Proportions and Normality

Grinnell College

March 6, 2024

Review

Key takeaways from Central Limit Theorem:

1. For any variable X with mean μ and population standard deviation σ, the sample mean will have a sampling distribution of

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

2. CLT justifies use of 68-95-99 rule
3. This holds even if our population is not normally distributed
4. This *does not hold* for other statistics (i.e., median)
5. If population non-normal, may need more samples for better approximation

Central Limit Theorem

In particular, we have that as the size of our sample increases, the sampling distribution for $\overline{x}$ approaches

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

If we were to *standardize* our variable

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

so that the mean value of the sampling distribution is 0 with standard deviation 1, we end up with a standard normal:

$$\overline{Z} \sim N(0, 1)$$

Sample Means and Proportions

There is an interesting relationship between means and proportions

For example, consider taking a fair coin and flipping it 10 times. How many heads would you expect to see?

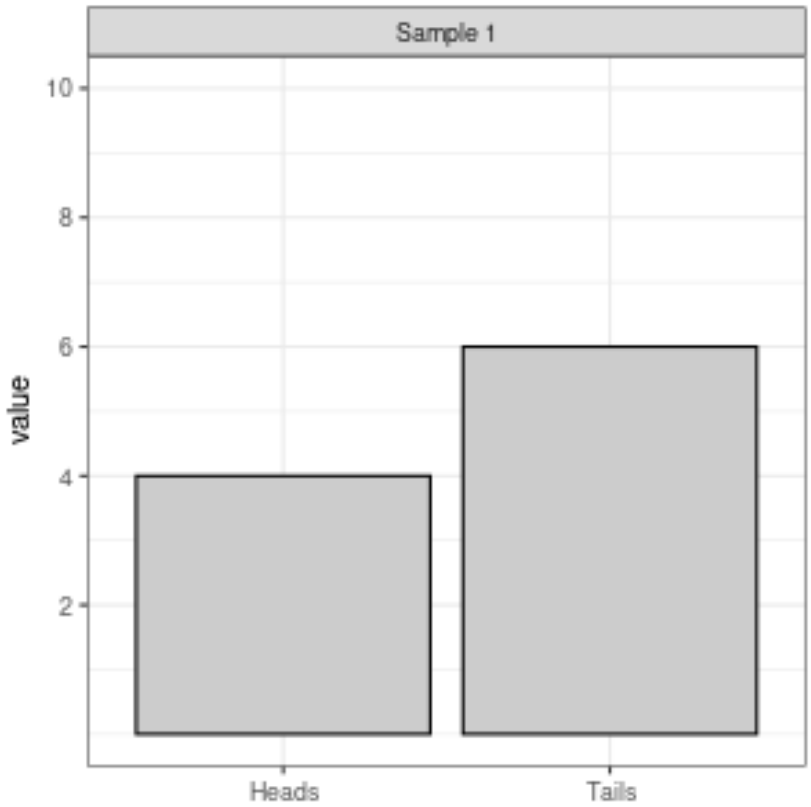
$$\mathcal{S} = \{H, H, T, T, H, T, H, T, T, T\}$$
$$X = \{1, 1, 0, 0, 1, 0, 1, 0, 0, 0\}$$

We can find the *proportion* of heads from our sample $\mathcal{S}$ by simply taking the total number of heads and dividing by the total number of flips, giving

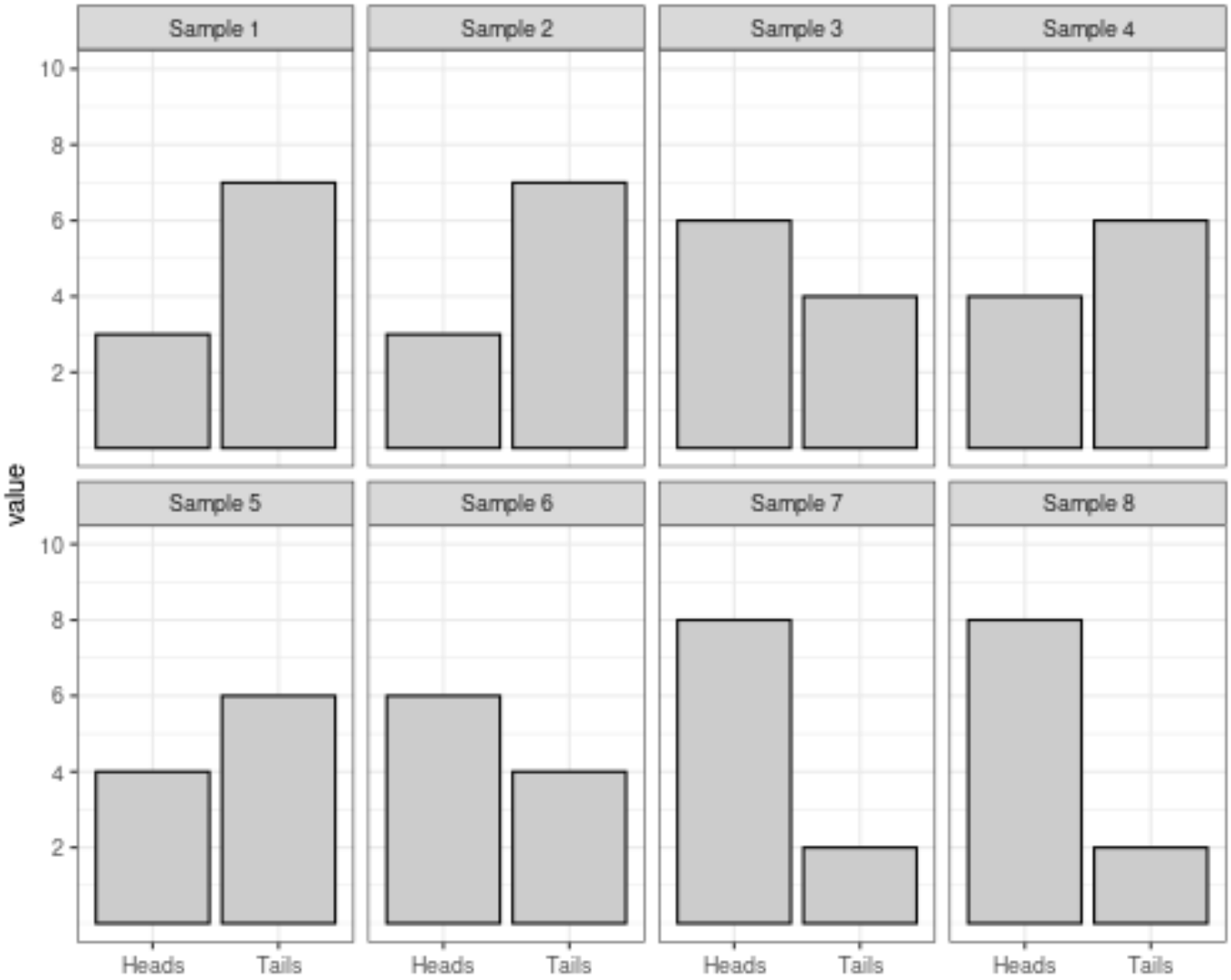
$$\hat{p} = \frac{4}{10}$$

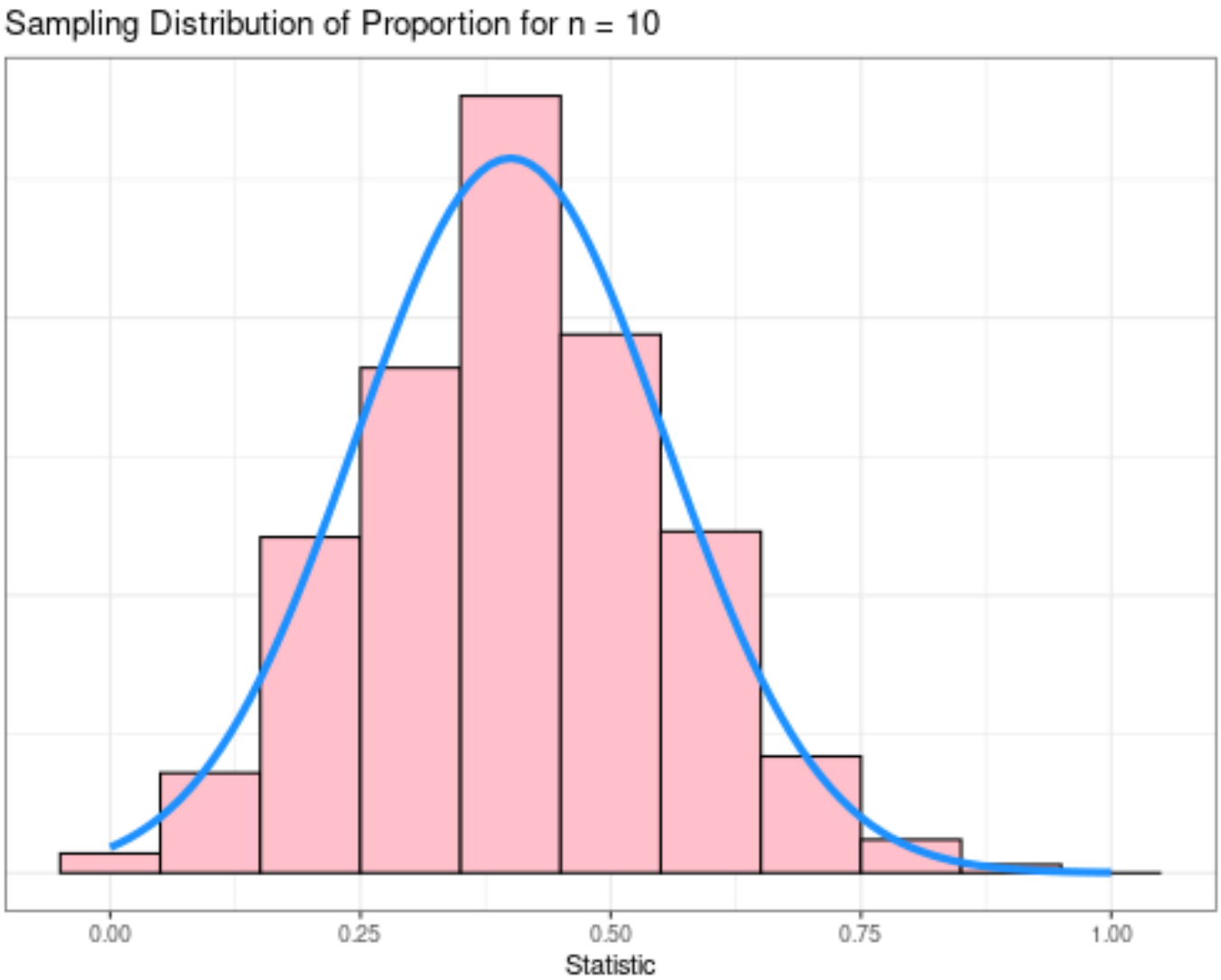
However, if we consider X, which defines H as 1 and T as 0, we can also find the sample mean:

$$\begin{aligned} \overline{x} &= \frac{1}{10} \sum_{i=1}^{n} x_i \\ &= 0.4 \end{aligned}$$



Repeated Samples




Sampling Distribution of Proportion for n = 10
0.00
0.25
0.50
0.75
1.00
Statistic

Central Limit Theorem

For a sample with one proportion, the sampling distribution of our proportion statistic, $\hat{p}$ is approximately

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

There a few rules of thumb relating to the size and the proportion:

1. $n \times p \geq 10$
2. $n \times (1-p) \geq 10$

In particular, it is often difficult to estimate proportions precisely that are near the boundaries (0 and 1)

Example

In a study conducted by Johns Hopkins University researchers investigated the survival of babies born prematurely. They searched their hospital's medical records and found 39 babies born at 25 weeks gestation (15 weeks early), 31 of these babies went on to survive at least 6 months. With your group:

1. Use a normal approximation to construct a 95% confidence interval estimate for the true proportions of babies born at 25 weeks gestation that are expected to survive
2. An article on Wikipedia suggetss that 70% of babies born at a gestation period of 25 weeks survive. Is this claim consistent with the Johns Hopkins study?

Example

1. We find that

$$\hat{p} = \frac{31}{39} = 0.795$$

$$SE = \sqrt{\frac{0.795(1 - 0.795)}{39}} = 0.065$$

   From here, we found our 95% CI:

$$0.795 \pm 2 \times 0.065 = (0.668, 0.922)$$

2. As 0.7 is contained within our constructed 95% CI, it is consistent with the results of the study by Johns Hopkins

Example cont.

We have seen several questions now in the labs and homework in which we ask, "Based on the data we have seen, is such and such a reasonable value?"

One method which we could employ is to construct a confidence interval and see if the value in question is contained within, i.e., "is 0.7 contained within the interval (0.668, 0.922)?"

Alternatively, we could investigate z-scores

z-scores

We found in the Johns Hopkins example that $\hat{p} = 0.795$ and $SE = 0.065$. If we were to construct a z-score of the form

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

We would find a z-score for the estimate of 70% to be

$$Z = \frac{0.7 - 0.795}{0.065}$$

$$= -1.4615$$

indicating that an estimate of 70% is less than one and a half standard deviations away from the mean

Sampling Distributions with Normal Approximation

Statistic	Std. Error	Conditions
$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$np \geq 10$ and $n(1-p) \geq 10$
$\overline{x}$	$\frac{\sigma}{\sqrt{n}}$	Normal population $n \geq 30$
$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	Same
$\overline{x}_1 - \overline{x}_2$	$\frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}$	Same

"Approximations"

Until now, we have been dealing with approximations

In particular, we have primarily dealt with cases in which our sample sizes were "large enough"

But what happens when they're not?

CLT for Sample Mean

We saw from the Central Limit Theorem that our sample mean approaches the distribution

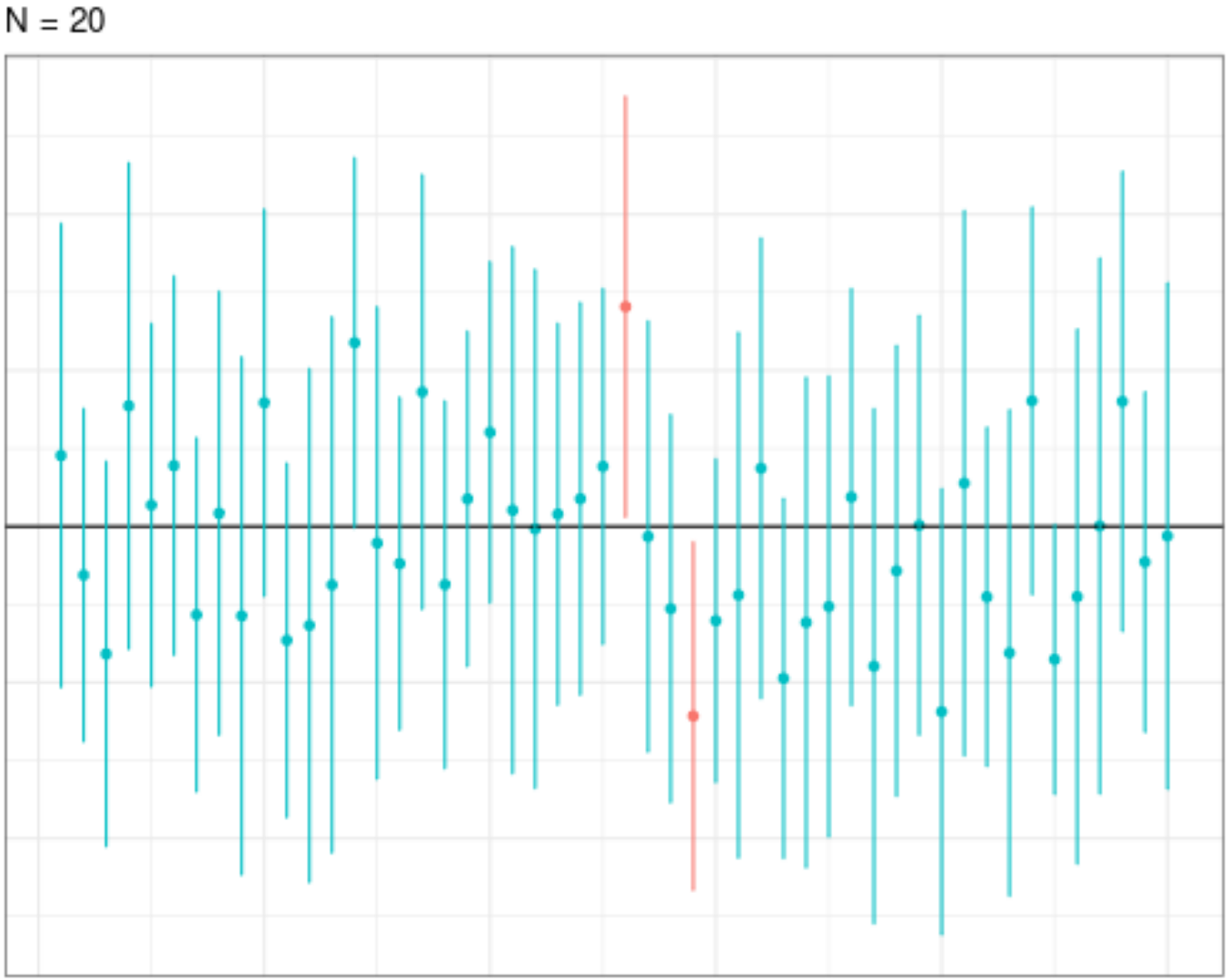
$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

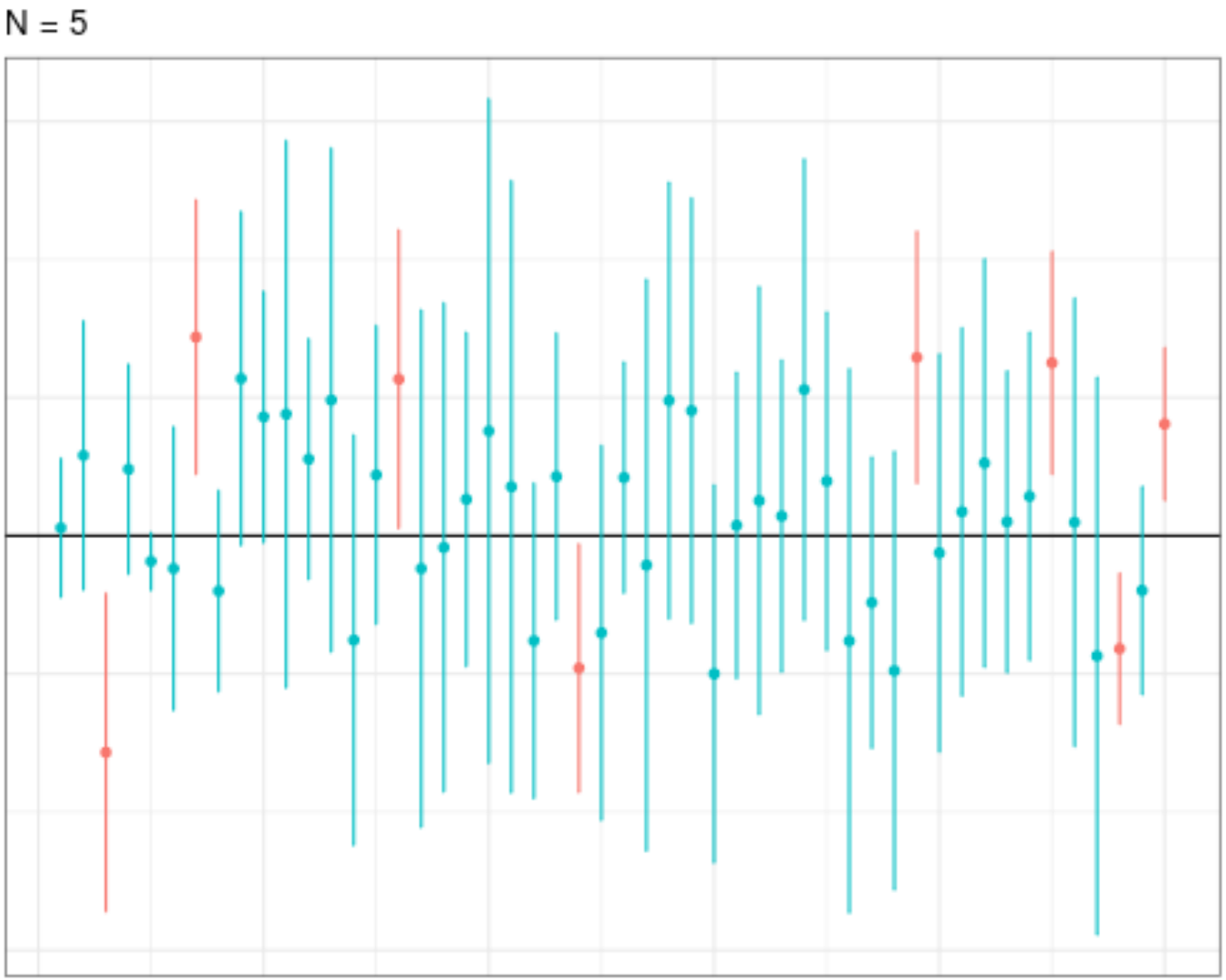
and, from this information, we have concluded that

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

would provided the intended coverage for an appropriate selection of C.

Indeed, we showed this empirically with the coverage of confidence intervals from a sample of size $n = 20$ from a normal distribution $N(50, \sigma = 3.8)$


N = 20


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

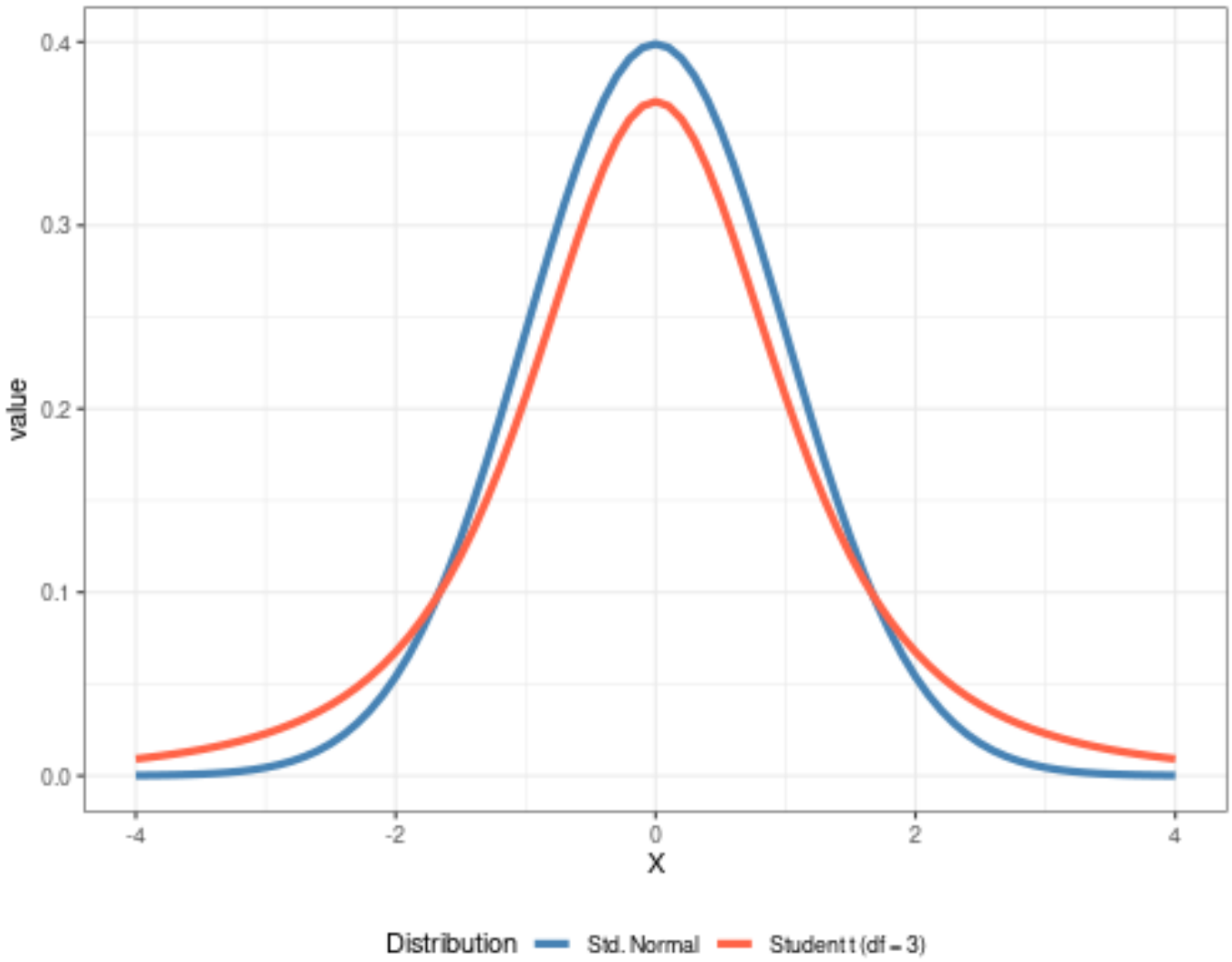
However, because Guinness was not keen on its competitors finding out that it was hiring statisticians, he was forced to publish his new distribution under the pseudonym "student", hence "Student's t-distribution"

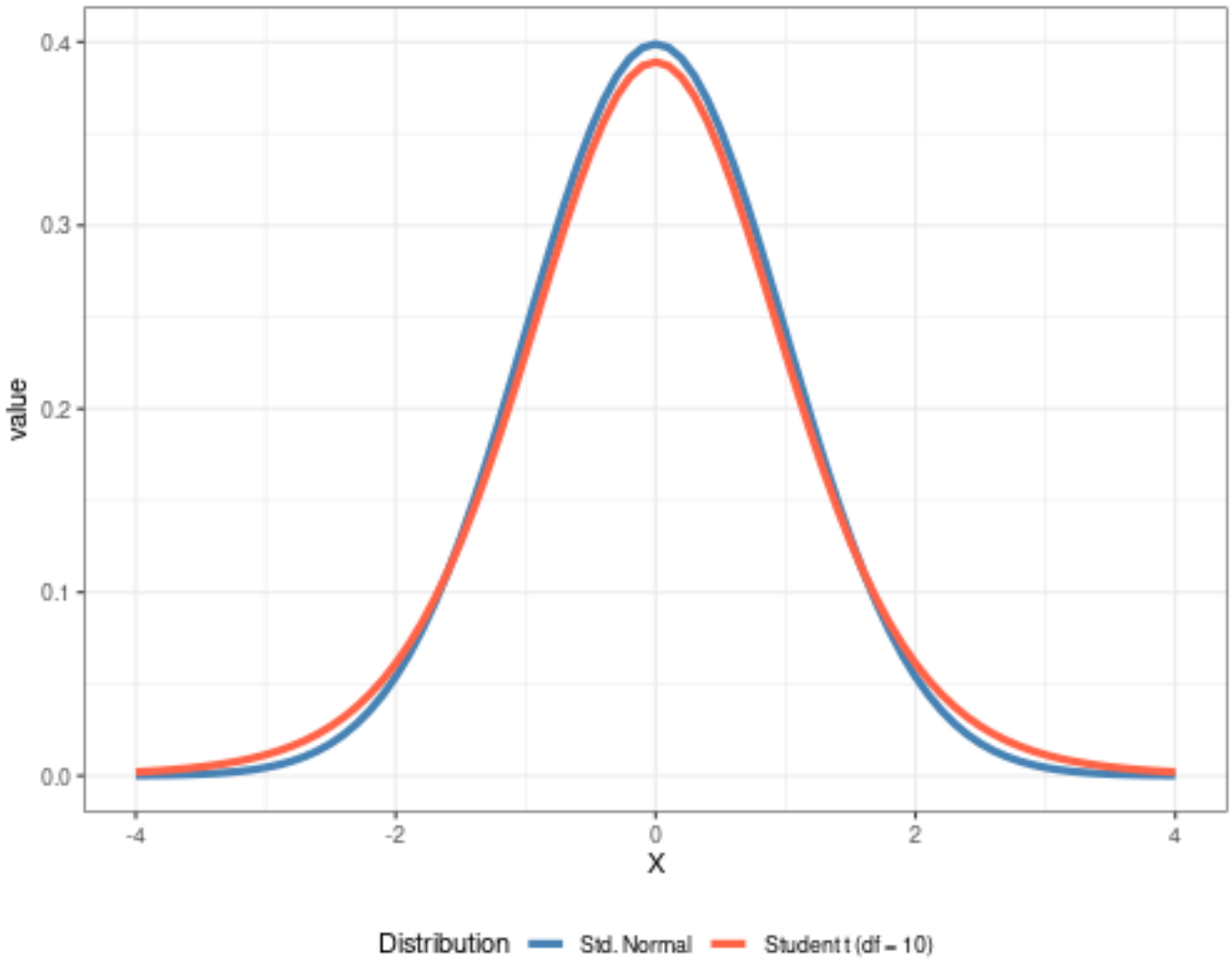
Student's t-distribution

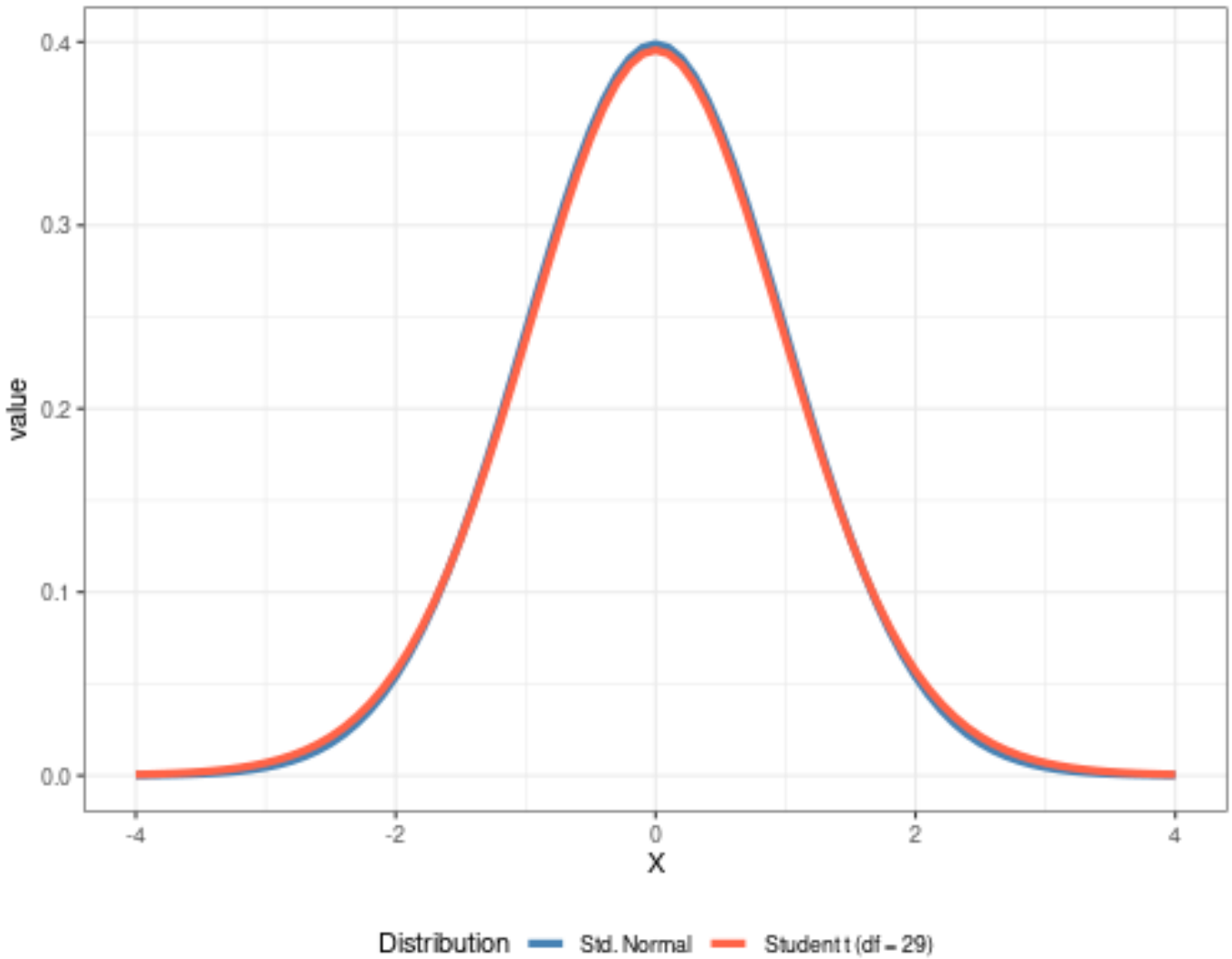
Student's t Distribution:

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

1. The t distribution has only one 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 normal as $n \to \infty$

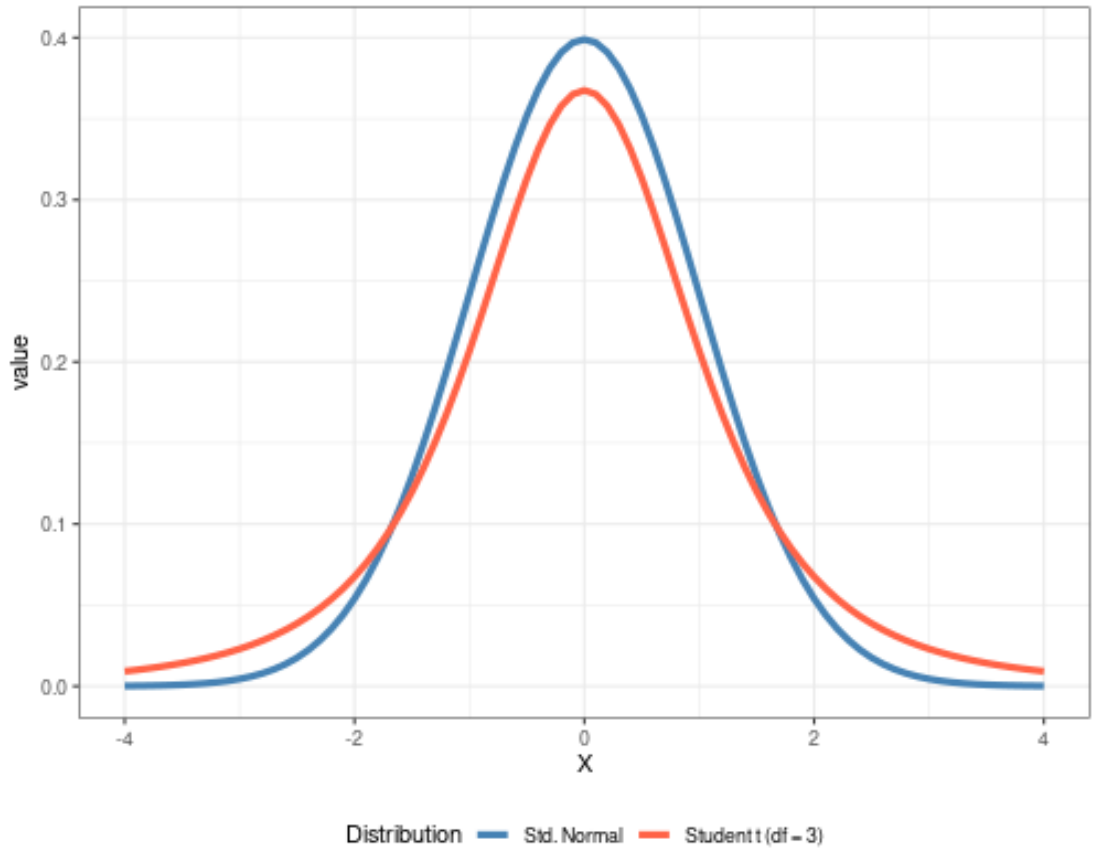

0.4
0.3
0.2
0.1
0.0
value
-4
-2
0
2
4
X
Distribution
Std. Normal
Student t (df = 3)

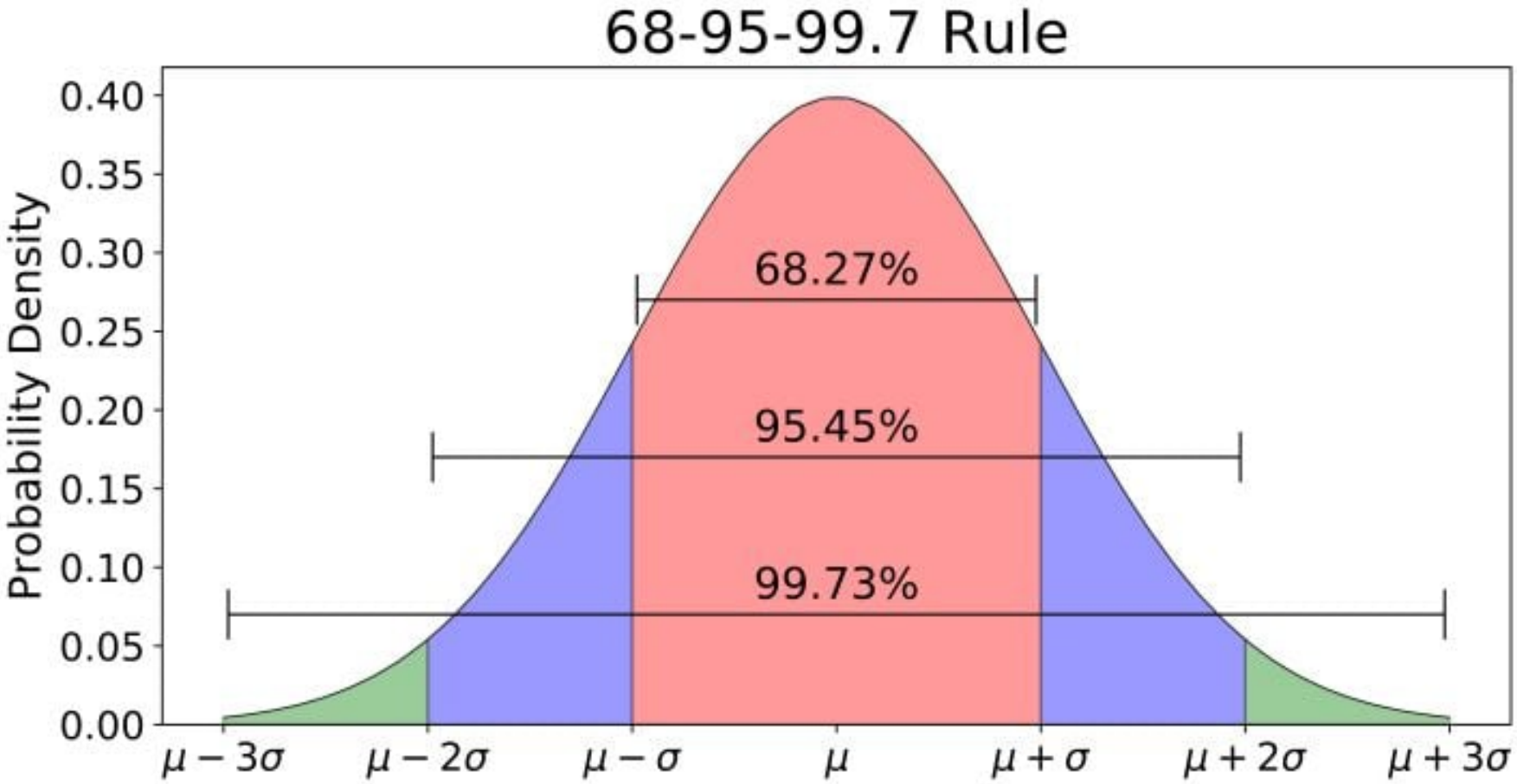

0.4
0.3
0.2
0.1
0.0
value
-4
-2
0
2
4
X
Distribution
Std. Normal
Student t (df = 10)


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

Implications?

What are the implications of this for our confidence intervals?




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
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

stuff

There are special quantile functions in R associated with distributions. In the same way `rnorm()` was used to generated RandomNORMals, the function `qnorm()` will return the quantiles of a normal distribution

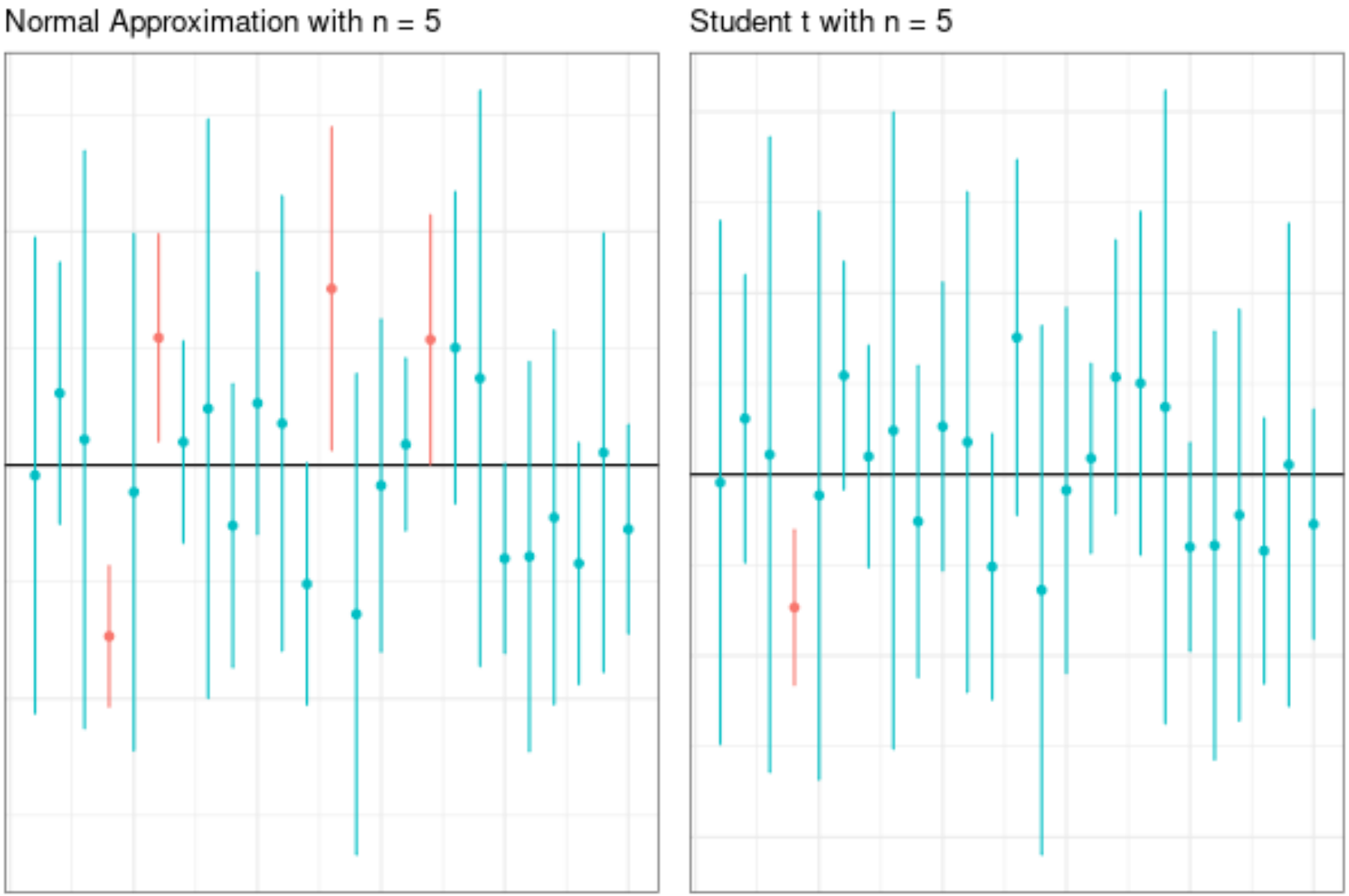
```
> qnorm(c(0.025, 0.975))
[1] -1.96  1.96
```

Here, we see the "2" that we have been using up to this point

Now consider the same quantile for the t distribution (using `qt()`) with 9 degrees of freedom ($n = 10$ observations):

```
> qt(c(0.025, 0.975), df = 9)
[1] -2.2622  2.2622
```

How has the uncertainty in estimating σ impacted our intervals?


Normal Approximation with n = 5
Student t with n = 5

t-Distribution, Key Takeaways

- Normal distribution is approximate, t distribution is exact
- This is true for *all* sample sizes
- Not considering $\hat{\sigma}$ causes us to *underestimate* variability and width of CI
- We can still use Point $\pm$ MOE to find CI
- Instead of $C = 2$, we need to use `C = qt(., df = n - 1)`
- If population is very skewed *and* sample size small, we may need other options