Standardized Distribution

Grinnell College

October 13, 2025

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,

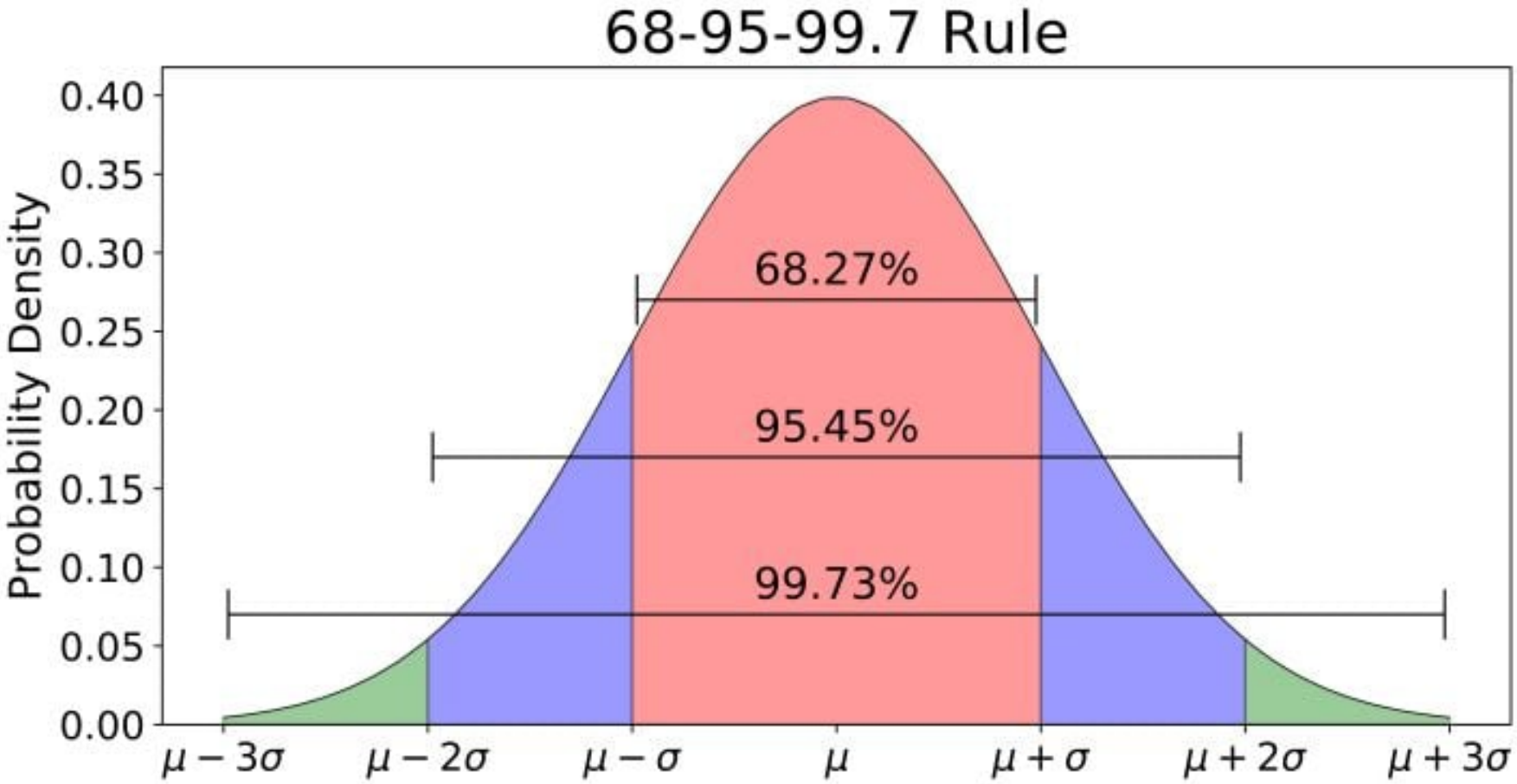
$$\overline{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}} = \left(\overline{x} - 1.96 \times \frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$$

what quantiles of the distribution would each endpoint represent?

Review

Key Items Last Week:

- Vocabulary
- Expression for confidence interval
- Which terms impact location, size, and confidence


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

Consider an expression written:

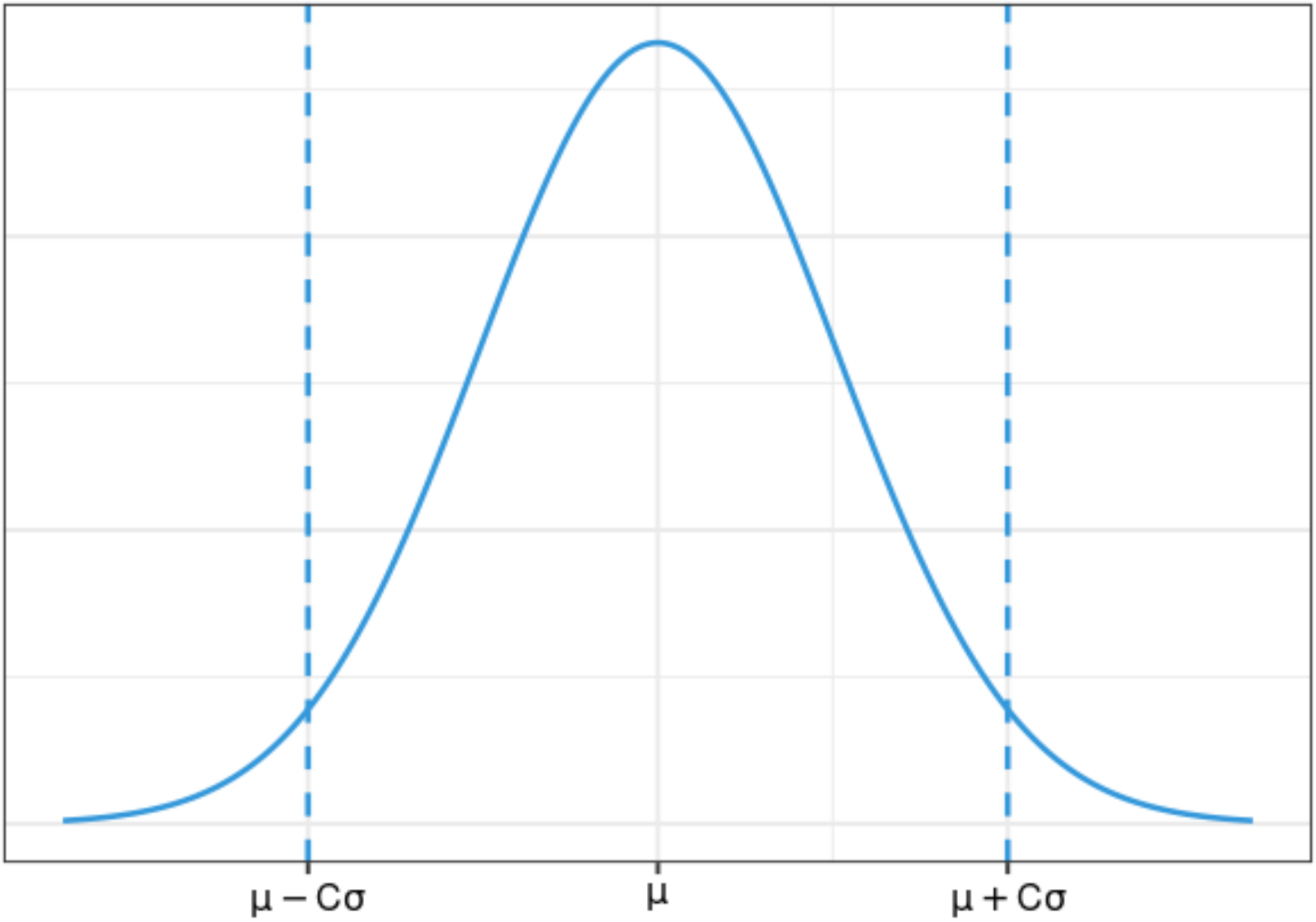
$$\overline{x} \pm C \times \frac{\hat{\sigma}}{\sqrt{n}}$$

Recall from last week that C, the **critical value**, is the value the mediates the relationship between my **confidence** and my standard error

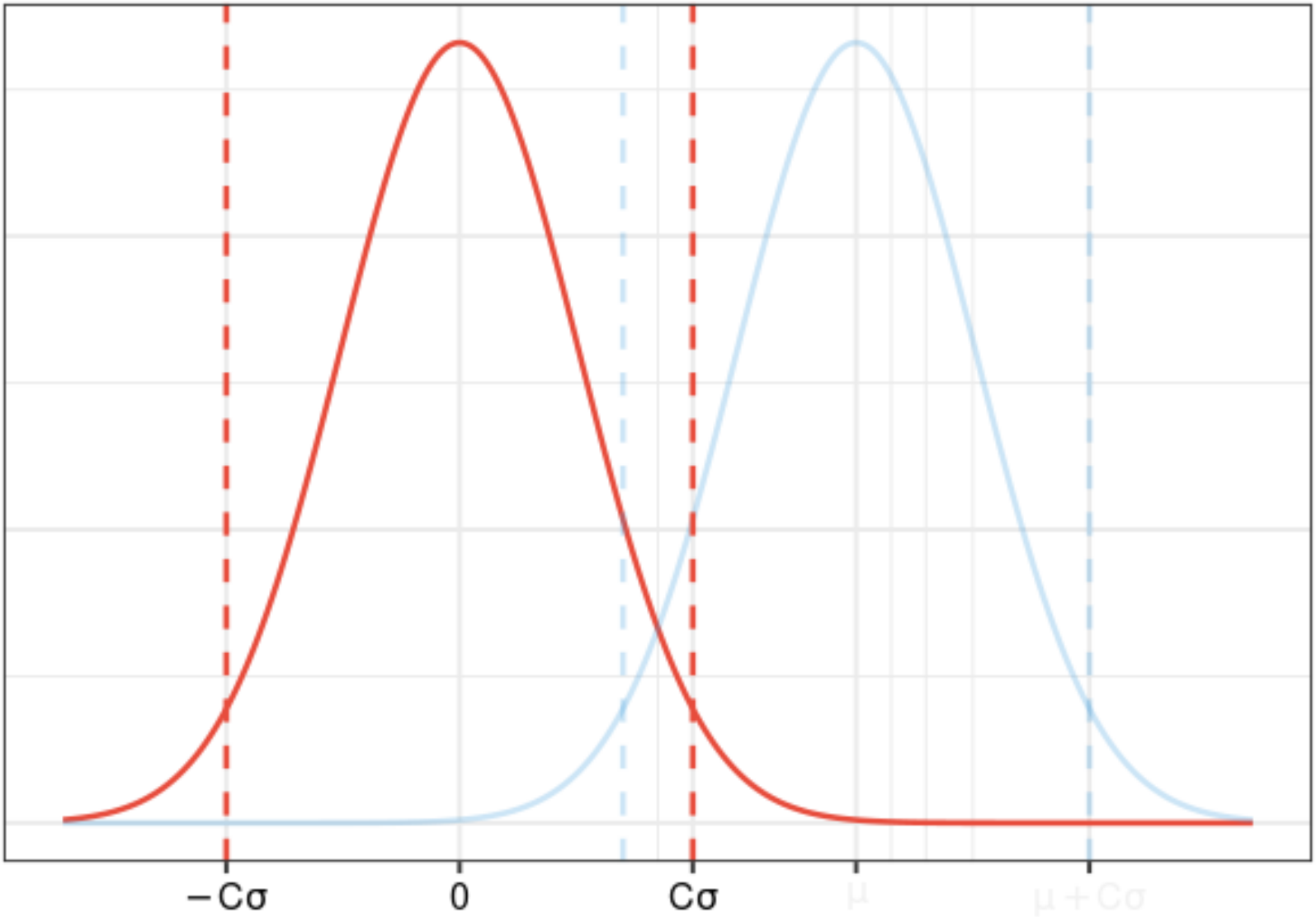
For example, based on the empirical rule on the previous slide, we know that setting $C = 2$ will cover 95.45% of our distribution

What would be of interest is a way to find values of C for any level of confidence

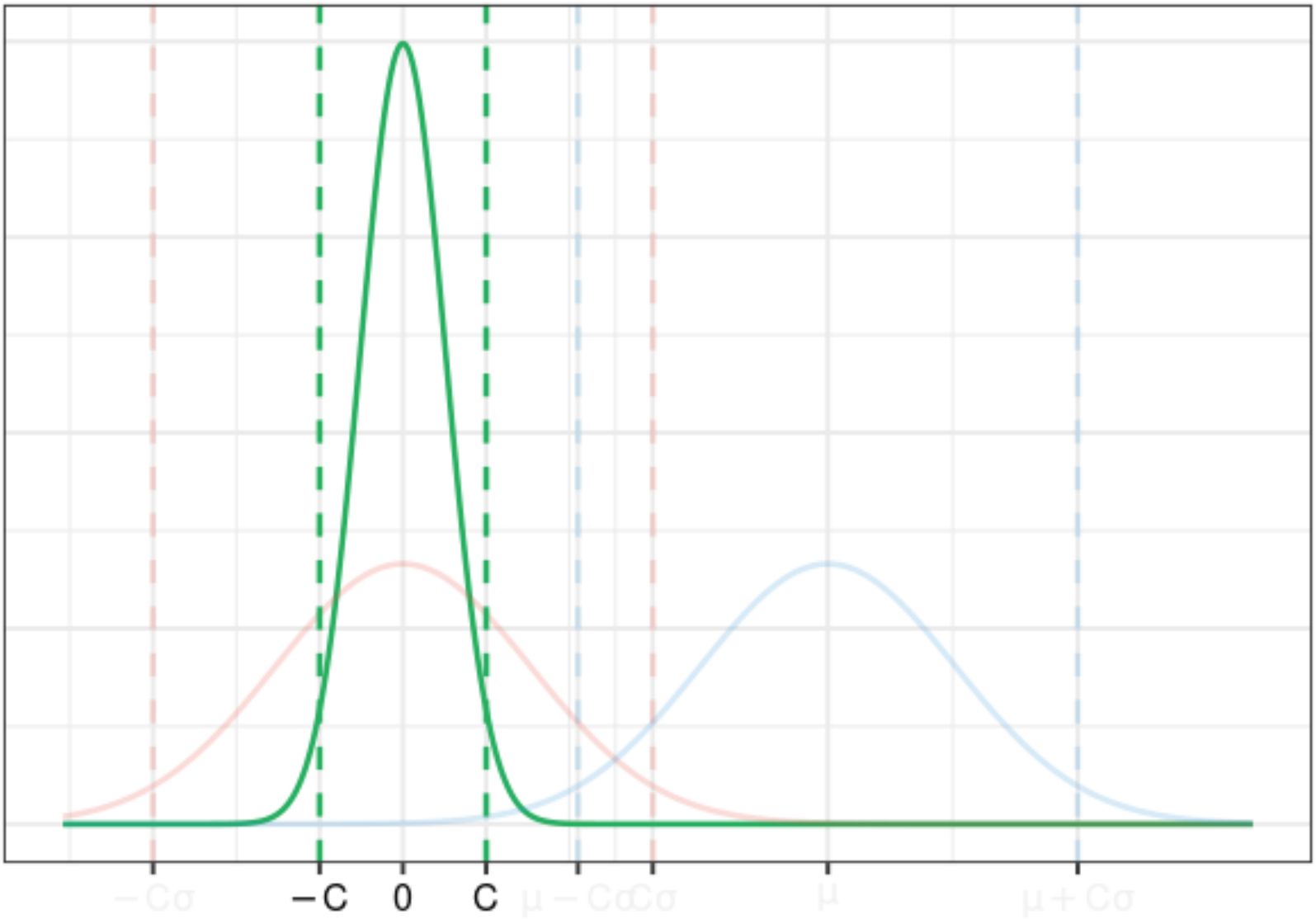
$$X \sim N(\mu, \sigma)$$



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



$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$



Standardization

Previously we saw that

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

would create a standardized variable with mean value of 0 and a standard deviation of 1. For a normally distributed variable like $\overline{X}$, this results in the following:

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

We call this a **z statistic** which follows a **standard normal** 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 of the standard normal distribution (this is the same 2.5th and 97.5th percentiles)

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:

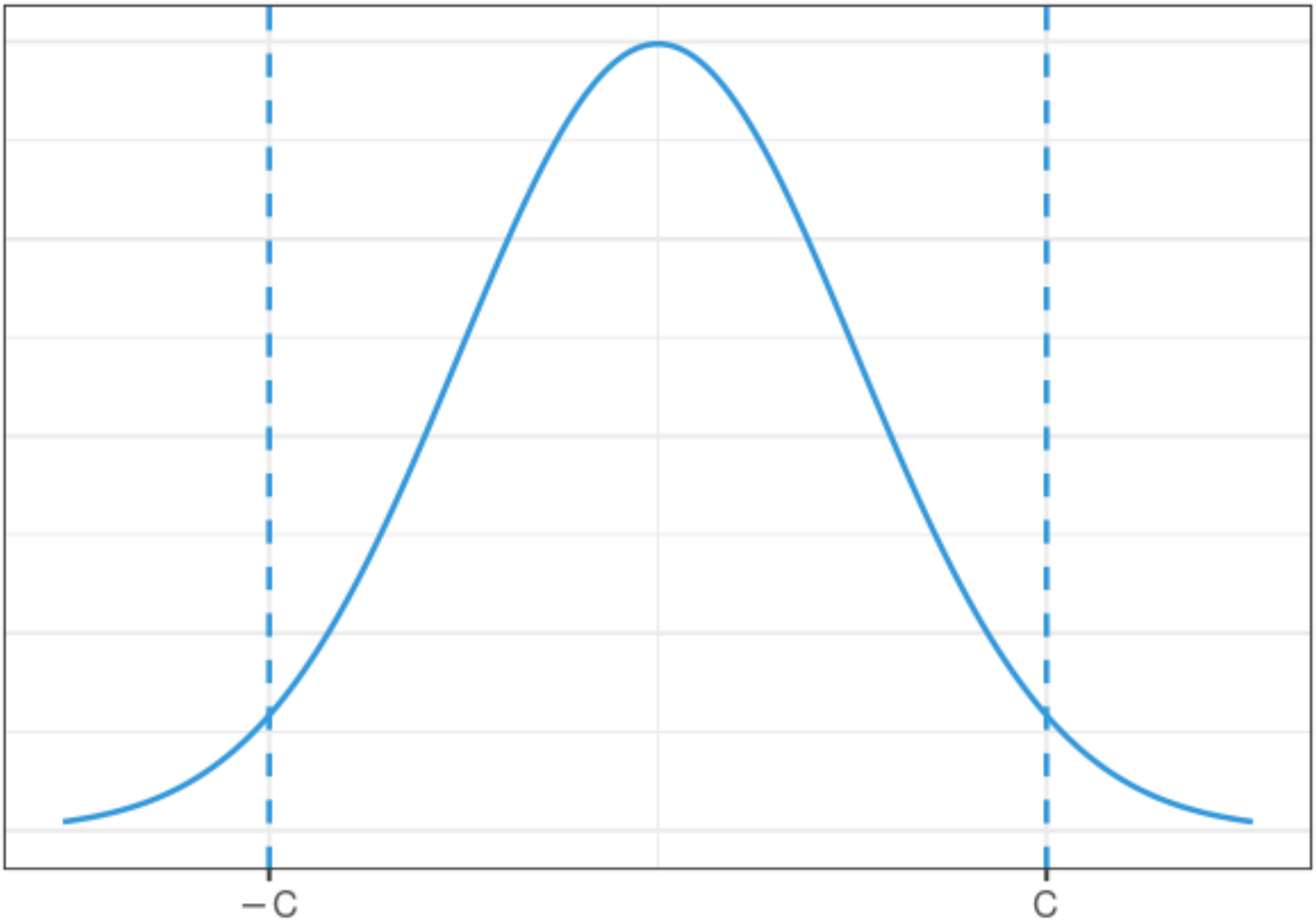
```
> quants <- c(0.025, 0.975)

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

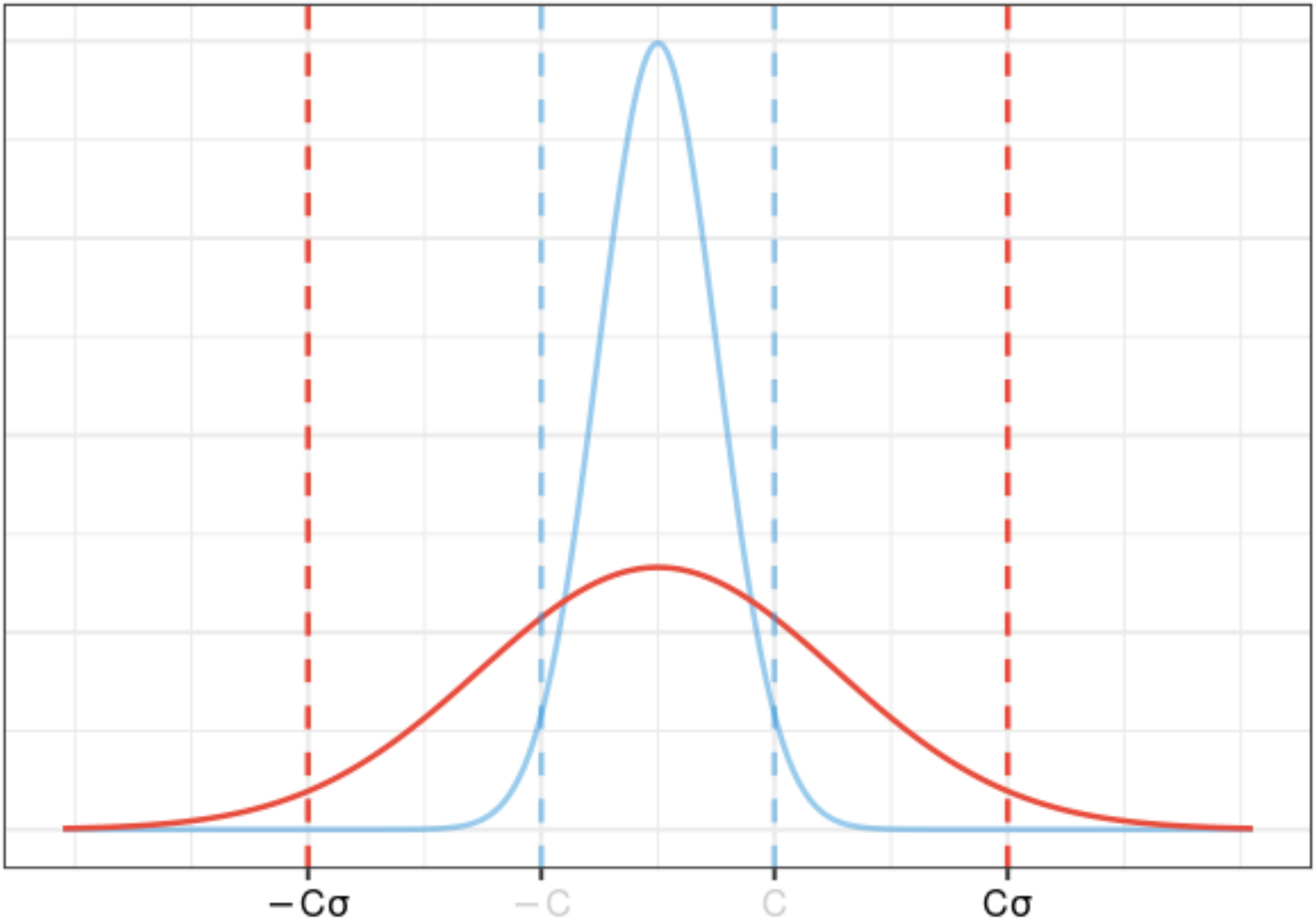
This means that for a *true* 95% confidence interval, we should be using

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

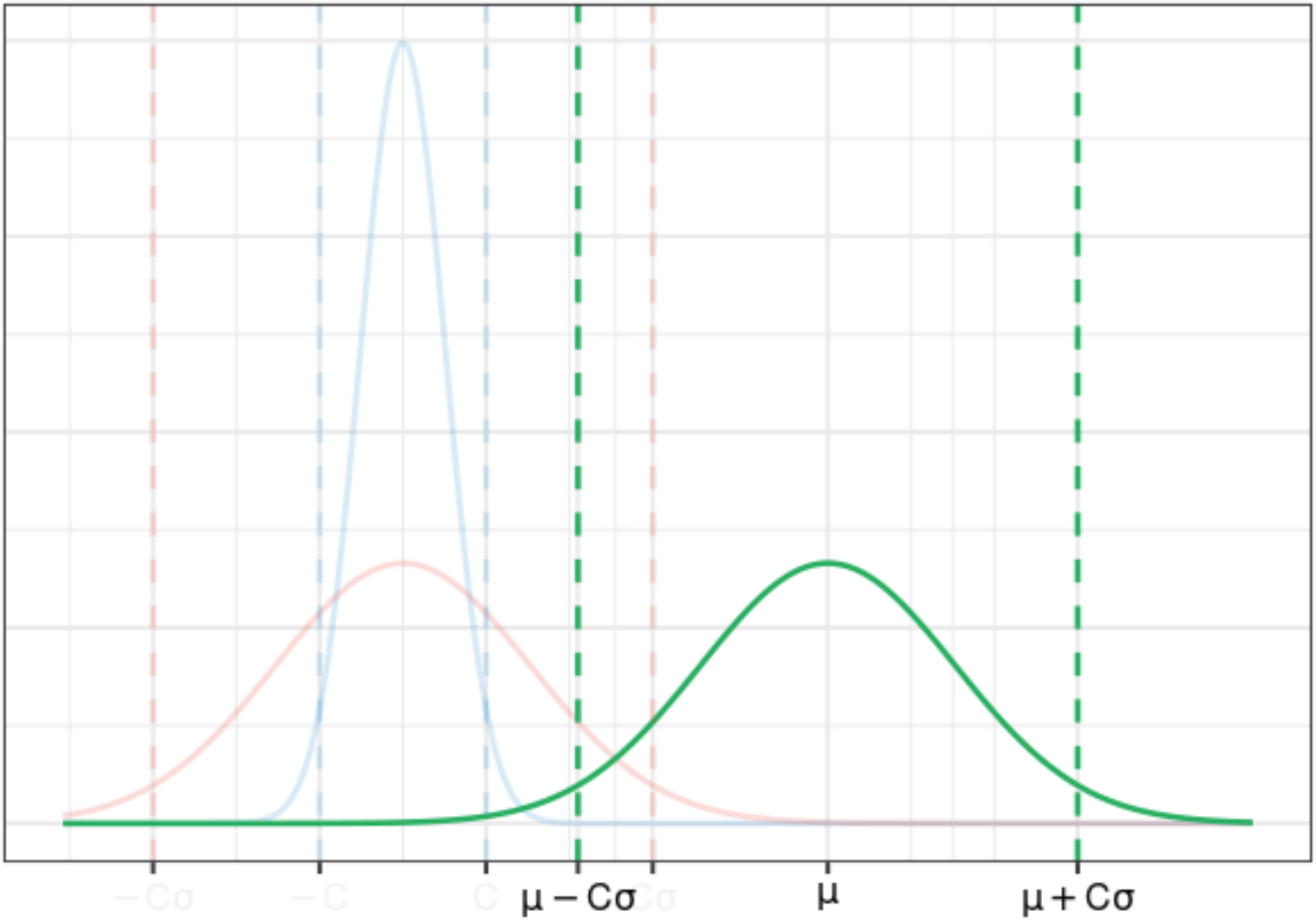
$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$



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



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



Z-Scores

In addition to providing us with critical values, standardizing values of $\overline{X}$ allow us to determine their proximity to μ in terms of standard deviations

Consider measurements of bill length (mm) on the male Adelie penguin, where we know our true population population parameters to be

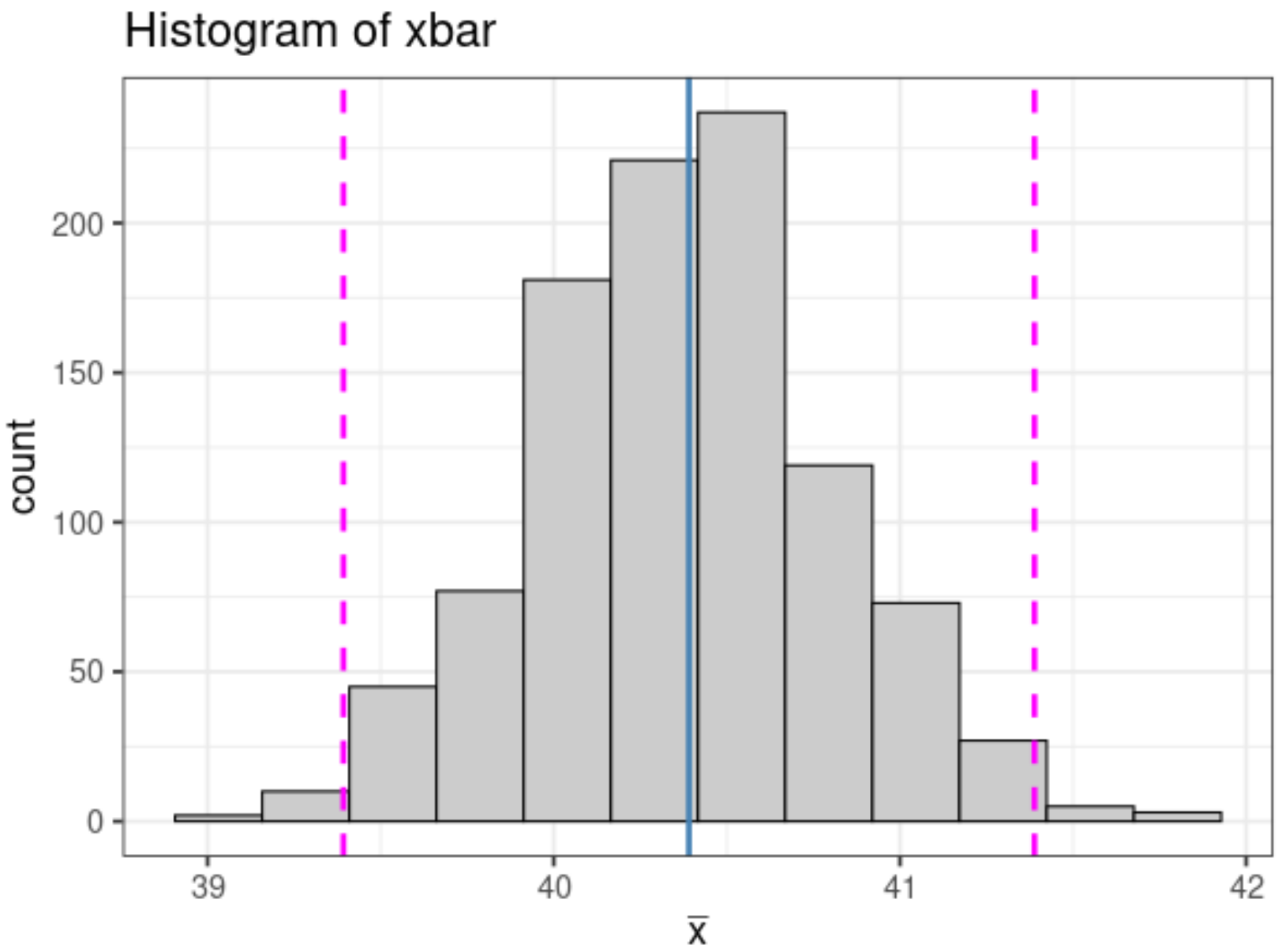
$$\mu = 40.39, \quad \sigma = 2.28$$

Suppose I collect 10 different samples of size $n = 20$. What kind of behavior can I expect from my collection of $\overline{x}$?

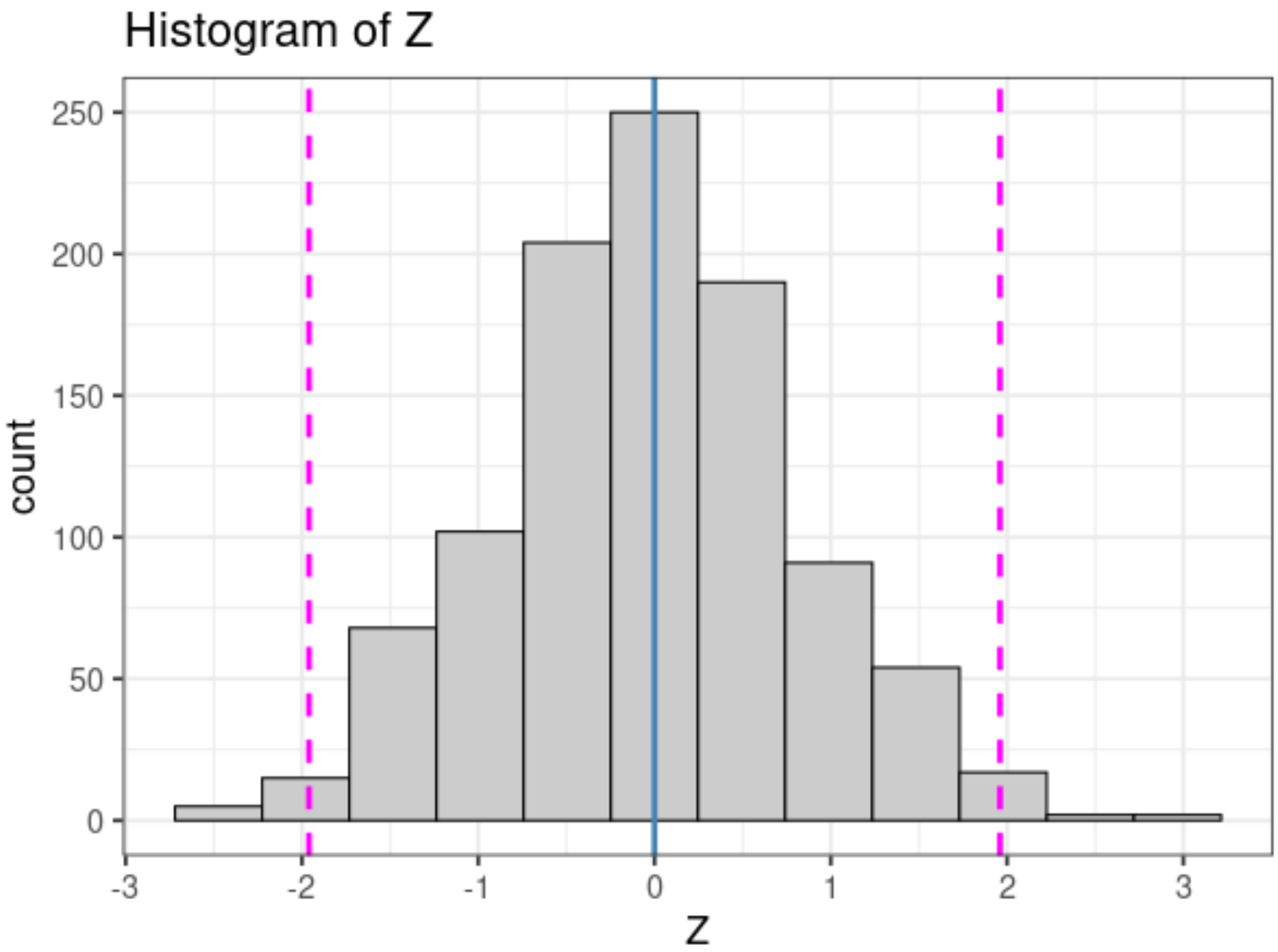
10 samples means from a population with $\mu = 40.39$ and $\sigma = 2.28$

Sample	$\overline{X}$	z-score
Sample 1	40.02	-0.737
Sample 2	40.5	0.215
Sample 3	40.76	0.736
Sample 4	40.48	0.186
Sample 5	40.58	0.382
Sample 6	40.52	0.255
Sample 7	40.49	0.196
Sample 8	40.42	0.058
Sample 9	39.17	-2.387
Sample 10	40.07	-0.629

1000 samples means from a population with $\mu = 40.39$ and $\sigma = 2.28$



1000 samples means from a population with $\mu = 40.39$ and $\sigma = 2.28$



Recap

Key Takeaways:

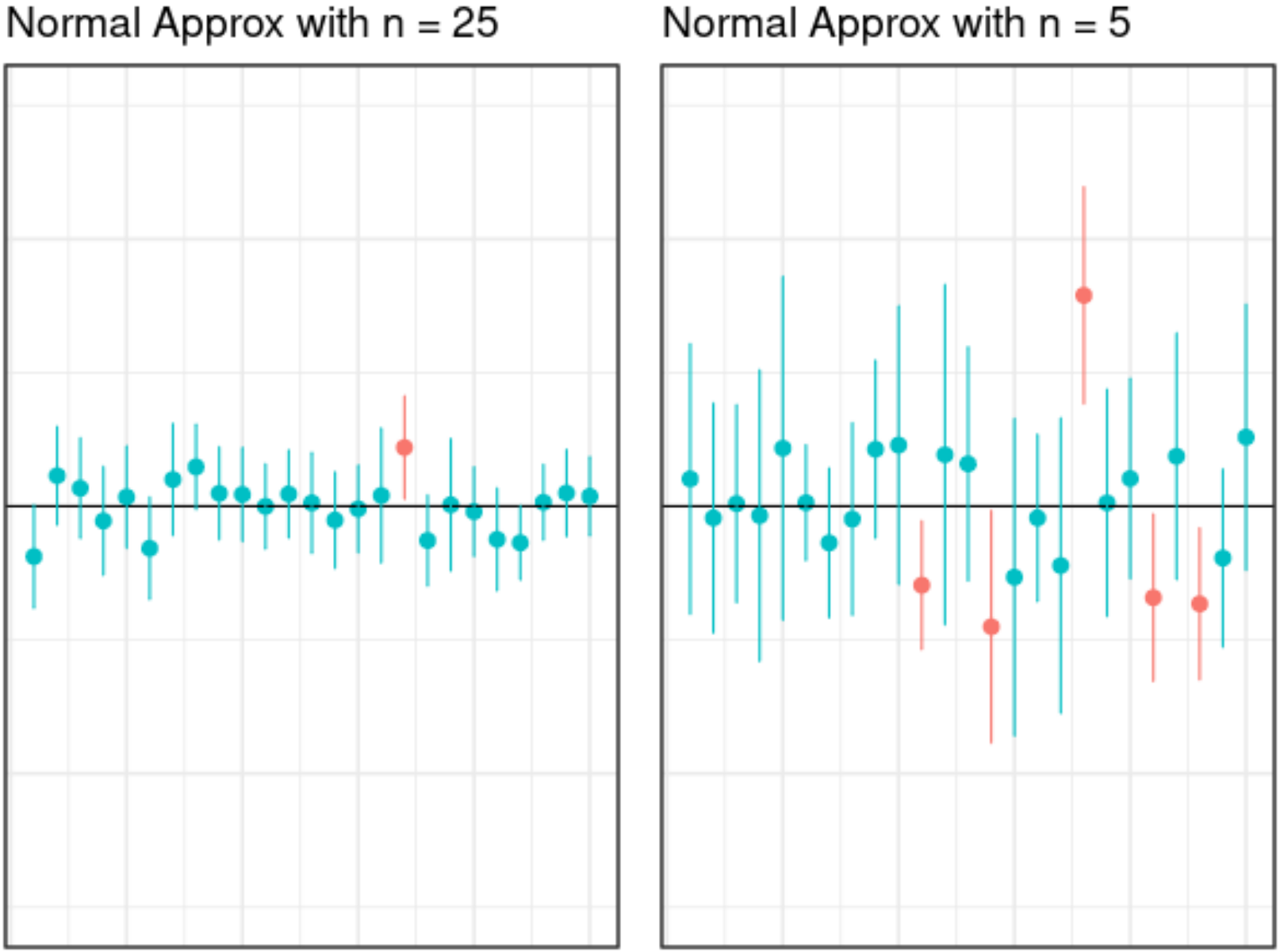
- We can standardize a normal distribution to create a **standard normal distribution**
- Critical values, C, are based on the quantiles of the standard normal
- If we knew μ and σ, we could standardize values of $\overline{X}$ to measure their relation to μ

Issues with Approximation

It's important to understand that 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

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 *over-estimate* 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
- When our sample size is smaller, we *overestimate* how certain we are about our estimation of σ

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

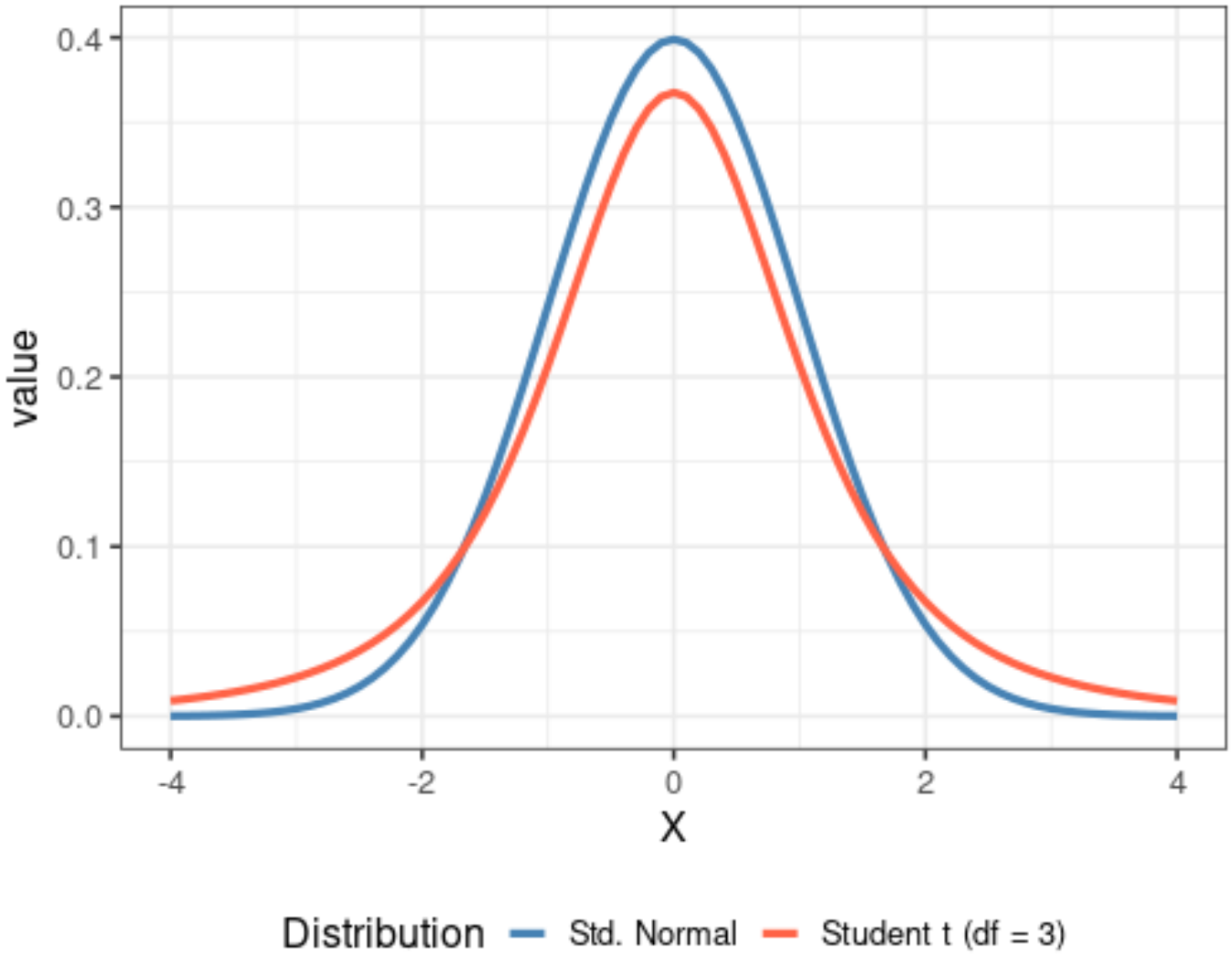
The sampling distribution of the t-statistic is known as the t-**distribution**

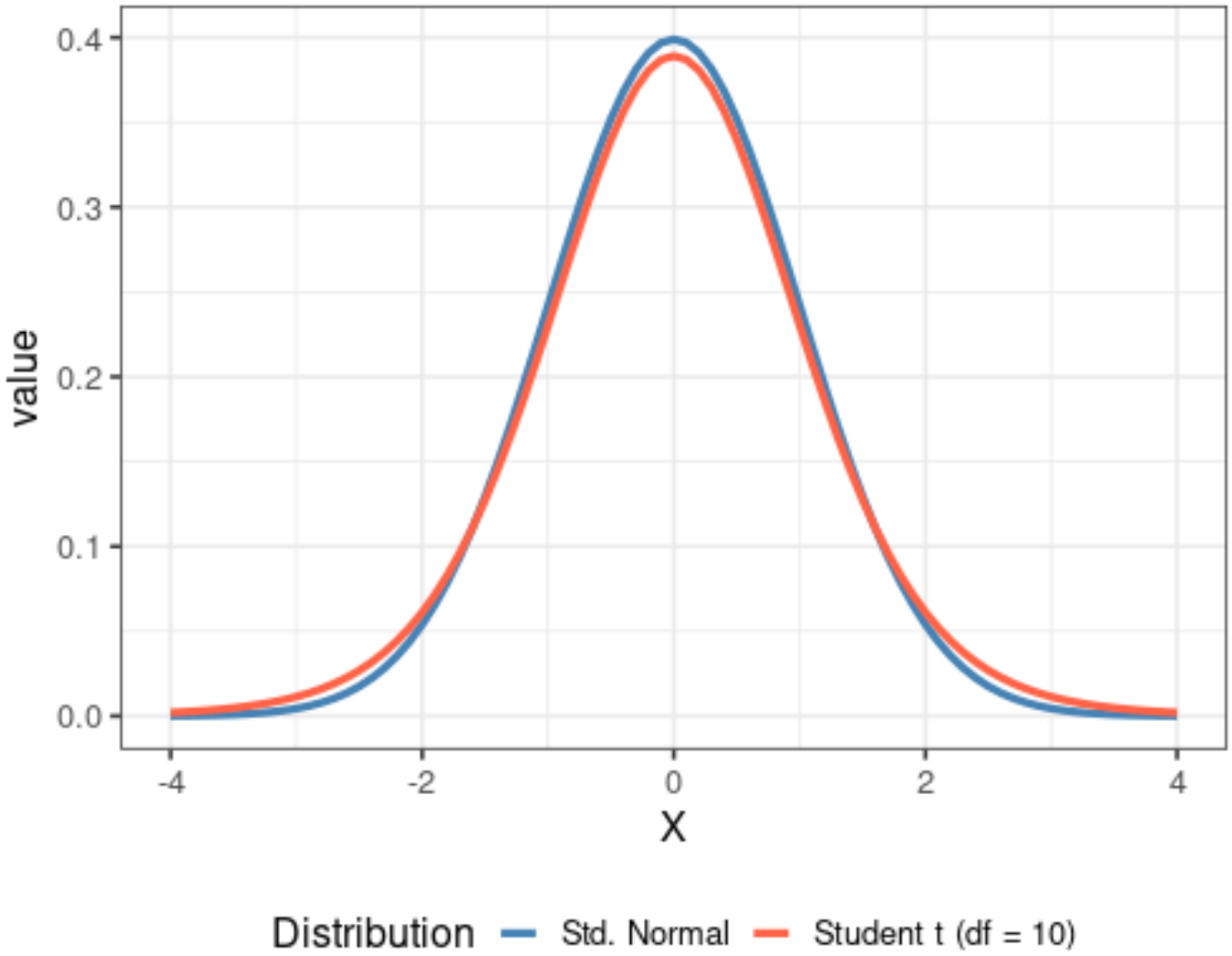
Student's t-distribution

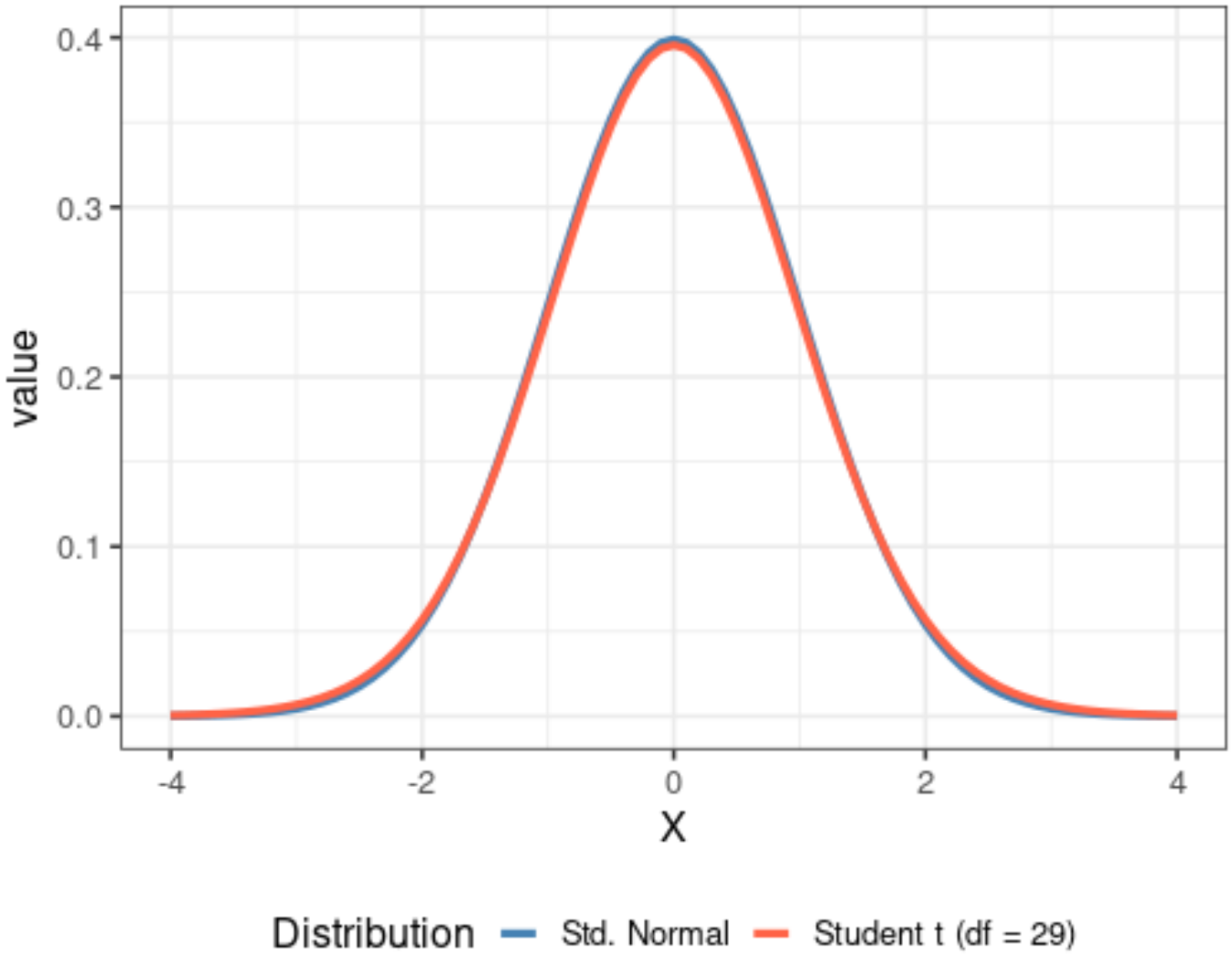
The t statistic and the t-distribution are very similar to a z statistic and a standard normal distribution:

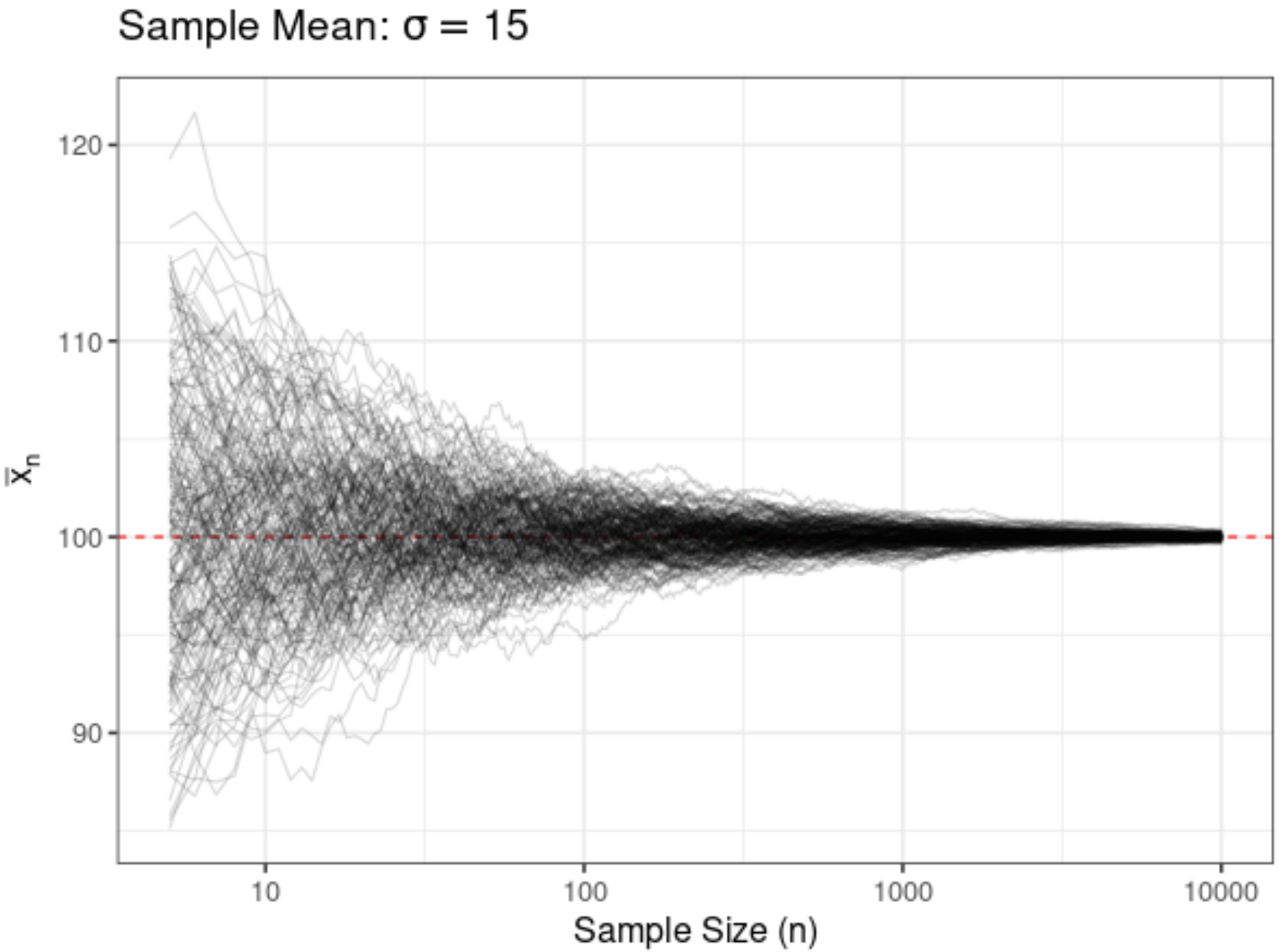
$$t = \frac{\overline{x} - \mu}{\hat{\sigma}/\sqrt{n}}, \qquad t \sim t(n-1)$$

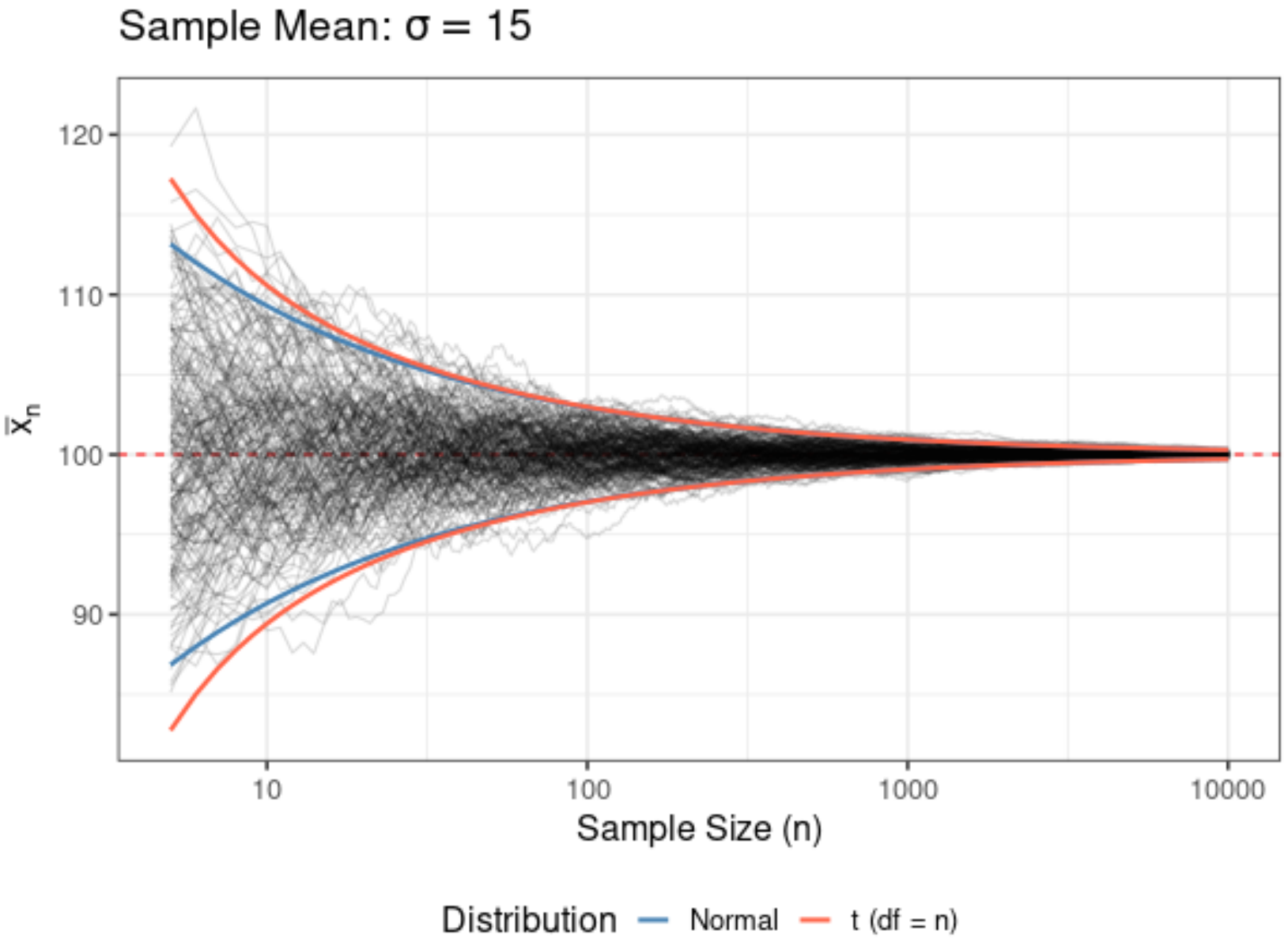
1. The t distribution is symmetric around 0
2. The t distribution has only one *distributional parameter* called the *degrees of freedom*, equal to $n-1$. This controls the variability
3. The t distribution has "fatter tails" than the normal distribution, allowing for the possibility of larger values
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)

```
> 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 our estimate with a standard normal,

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

Review

Big day today:

- By assuming a distribution, we can use quantiles to determine our **critical values** for constructing confidence intervals
- We can standardize sampling distribution to derive the **standard normal distribution**
- Standardized values of $\overline{X}$ give us a sense of how close our sample mean is to the true parameter
- Our estimation of $\hat{\sigma}$ necessitates accommodating 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