Hypothesis Testing

Grinnell College

March 25, 2026

Warm-up

We have a lot to do today

Goals for Today

1. t Distribution
2. Introduce mechanics of hypothesis testing
3. Identify null distribution for a given hypothesis
4. Learn to use test statistics to evaluate plausibility of a given hypothesis

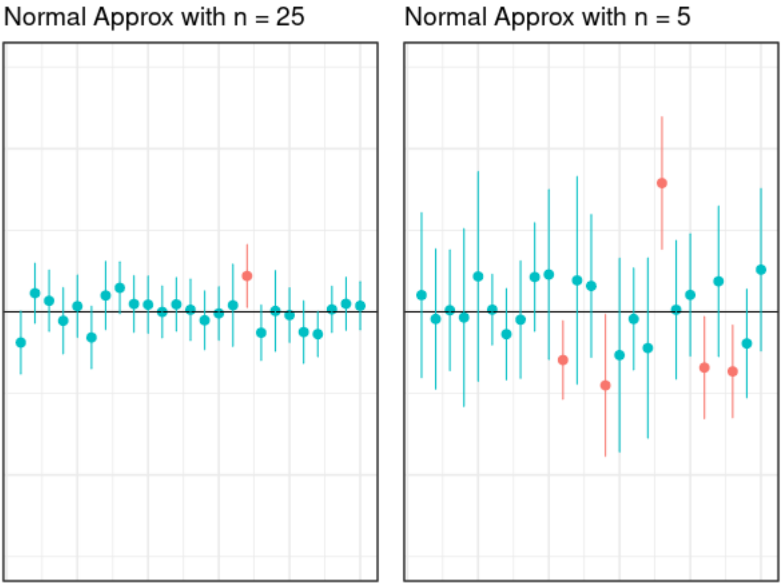
Part I: t Distribution

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

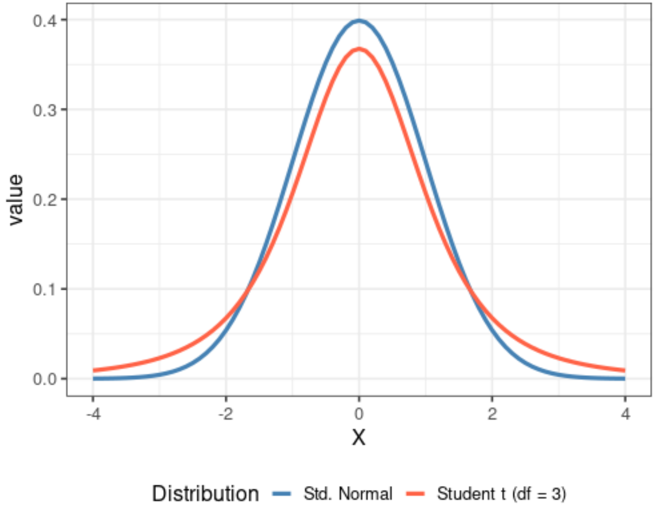
The sampling distribution of the t-statistic is known as the t-**distribution**

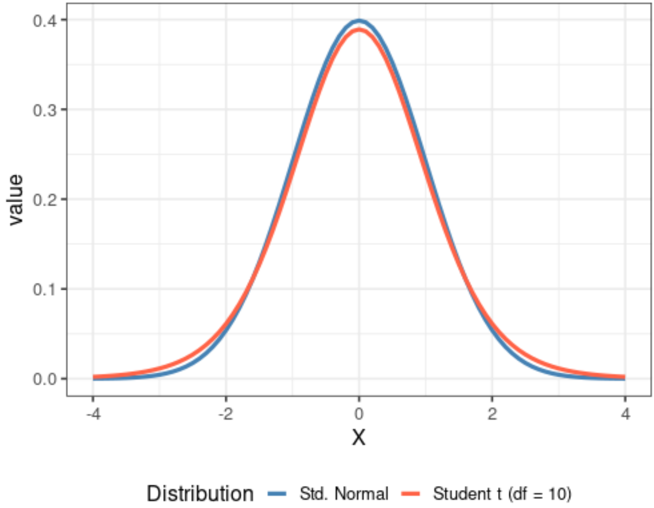
Student's t-distribution

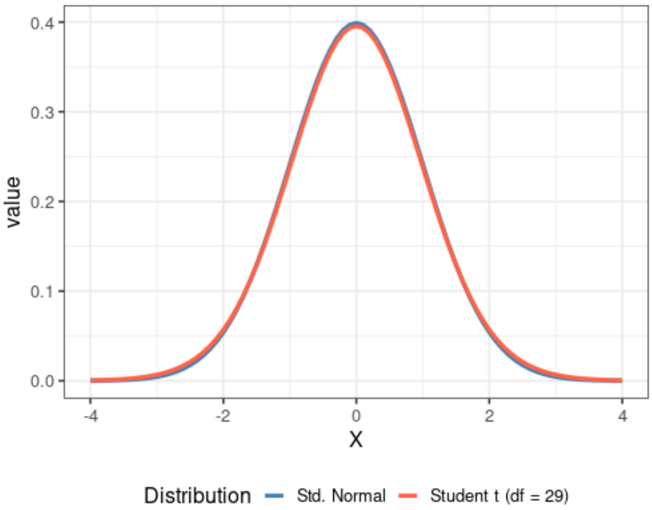
The t statistic and the t-distribution are very similar to a z statistic and a standard normal distribution:

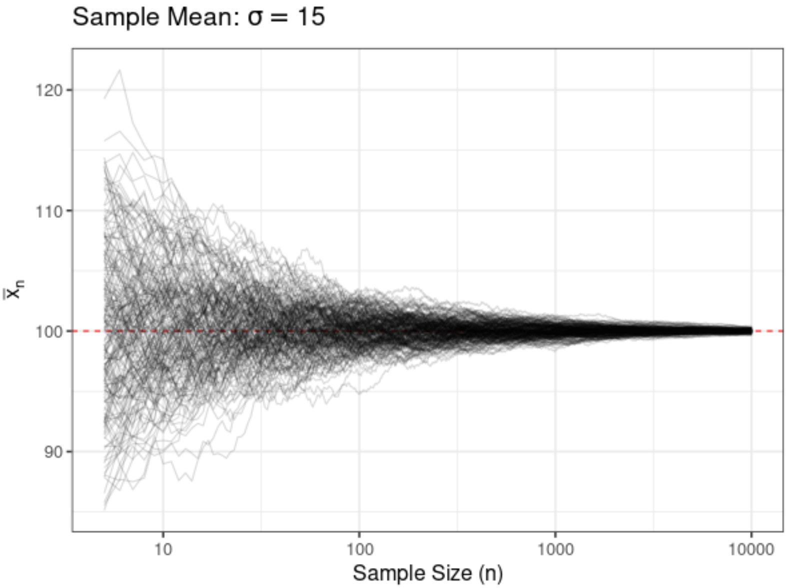
$$t = \frac{\overline{x} - \mu}{\hat{\sigma}/\sqrt{n}}, \qquad t \sim t(n-1)$$

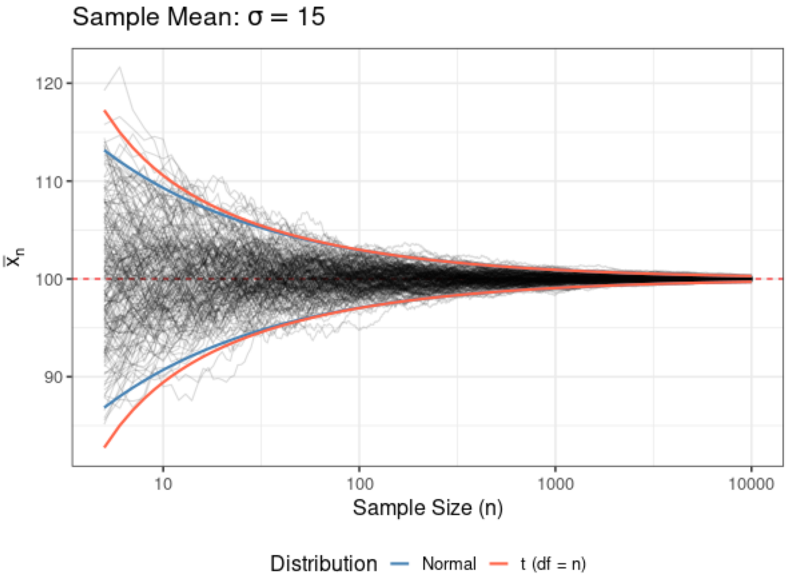
1. The t distribution is symmetric around 0
2. The t distribution has only one *distributional parameter* called the *degrees of freedom*, equal to $n-1$. This controls the variability
3. The t distribution has "fatter tails" than the normal distribution, allowing for the possibility of larger values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

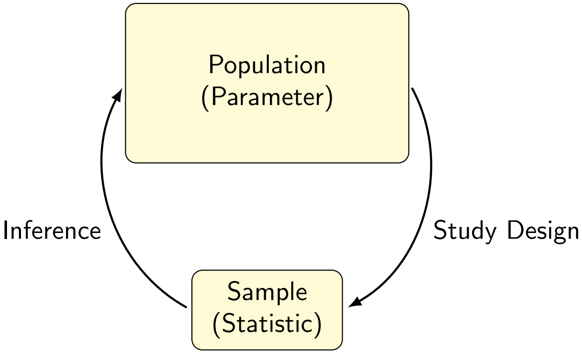
$$\overline{x} \pm 2.5706 \times \frac{\hat{\sigma}}{\sqrt{6}}$$

As opposed to our estimate with a standard normal,

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

Part II: Hypothesis Testing

The Statistical Framework



Hypothesis Testing

Hypothesis testing involves:

1. Formulating an *unambiguous* statement about a population parameter, called our **null hypothesis**
2. Collecting observational or experimental data
3. Determining if the data collected is consistent with our hypothesis
4. Either *rejecting* or *failing to reject* our hypothesis based on the *strength* of the evidence

Null Hypothesis

Our hypothesis about a parameter prior to seeing any data is called our **null hypothesis**, typically expressed in the form

$$H_0 : \mu = \mu_0$$

where μ_0 ("mew-naught") represents a specific quantity.

This addresses the fact that, in actuality, *we do not know the value of* μ. Instead, our goal will be to posit a value and determine whether or not this hypothesis is consistent with the data observed.

Test statistics

We relate the data that we have observed (i.e., $\overline{x}$, $\hat{\sigma}$) with our null hypothesis with the use of **test statistics**

For example, if $H_0 : \mu = \mu_0$ is correct, then

$$t = \frac{\overline{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

will be centered at zero and will follow a t-distribution. If H_0 is *not* correct, the distribution of t statistics will be centered at $\overline{x} - \mu_0$ instead

The **null distribution** describes the distribution that our test statistics will follow if the null hypothesis is true

This is what it looks like when we subtract true mean and this is what it looks like when we subtract WRONGO MEAN. The more wrong, the more different it is. This is a pregnant idea

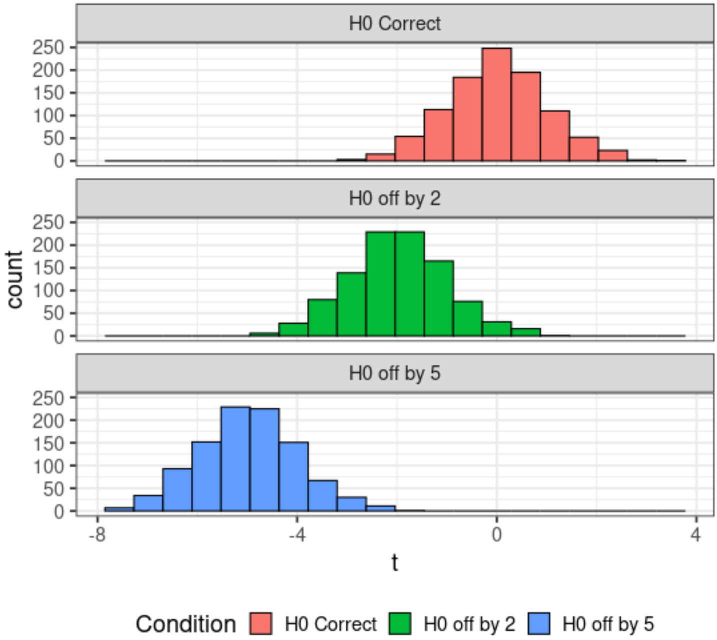
Null vs Sampling Distribution

Suppose there are true population parameters $\mu = 10$ and $\sigma = 5$. Further, suppose we have three separate hypotheses for H_0:

1. $H_0 : \mu = 10$
2. $H_0 : \mu = 12$
3. $H_0 : \mu = 15$

If we were to collect many samples of $\overline{X}$ and construct t statistics from each one, only one of the resulting distributions will be centered at 0. For both $H_0 : \mu = 12$ and $H_0 : \mu = 15$, will we be subtracting too much

This will impact how likely we are to reject our null hypothesis

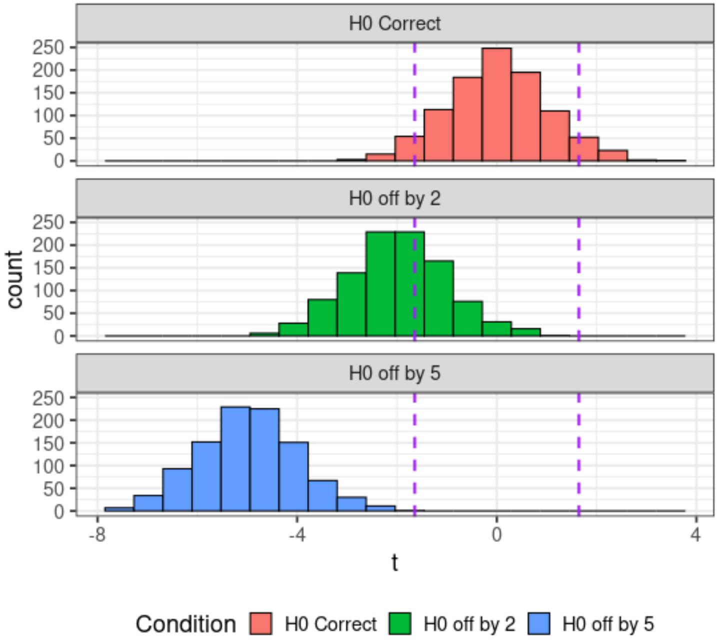

H0 Correct
H0 off by 2
H0 off by 5
count
250
200
150
100
50
0
-8
-4
0
4
t
Condition
H0 Correct
H0 off by 2
H0 off by 5

Hypothesis Testing and Confidence

How do we go about using a null distribution?

Assuming the null hypothesis is true, we know what distribution our statistic should follow and, accordingly, the *critical values* associated with the bounds of our distribution

Test statistics that fall sufficiently outside of the bounds of where we expect our data to fall may be considered evidence *against* the null hypothesis


H0 Correct
H0 off by 2
H0 off by 5
count
250
200
150
100
50
0
-8
-4
0
4
t
Condition
H0 Correct
H0 off by 2
H0 off by 5

The General Idea

The general idea is this:

1. We specify some null hypothesis $H_0 : \mu = \mu_0$
2. *Assuming that the null hypothesis is true*, the test statistic

$$t = \frac{\overline{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

   will follow a t-distribution, centered at 0
3. From our data, we will compute a test statistic using both $\overline{x}$ and μ_0
4. We will then check our test statistic t against critical values C
5. If $C < |t|$, our test statistically is sufficiently far from what we expect and we reject our null hypothesis

Penguin Example

Suppose we are studying the length (mm) of male penguin flippers, beginning with two competing hypotheses:

1. $H_{0_1} : \mu = 203$
2. $H_{0_2} : \mu = 215$

We collect sample of size $n = 20$, finding sample statistics of:

1. $\overline{x} = 209.7$
2. $\hat{\sigma}/\sqrt{n} = 3.342$

How would we test these hypotheses at confidence levels of 80%, 90%, and 95%?

Penguin Example

We essentially have two ways that are mathematically equivalent:

1. Check if μ_0 is in our confidence interval $\overline{X} \pm C \times \left(\frac{\hat{\sigma}}{\sqrt{n}}\right)$
2. Check if our t statistic falls between critical values $|t| < C$

With $n = 20$ we have $df = 19$, and from our critical value sheet we find critical values for 80, 90, and 95 to be

$$C_{80} = 1.33, \quad C_{90} = 1.73, \quad \text{and} \quad C_{95} = 2.09$$

Confidence Interval Method (clunky)

We have $\overline{x} = 209.7$ and $\hat{\sigma}/\sqrt{n} = 3.34$, with critical values

$$C_{80} = 1.33, \quad C_{90} = 1.73, \quad \text{and} \quad C_{95} = 2.09$$

Using

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

we find intervals:

- 80%: (205.26, 214.14)
- 90%: (203.92, 215.48)
- 95%: (202.72, 216.68)

We can simply check inclusion of $H_{0_1} : \mu = 203$ and $H_{0_2} : \mu = 215$

Critical value method (elegant)

We have $\overline{x} = 209.7$ and $\hat{\sigma}/\sqrt{n} = 3.34$, with critical values

$$C_{80} = 1.33, \quad C_{90} = 1.73, \quad \text{and} \quad C_{95} = 2.09.$$

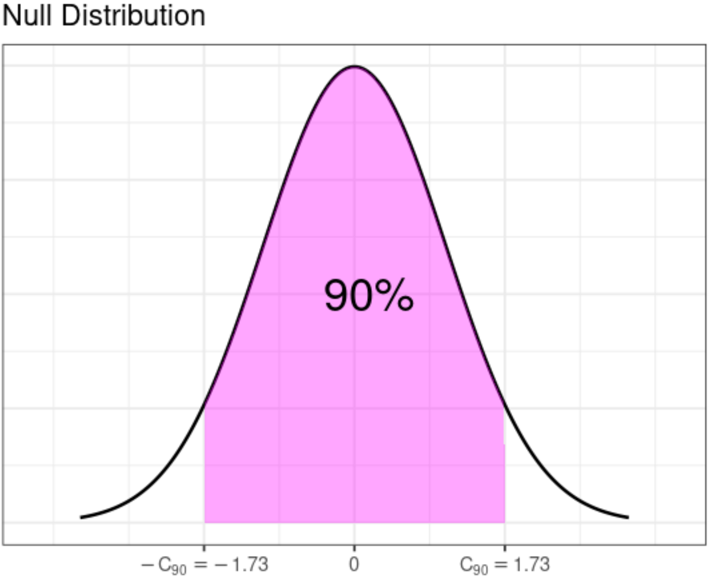
We can then construct t-statistics from each of our hypotheses $H_{0_1} : \mu = 203$ and $H_{0_2} : \mu = 215$

$$t_1 = \frac{209.7 - 203}{3.34} = 2.006, \qquad t_2 = \frac{209.7 - 215}{3.34} = -1.587$$

By checking if $|t| < C$ we will come to the same conclusions:

1. Reject both with 80% confidence
2. Reject H_{0_1} with 90% confidence but not H_{0_2}
3. Reject neither with 95% confidence

$$t_1 = 2.006 \text{ and } t_2 = -1.587$$



Review Steps

- Formulate a null hypothesis H_0
- Use this and your sample data to construct a test statistic (i.e., t-statistic)
- If the null hypothesis is true, the t-statistic will follow a t-distribution centered at zero. This is our *null distribution*
- Find the critical values of the null distribution, i.e., *if the null hypothesis were true*, what are the bounds of where we would expect to see our data?
- Compare your statistic to the critical values. If our statistic exceeds those bounds, we reject H_0