Confidence Intervals

Conditions, t-distribution, Different Confidence %'s

Grinnell College

October 18, 2024

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

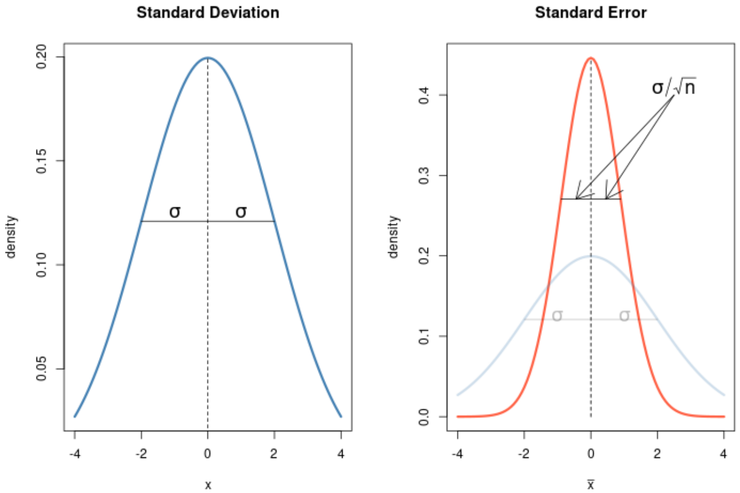
Review – Central Limit Theorem

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 $\sigma/\sqrt{n}$.

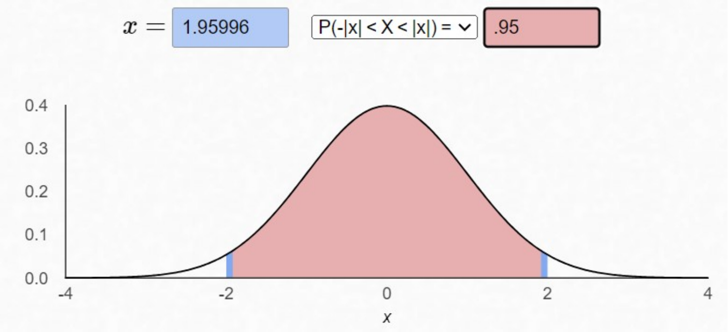
$$\overline{X} \sim \mathrm{N}(\mu, \sigma/\sqrt{n})$$

More on CLT


Standard Deviation
density
0.20
0.15
0.10
0.05
σ
σ
-4
-2
0
2
4
x
Standard Error
density
0.4
0.3
0.2
0.1
0.0
$\sigma/\sqrt{n}$
σ
σ
-4
-2
0
2
4
$\bar{x}$

CI's using Normal Distribution

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

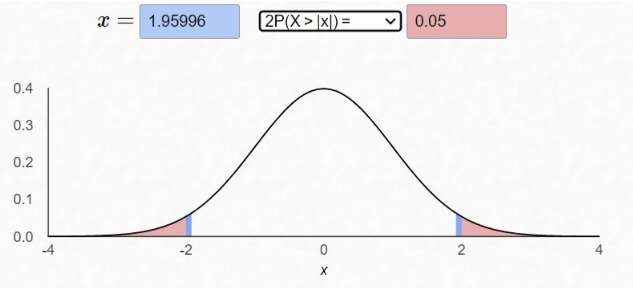


$x =$ 1.95996

P(-|x| < X < |x|) = ˅ .95

Note: I lied about using ±2×SE, we actually should be using ±1.96×SE

CI's using Normal Distribution



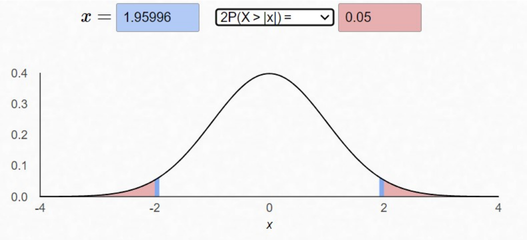
$x =$ 1.95996

2P(X > |x|) = 0.05

We could find the value we want to use for the ME $= \pm C \times \mathrm{SE}$ using R

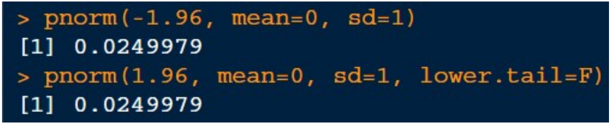
- We want 95% of cases within the values
- equivalent we want 5% outside the values
- cut that 5% in half (2.5%) to get how much probability we need on each side of the normal distribution

qnorm() in R



$x =$ 1.95996 2P(X > |x|) = 0.05

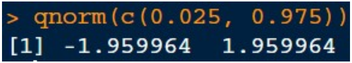
`pnorm()` let us find probabilities associated with specific values in R



```
> pnorm(-1.96, mean=0, sd=1)
[1] 0.0249979
> pnorm(1.96, mean=0, sd=1, lower.tail=F)
[1] 0.0249979
```

`qnorm()` is a function that does the opposite

- gives values that correspond to specific probabilities



```
> qnorm(c(0.025, 0.975))
[1] -1.959964  1.959964
```

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):

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

Conditions

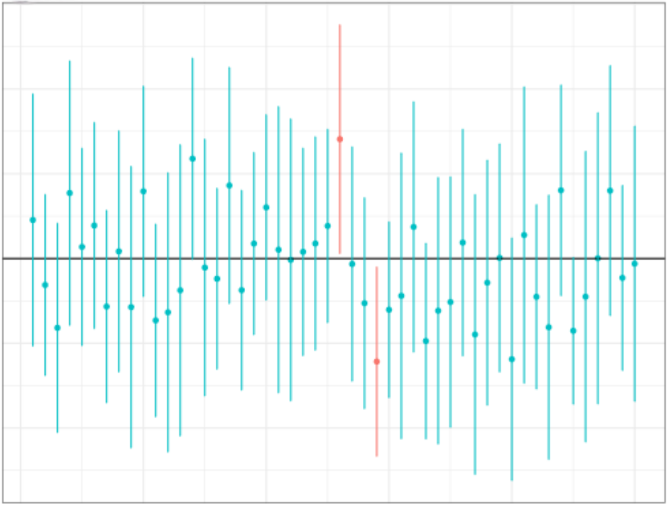
When we don't know the value of σ...

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

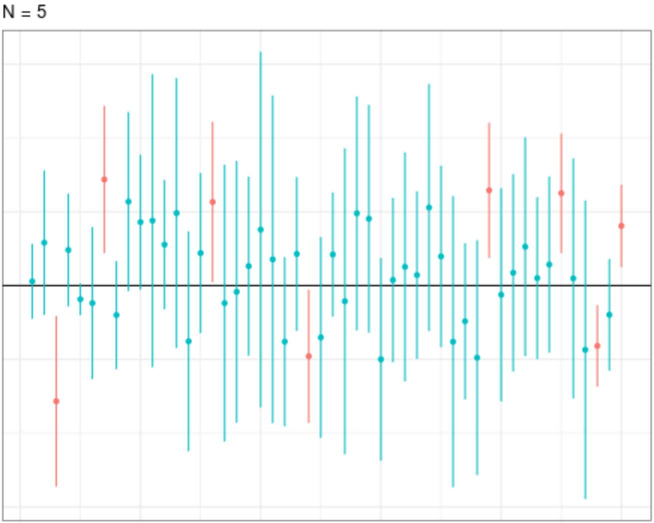
- The *population* must be Normal
  - we will almost always never know if this is true
- sample size n $\geq$ 30

Conditions

N = 50



Conditions



Conditions

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

- random sample
- sample size n $\geq$ 30

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

Estimating Variance

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

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"

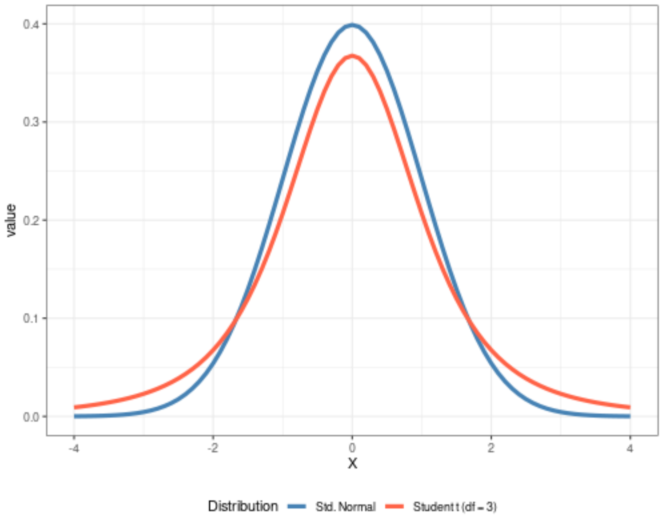
Student's *t*-distribution

Student's t Distribution:

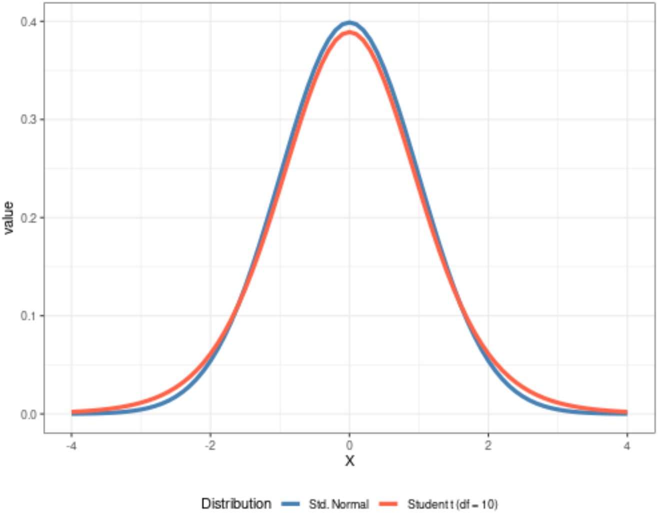
$$X \sim t(n-1)$$

1. symmetric around zero
2. bell-shaped
3. one parameter called the *degrees of freedom*, equal to $n-1$
4. has more probability in the tails than the normal distribution
5. The t distribution will become normal as $n \to \infty$

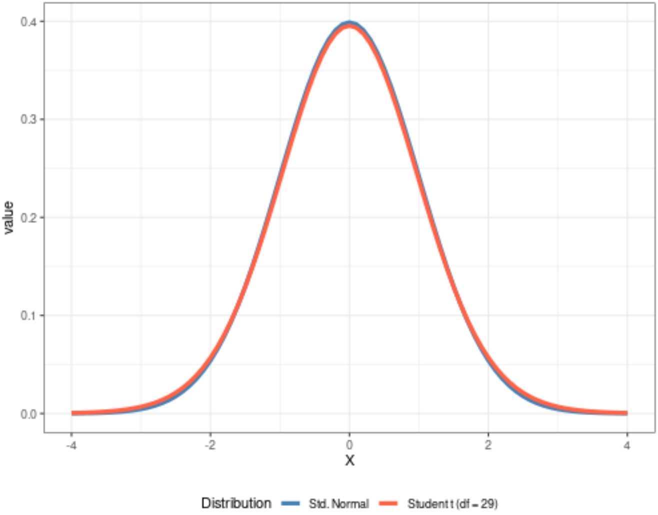
t-distribution with df=3



t-distribution with df=10

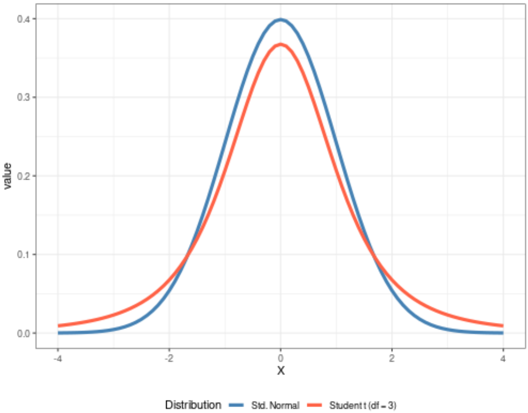


t-distribution with df=29



Implications?

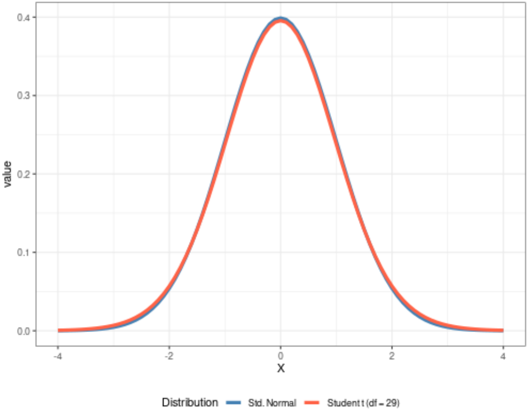
What are the implications of this for our confidence intervals?

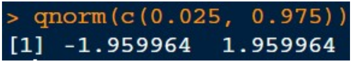


- we can't just use ±1.96×SE
- we need to go out further because there is more probability in the tails

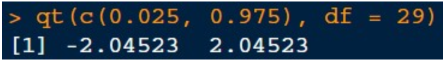
CI's using t-distribution

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





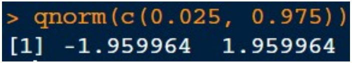
```
> qnorm(c(0.025, 0.975))
[1] -1.959964  1.959964
```



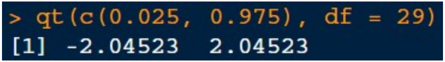
```
> qt(c(0.025, 0.975), df = 29)
[1] -2.04523  2.04523
```

CI's using t-distribution

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



```
> qnorm(c(0.025, 0.975))
[1] -1.959964  1.959964
```



```
> qt(c(0.025, 0.975), df = 29)
[1] -2.04523  2.04523
```

Instead of using:

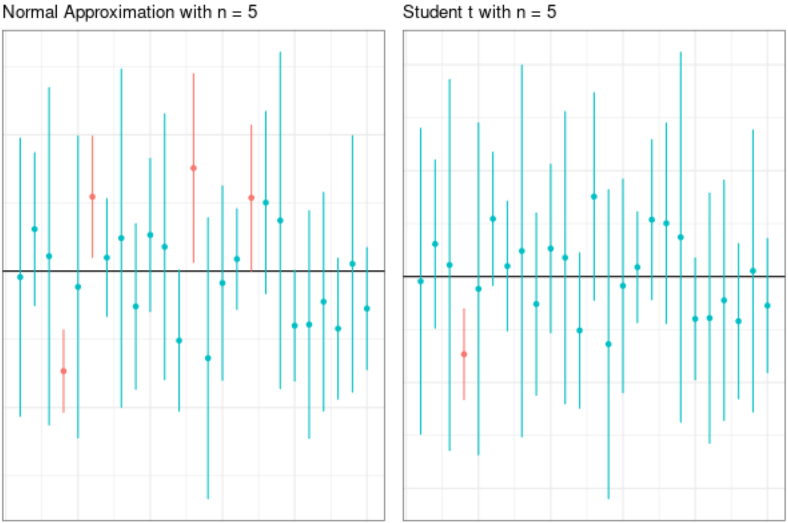
$$\overline{x} \pm 1.96 \times \frac{s}{\sqrt{n}},$$

We will use:

$$\overline{x} \pm t_{(.975, df=n-1)} \times \frac{s}{\sqrt{n}},$$

- $t_{(.975, df=n-1)}$ is the value corresponding to the .975 quantile of a t-distribution using df=n-1
- we use `qt(.975, df = n - 1)` in R to get this value

Comparing CI's between Normal and t-dist



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.

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 qt() function in R
- the appropriate value to put in the funtion is $(1 - \frac{\alpha}{2})$

General Formula for 100(1-α)% CI

$$\overline{x} \pm t_{(1-\frac{\alpha}{2}, df=n-1)} \times \frac{s}{\sqrt{n}},$$