Inference for Means in Small Samples

David Gerard

2017-11-01

Learning Objectives

- Introduce t-distribution.
- CI's and testing using the t-distribution.
- Section 5.1 of DBC.

Review Normal-based Confidence Intervals

Normal CI

When we wanted a $(1-\alpha)$ confidence interval for a mean, and we had a sample $X_1, X_2, \ldots, X_n$ such that $E[X_i] = \mu$ and $\text{var}(X_i) = \sigma^2$, we used the fact that for large n

$$\bar{X} \approx N(\mu, \sigma^2/n).$$

i.e. that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1).$$

Normal CI

Based on α, we found a z_α such that

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \in [-z_\alpha, z_\alpha]\right) = 1-\alpha$$

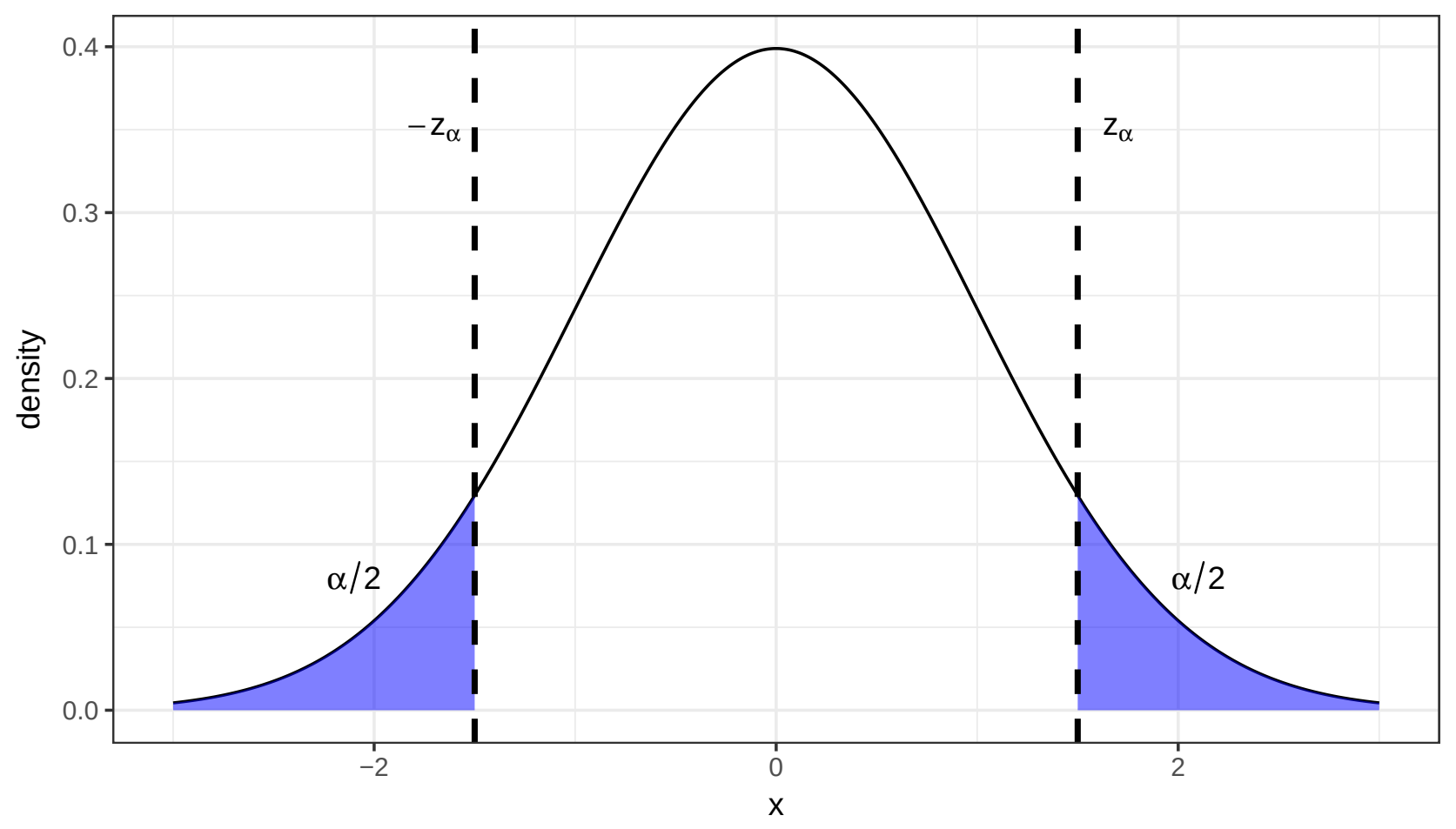
Rearranging terms, we got

$$P(\bar{X} - z_\alpha\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_\alpha\sigma/\sqrt{n}) = 1-\alpha.$$

And so if we know the population standard deviation (σ), our $(1-\alpha)$ confidence interval was

$$\bar{X} - z_\alpha\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_\alpha\sigma/\sqrt{n}$$

Finding z_α



You can use `qnorm` to find z_α.

Variance Unkown

This CI is valid only if the variance σ^2 is known.

Most of the time, σ^2 is not known.

If n is large enough, we can replace σ with s and the CI is still approximately correct. Mainly because of the Law of the Large Numbers

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \underset{n\to\infty}{\longrightarrow} \sigma^2$$

Variance Unknown

That is, for large n, we have

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \approx N(0, 1).$$

and so we find a z_α such that

$$P\left(\frac{\bar{X} - \mu}{s/\sqrt{n}} \in [-z_\alpha, z_\alpha]\right) = 1 - \alpha$$

Rearranging terms, we got

$$P(\bar{X} - z_\alpha s/\sqrt{n} \leq \mu \leq \bar{x} + z_\alpha s/\sqrt{n}) = 1 - \alpha.$$

t-based Confidence Intervals

Problem

However, for small n (rule of thumb $n \leq 30$), this approximation is not accurate! Not even when the $X_1, X_2, \ldots, X_n$ are exactly $N(\mu, \sigma^2)$!

Note:

To perform inference with small n, we will require that the X_i's are well approximated by a normal distribution.

Recall that for $X_1, X_2, \ldots, X_n$, independent with $X_i \sim N(\mu, \sigma^2)$, we have exactly

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

But we want the distribution of

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}.$$

t-distribution

Theorem

$X_1, X_2, \ldots, X_n$, independent with $X_i \sim N(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_\nu,$$

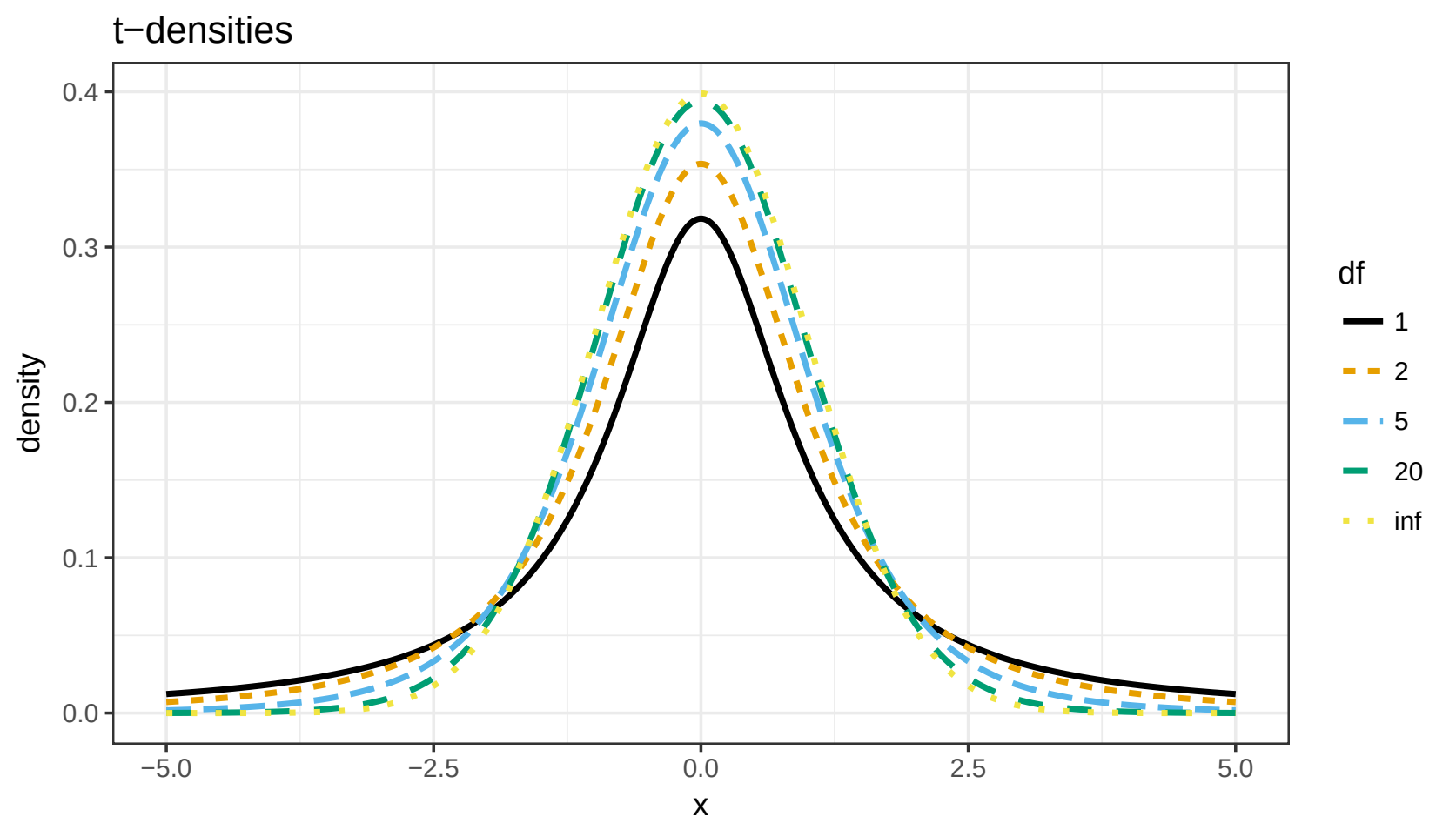
where t_{df} represents the t-distribution with ν degrees of freedom. Here, $\nu = n - 1$, one minus the sample size.

Properties of t

(Unlike the Normal or Binomial distributions, each of which has two parameters, the t-distribution has only one parameter, called the degrees of freedom.)

- Symmetric about zero
- Bell-shaped - similar to normal distribution
- More spread out than normal - heavier tails
- Exact shape depends on the degrees of freedom
- As the number of degrees of freedom (ν) increases, the t-distribution converges to the Normal distribution.
- ν must be greater than 0.

t-shape



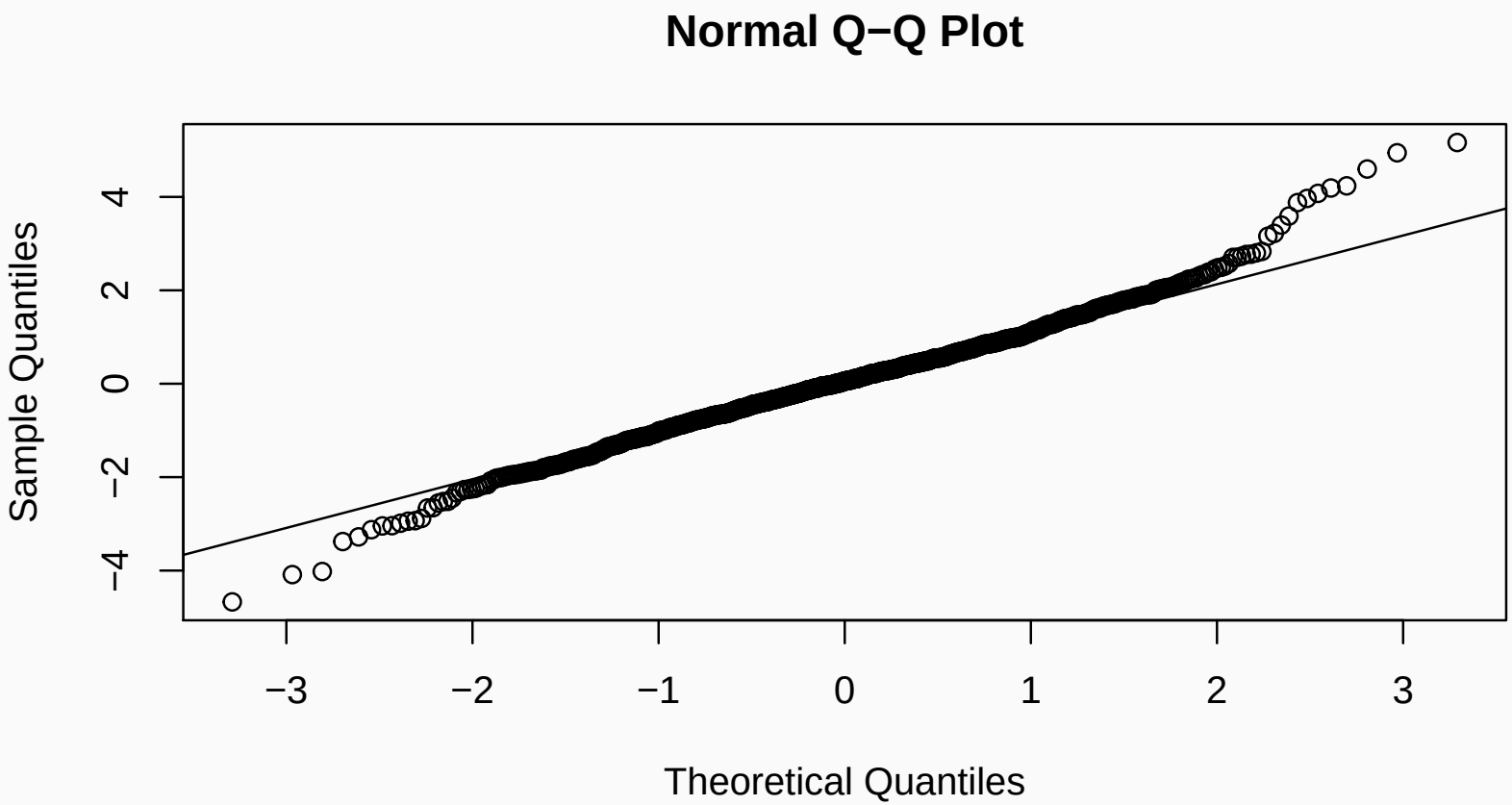
Empirical Example

```
x_matrix <- replicate(1000, rnorm(10))
xbar     <- colMeans(x_matrix)
s        <- apply(x_matrix, 2, sd)
tstat    <- xbar / (s / sqrt(10))
```

qq-plot using normal quantiles

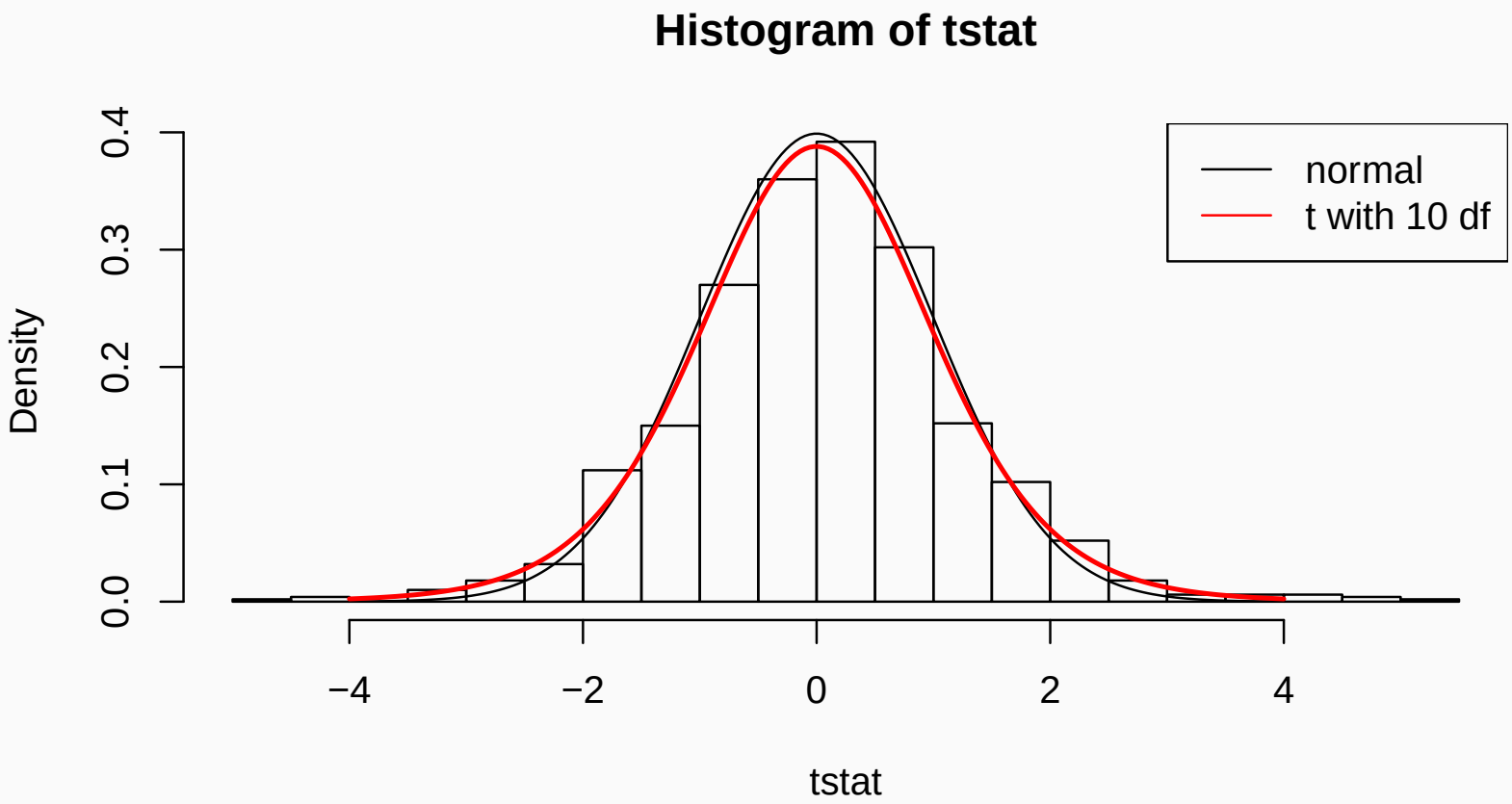
See heavier tails than expected under normal model

```
qqnorm(tstat)
qqline(tstat)
```



Histogram of t-statistics

t-distribution fits better in the tails



Confidence intervals with unknown σ

The goal is to find a confidence interval for μ when σ is unknown.

That is, we want a random interval that captures μ in $(1-\alpha)$ of repeated samples.

Since

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1},$$

we need to find a t^* such that

$$P\left(\frac{\bar{X}-\mu}{s/\sqrt{n}} \in [-t^*, t^*]\right) = 1-\alpha,$$

Confidence intervals with unknown σ

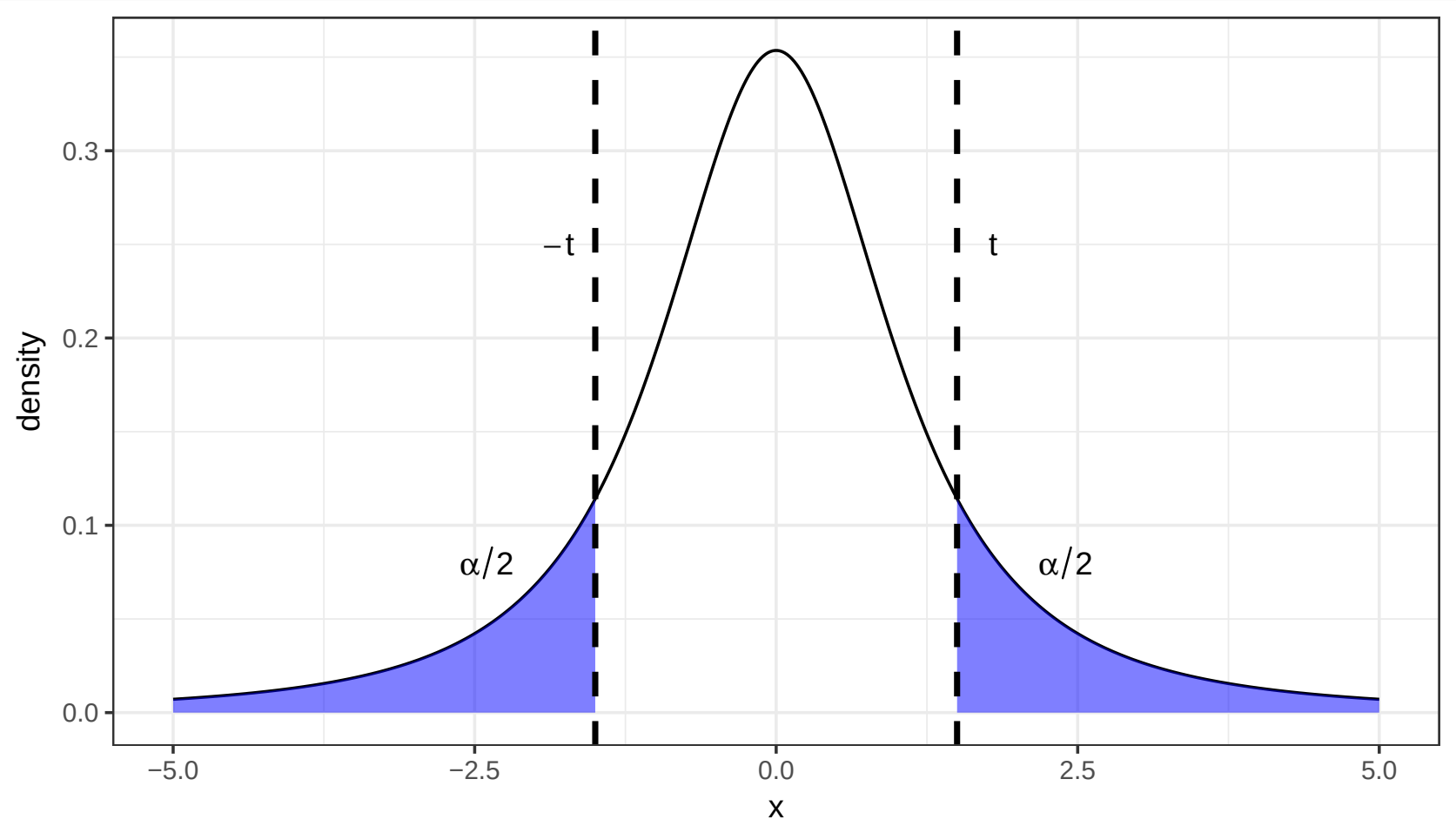
- Rearranging terms, we have

$$P\left(\bar{X} - t^* s/\sqrt{n}, \bar{X} + t^* s/\sqrt{n}\right) = 1 - \alpha.$$

- So $(\bar{X} - t^* s/\sqrt{n}, \bar{X} + t^* s/\sqrt{n})$ is a $(1-\alpha)$ confidence interval for the mean when σ is not known.
- You can use this for any sample size n, not just when n is small.
- But it will approximately equal the normal-based CI when n is large.
- These confidence intervals are again random. In addition to having a random center $\bar{X}$, they have a random width $t^* S/\sqrt{n}$.
- The t intervals are wider than the normal intervals because the t distribution has larger tails. This corrects for uncertainty in estimating σ.

How do you get t^*?

The critical value, $t^* = t_{n-1,\alpha}$ is chosen such that $(100(1-\alpha)\%$ of the area under the t_{n-1} density lies between $-t^*$ and t^*.



You can use the R function qt to find t^*.

Some notes on Approximation

1. If the underlying population is Normally distributed, the interval is exact. (i.e. exact if $X_1, X_2, \ldots, X_n$ are $N(\mu, \sigma^2)$).
2. Otherwise, the interval is approximately correct if n is not too small (say, $n \geq 15$), the data are not strongly skewed, and there are no outliers.
3. With n sufficiently large (say $n \geq 30$), the approximation is correct even if the data are clearly skewed.
4. For small sample sizes, this motivates taking transformations to make the data look more normal.

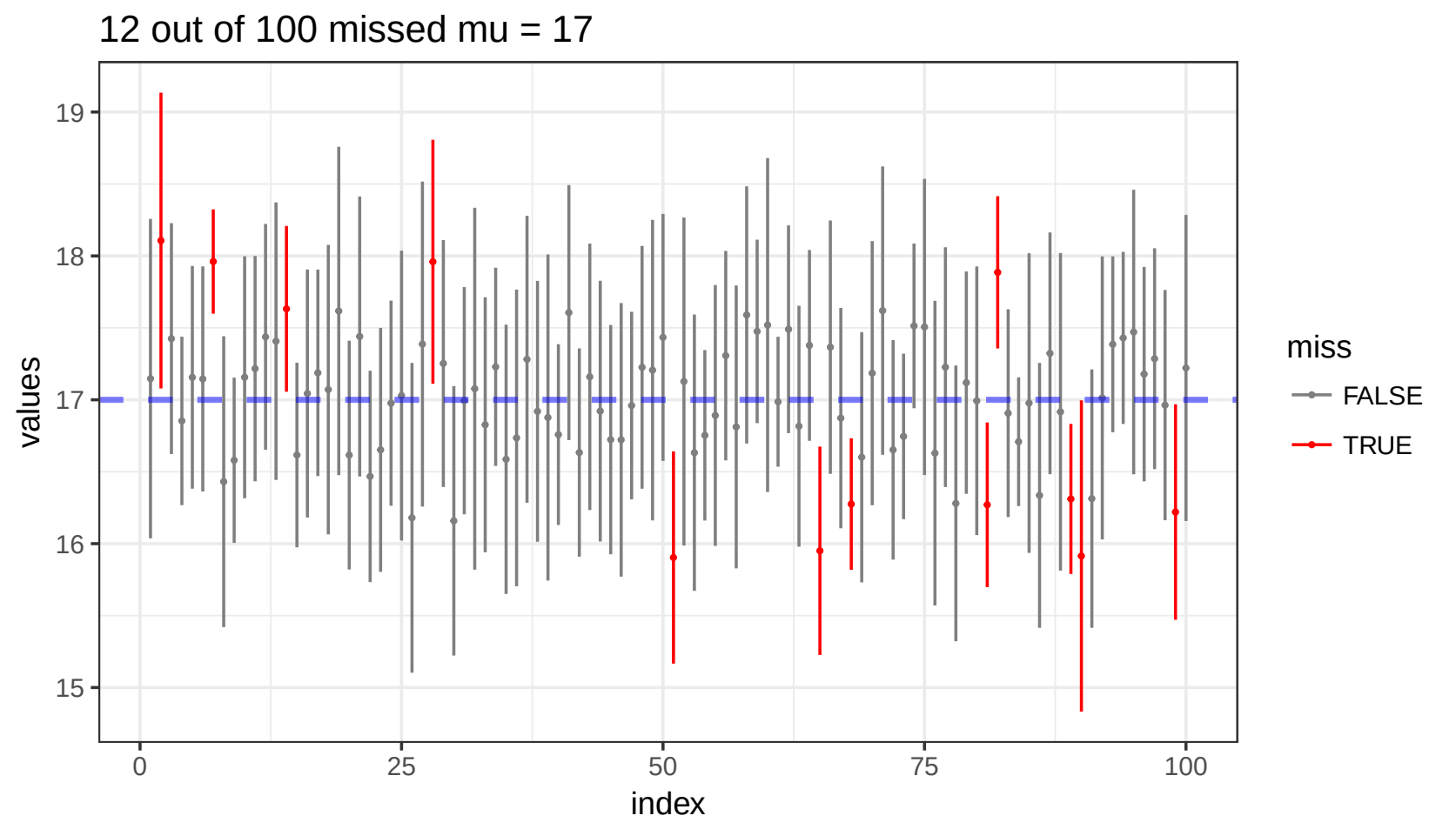
Does this really matter?

Simulate 100 samples and calculate their corresponding normal and t 95% confidence intervals:

```
mu       <- 17
sigma2   <- 2
n        <- 10
alpha    <- 0.05
x_matrix <- replicate(100, rnorm(n, mu, sqrt(sigma2)))
xbar     <- colMeans(x_matrix)
s        <- apply(x_matrix, 2, sd)
z_alpha  <- abs(qnorm(alpha / 2))
t_alpha  <- abs(qt(alpha / 2, df = n - 1))
lower_z  <- xbar - z_alpha * s / sqrt(n)
upper_z  <- xbar + z_alpha * s / sqrt(n)
lower_t  <- xbar - t_alpha * s / sqrt(n)
upper_t  <- xbar + t_alpha * s / sqrt(n)
```

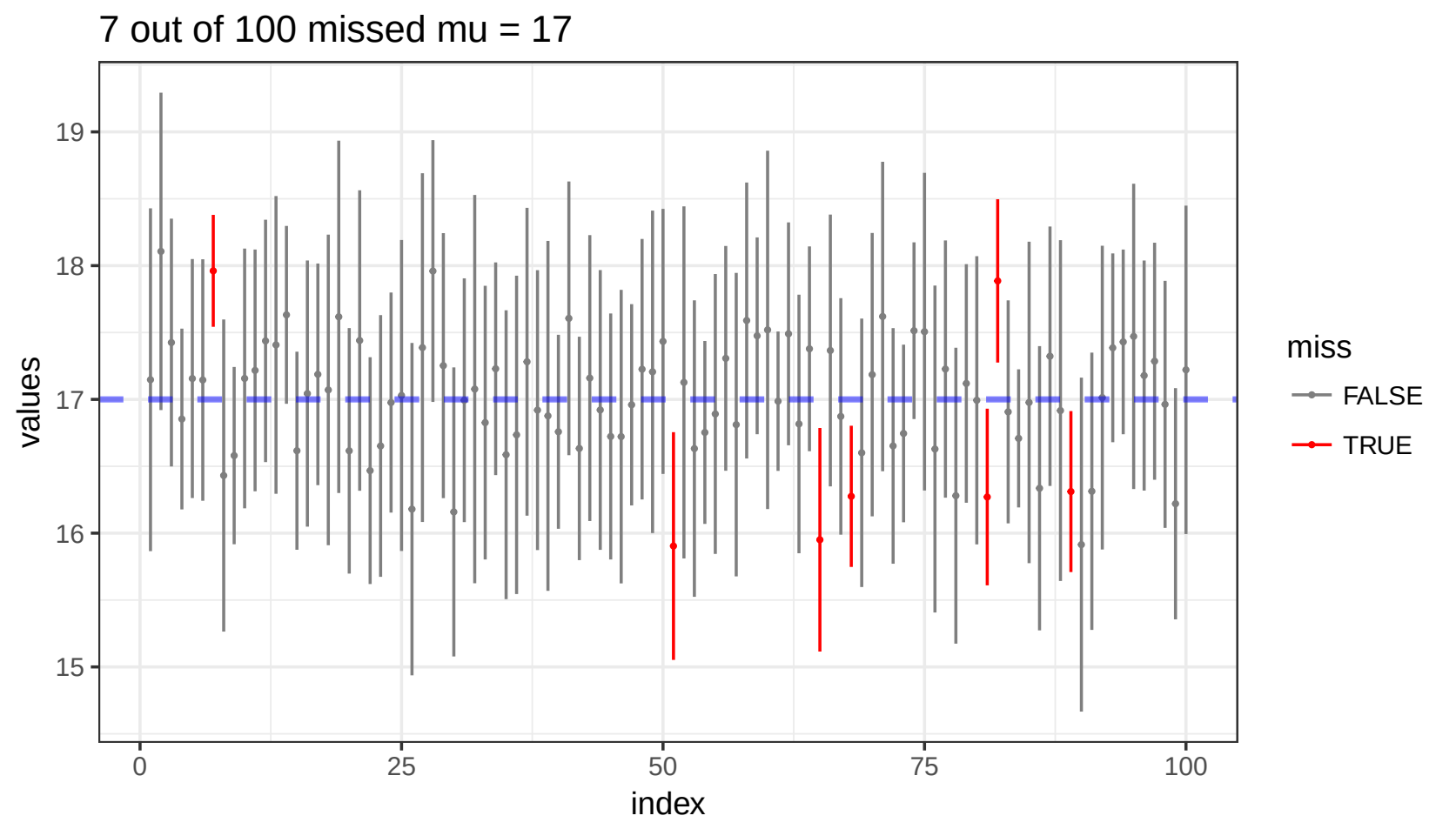
Does this really matter?

Normal based intervals



Does this really matter?

t based intervals



t-tests

One t-tests

- We can also use the t-distribution for hypothesis testing.
- Suppose $X_1, X_2, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ (e.g. from an SRS of a population that is normal).
- We want to test
  - H_0: $\mu = \mu_0$ versus
  - H_A: $\mu \neq \mu_0$.
- Then we know under H_0 that the test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

  has a t-distribution with $n - 1$ d.f.

One sample t-tests

- The p-value for this test is the probablity that T is as extreme or more exteme than our observed test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

- For the two-sided alternative hypothesis H_A: $\mu \neq \mu_0$, we calculate the two tail probabilities

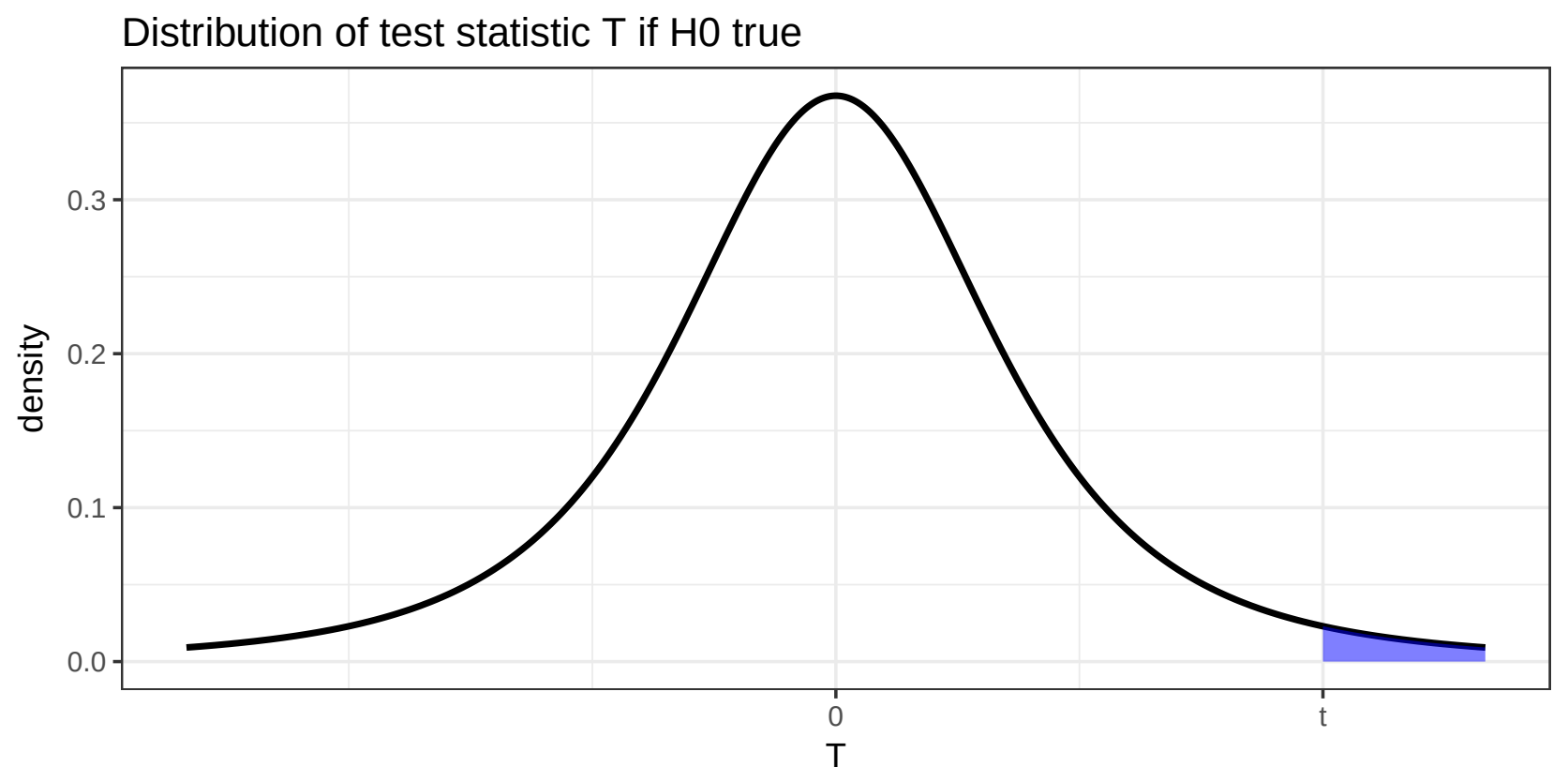
$$2P(T_{n-1} \geq |t|).$$

p-values

This is equal to: `2 * pt(-abs(t), df = n - 1)`.

One-sided alternative

For a one-sided alternative H_A: $\mu > \mu_0$, the p-value is $P(T_{n-1} \geq t)$.



This is equal to: `pt(t, df = n - 1, lower.tail = FALSE)`.

one-sided alternative

For a one-sided alternative H_A: $\mu < \mu_0$, the p-value is $P(T_{n-1} \leq t)$.

This is equal to: `pt(t, df = n - 1)`.

Tumor Growth Example: setup

- Let X (in mm) denote the growth in 15 days of a tumor induced in a mouse. It is known from a previous experiment that the average tumor growth is 4mm.
- A sample of 20 mice that have a genetic variant hypothesized to be involved in tumor growth yielded $\bar{x} = 3.8$mm and $s = 0.3$mm.
- Test whether $\mu = 4$ or not, assuming growths are normally distributed.

Tumor Growth Example: solution

1. State the hypotheses:

$$H_0 : \mu = 4 \text{ versus } H_A : \mu \neq 4.$$

2. Calculate the t-statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.8 - 4.0}{0.3/\sqrt{20}} = -2.98$$

3. Determine the p-value

$$p = 2P(T_{19} \geq 2.98) = 0.008.$$

Significance Level

- We could have chosen a significance level α ahead of time (usually $\alpha = 0.05$) and then reject H_0 if our p-value fell below this threshold. Ideally you choose this before running the hypothesis test.
- E.g., we could reject H_0 at level $\alpha = 0.01$ and conclude that the population mean growth is not 4mm.
- Note: Since we reject H_0 if $p \leq \alpha$, the p-value has the interpretation of being the smallest signficance level at which we would reject H_0.

99% CI

- Remember the relationship between hypothesis testing and confidence intervals?
- Let's construct a 99% CI for μ:

$$\begin{aligned}
&(\bar{x} - t^* s/\sqrt{n}, \bar{x} + t^* s/\sqrt{n}) \\
&= (3.8 - 2.861 \times 0.3/\sqrt{20}, 3.8 + 2.861 \times 0.3/\sqrt{20}) \\
&= (3.61, 3.99),
\end{aligned}$$

  where $t^* : P(|T_{19} > t^*) = 0.01$.
- Using `abs(qt(0.005, df = 19))` in R, this is 2.8609.
- Note that 4 is outside this CI. From this, we can draw the same conclusion as from the test. Namely, at significance level $\alpha = 0.01$, the mean growth is not equal to 4mm.

Relationship between CI and hypothesis tests

- A two-sided hypothesis test with significance level α rejects the null hypothesis $H_0 : \mu = \mu_0$ if and only if the value of μ_0 falls outside the $100(1 - \alpha)\%$ CI for μ.

- Reporting a CI is generally more informative than just reporting a *p*-value or the decision made on the basis of a hypothesis test since if tells the reader about your level of uncertainty (MOE).

One-sided alternatives

- In the previous example, suppose we wished to test $\mu < 4$ as our alternative.

1. State Hypotheses. $H_0 : \mu = 4$ versus $H_A : \mu < 4$.
2. Calculate the t-statistics. $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{3.8-4}{0.3/\sqrt{20}} = -2.98$.
3. Determine the p-value. $p = P(T_{19} \leq -2.98) =$ `pt(-2.98, df = 19)` $= 0.0038$.

- Since $0.0038 = p \leq \alpha = 0.1$, we reject H_0 at significance level 0.01 and conclude that mean growth is less than 4mm.