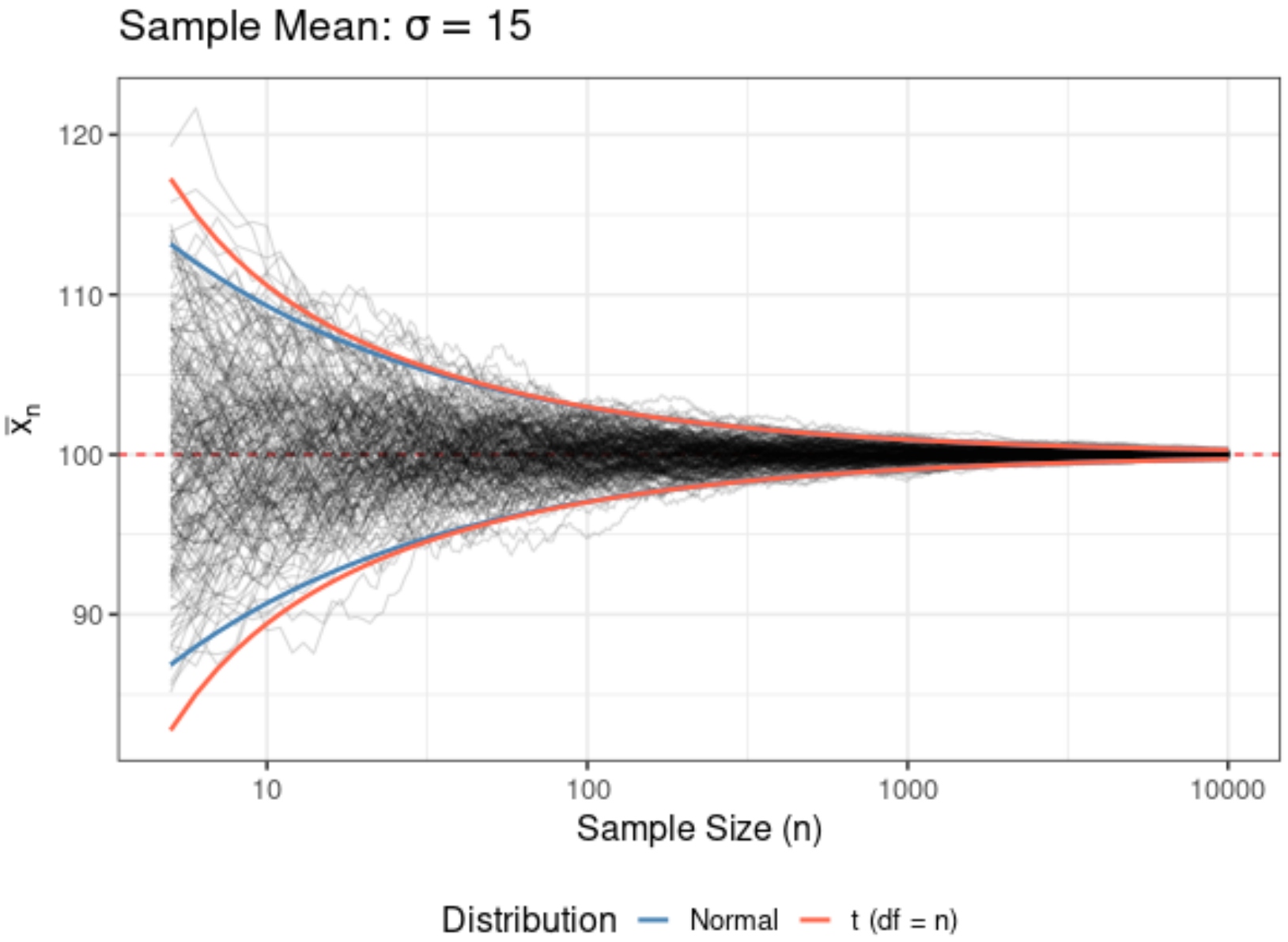
1. The t distribution is centered around 0
2. The t distribution has only one *distributional parameter* called the *degrees of freedom*, equal to $n-1$. This controls the variability
3. The t distribution has "fatter tails" than the normal distribution, allowing for the possibility of larger values
4. The standard error of a t distribution is $\sqrt{\frac{n-1}{n-3}}$ which gets closer to 1 as n increases
5. 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)

```
> 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?

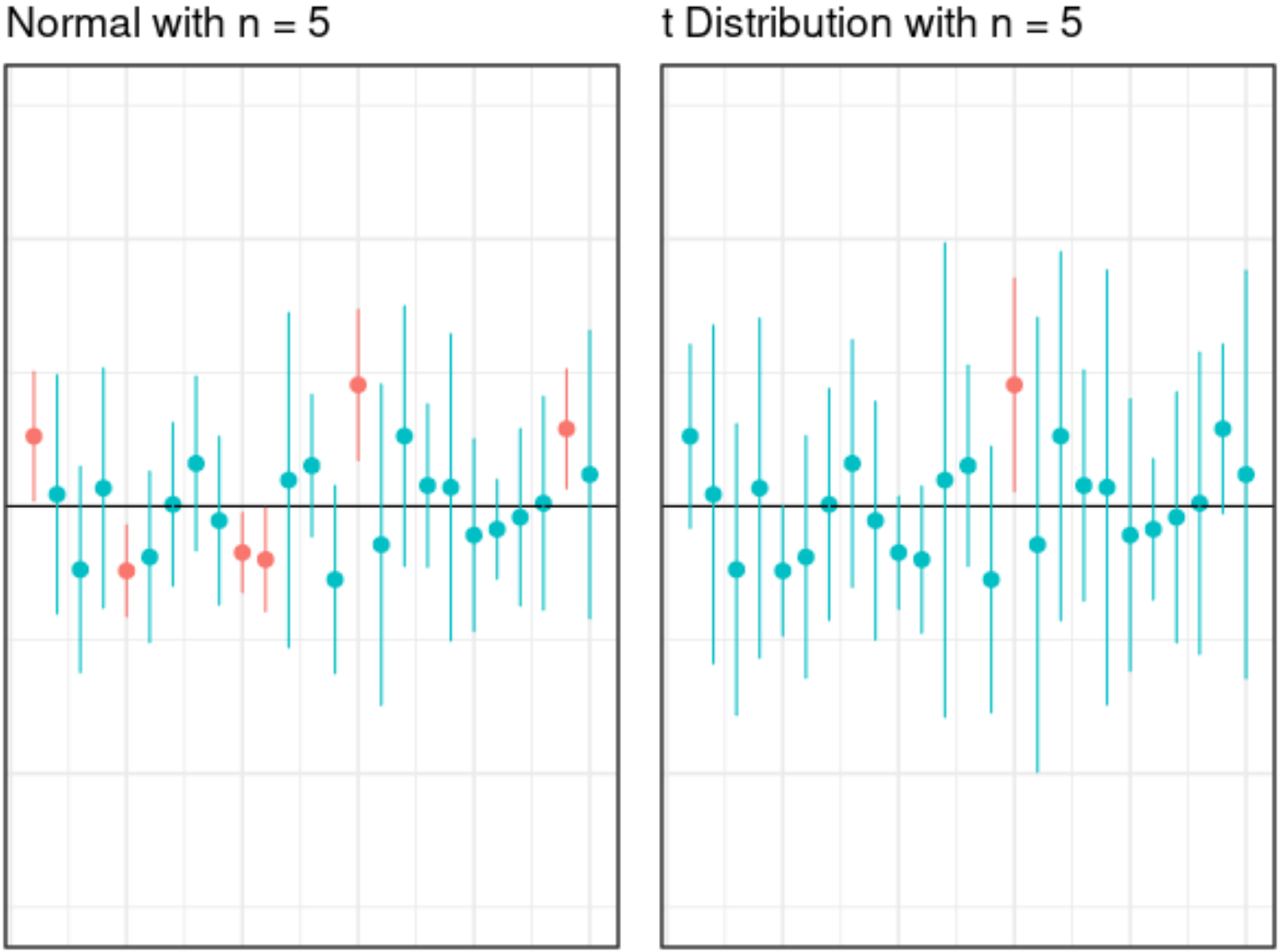
```
> quants <- c(0.025, 0.975)
> qt(quants, df = 5)
[1] -2.5706  2.5706

> 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

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

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

$$\bar{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}} = 199.94 \pm 2.0345 \times (5.9766/\sqrt{34})$$
$$= (197.82, 202.03)$$

Compare this with our estimate using the normal distribution:

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

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